Statistica Sinica **22** (2012), 1123-1146
doi:http://dx.doi.org/10.5705/ss.2009.210

TUNING PARAMETER SELECTION FOR PENALIZED LIKELIHOOD ESTIMATION OF GAUSSIAN GRAPHICAL MODEL

Xin Gao, Daniel Q. Pu, Yuehua Wu and Hong Xu

York University

Abstract: In a Gaussian graphical model, the conditional independence between two variables are characterized by the corresponding zero entries in the inverse covariance matrix. Maximum likelihood method using the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)) has been proposed in the literature. In this article, we establish the result that when p is fixed, using the Bayesian information criterion (BIC) to select the tuning parameter in penalized likelihood estimation with the SCAD penalty can lead to consistent graphical model selection. When p increases with the sample size, a modified BIC with an extra penalty term is proposed. It can consistently select the true graphical model under the condition that p tends to infinity and all the true edges are included in a bounded subset. We compare the empirical performance of BIC with the cross validation method and demonstrate the advantageous performance of BIC criterion for sparse graphical models through simulation studies.

Key words and phrases: BIC, consistency, cross validation, Gaussian graphical model, model selection, oracle property, penalized likelihood.

1. Introduction

A multivariate Gaussian graphical model is also known as a covariance selection model. The conditional independence relationships between the random variables are equivalent to specified zeros among the inverse covariance matrix. More exactly, let $X = (X^{(1)}, \ldots, X^{(p)})$ be a p-dimensional random vector following a multivariate normal distribution $N_p(\mu, \Sigma)$, with μ denoting the unknown mean and Σ denoting the nonsingular covariance matrix. Denote the inverse covariance matrix as $\Sigma^{-1} = C = (C_{ij})_{1 \le i,j \le p}$. Zero entries C_{ij} in the inverse covariance matrix indicate conditional independence between the random variables $X^{(i)}$ and $X^{(j)}$ given all other variables (Dempster (1972), Whittaker (1990), Lauritzen (1996)). The Gaussian random vector X can be represented by an undirected graph $G = (V, E)$, where V contains p vertices corresponding to the p coordinates and the edges $E = (e_{ij})_{1 \le i < j \le p}$ represent the conditional dependency relationships between variables $X^{(i)}$ and $X^{(j)}$. It is of interest to identify

the correct set of edges, and estimate the parameters in the inverse covariance matrix simultaneously.

To address this problem, many methods have been developed. In general, there are no zero entries in the maximum likelihood estimate, which results in a full graphical structure. Dempster (1972) and Edwards (2000) proposed to use penalized likelihood with the L_0-type penalty $p_\lambda(|c_{ij}|)_{i\neq j} = \lambda I(|c_{ij}| \neq 0)$, where $I(\cdot)$ is the indicator function. Since the L_0 penalty is discontinuous, the resulting penalized likelihood estimator is unstable. Another approach is stepwise forward selection or backward elimination of the edges. However, this ignores the stochastic errors inherited in the multiple stages of the procedure (Edwards (2000)) and the statistical properties of the method are hard to comprehend. Furthermore, the computational complexity of this greedy search algorithm increases exponentially with the number of vertices in the graph. Meinshausen and Bühlmann (2006) proposed a computationally attractive method for covariance selection; it performs the neighborhood selection for each node and combines the results to learn the overall graphical structure. It has been shown that this method is related to the quadratic approximation of the loglikelihood with L_1 penalty (Yuan and Lin (2007)). Nevertheless this method performs model selection and parameter estimation separately. Yuan and Lin (2007) proposed penalized likelihood methods for estimating the concentration matrix with the L_1 penalty (LASSO) (Tibshirani (1996)). The method can be implemented through the maxdet algorithm in convex optimization. However, due to the inherent computational complexity, the maxdet algorithm can only handle matrices with small p.

Banerjee, Ghaoui, and D'Aspremont (2007) proposed a block-wise updating algorithm for the estimation of the inverse covariance matrix. For each block-wise update, the problem is a box-constrained quadratic program that can be solved by an interior-point procedure. They further showed that the problem that emerges from each step of block-wise update is equivalent to a linear regression under the L_1 penalty. Further in this line, Friedman, Hastie, and Tibshirani (2008) proposed the graphical LASSO algorithm to estimate the sparse inverse covariance matrix using the LASSO penalty through a coordinate-wise updating scheme. It is presently the fastest and most convenient algorithm to tackle this problem. Fan, Feng, and Wu (2009) proposed to estimate the inverse covariance matrix using the adaptive LASSO and the Smoothly Clipped Absolute Deviation (SCAD) penalty to attenuate the bias problem. They employed a local linear approximation method (Zou and Li (2008)) to approximate the LASSO penalty as weighted L_1 penalty, and the method is implemented through the graphical LASSO algorithm. The resulting methods with both SCAD and adaptive LASSO penalties are computationally convenient algorithms leading to asymptotically unbiased, sparse estimators that possess the oracle property.

In practice, the performance of the penalized likelihood estimator depends on the proper choice of the regularization parameter. Here we focus on the tuning parameter selection in penalized likelihood estimation of the sparse inverse covariance matrix. Wang, Li, and Tsai (2007) proposed using the Bayesian information criterion (BIC) to select the tuning parameter for the penalized likelihood method with SCAD penalty. They showed that BIC with the SCAD penalty is able to identify the true model consistently in the setting of linear regression and the partial linear model. Yuan and Lin (2007) used BIC to select the tuning parameter with the L_1 penalty in the estimation of the inverse covariance matrix. The consistency of BIC for the Gaussian graphic model has not yet been investigated. In this article we show that, for fixed p, the optimum tuning parameter selected by BIC with SCAD penalty yields the graphical structure of the true underlying graphical model with probability tending to one as $n \to \infty$. If p tends to infinity at a certain rate with the sample size, BIC needs to be modified with an extra penalty term. The modified BIC is consistent under the condition that $p \to \infty$ and the number of true edges d_T is bounded.

The rest of the article is organized as follows. In Section 2.1 we formulate the penalized likelihood function for the inverse covariance matrix. In Sections 2.2 and 2.3, we consider the case of p fixed, discuss the selection of tuning parameters through the BIC criterion, and prove its consistency in graphical model selection with SCAD and the adaptive LASSO penalty. In Section 3, we develop a modified BIC for consistent model selection when p tends to infinity. In Section 4, simulation studies are presented to demonstrate the empirical performance of the tuning parameter selection with BIC, compared with the cross validation method, in small p and large p scenarios. Throughout, we use $||\,||$ to denote the supreme norm, $||\,||_2$ the L_2 norm, and $||\,||_F$ the Frobenius norm, with $||A||_F = \sqrt{\text{tr}(A'A)}$.

2. Method

2.1. Penalized likelihood estimation of inverse covariance matrix

Given a random sample $X_1, \ldots, X_n$ from the multivariate normal $N_p(\mu, \Sigma)$, the loglikelihood for μ and $C = \Sigma^{-1}$ can be expressed as

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

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

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

Center the observations so that $\hat{\mu} = 0$. To obtain the maximum likelihood estimator of the concentration matrix, minimize

$$-\frac{2}{n}\ell(C) = -\log|C| + \mathrm{tr}(C\bar{A}).$$

In the penalized likelihood method, take $\hat{C}$ to minimize:

$$Q(C) = -\log|C| + \mathrm{tr}(C\bar{A}) + \sum_{i \neq j} p_\lambda(|c_{ij}|), \tag{2.1}$$

with p_λ some penalty function. Yuan and Lin (2007) proposed the LASSO penalty, $p_\lambda(|c_{ij}|) = \lambda|c_{ij}|$. Friedman, Hastie, and Tibshirani (2008) proposed a graphical LASSO algorithm using a coordinate descent procedure that is computationally very fast and guarantees the positive definiteness of the resulting estimate. As the LASSO penalty increases linearly with the size of its argument, this leads to biases for the estimates of nonzero coefficients. To attenuate such estimation biases, Fan and Li (2001) proposed the SCAD penalty p_λ, with $p_\lambda(0) = 0$ and first derivative

$$p'_\lambda(\theta) = \lambda\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\}, \text{ for } \theta > 0,$$

where a is some constant usually set to 3.7 (Fan and Li (2001)), and $(t)_+ = tI(t > 0)$ is the hinge loss function.

The SCAD penalty is a quadratic spline function with knots at λ and $a\lambda$. It is singular at the origin, which ensures the sparsity and continuity of the solution. The penalty function does not penalize as heavily as does the L_1 penalty function on large parameters. More importantly, the SCAD penalty not only selects the correct set of edges, but also produces parameter estimators as efficient as if the true underlying graphic structure is known, the oracle property.

Fan, Feng, and Wu (2009) proposed using local linear approximation (Zou and Li (2008)) to approximate the SCAD by a symmetric linear function. The proposed iterative re-weighted penalized likelihood method optimizes the objective function at step $(k+1)$ as:

$$Q(C)^{(k+1)} = -\log|C| + \mathrm{tr}(C\bar{A}) + \sum_{i \neq j} w_{ij}|c_{ij}|, \tag{2.2}$$

with $w_{ij} = p'_\lambda(|\hat{c}_{ij}^{(k)}|)$ and $\hat{c}_{ij}^{(k)}$ denoting the estimates obtained at previous step. The computation can be implemented by reiteratively using the graphical LASSO algorithm.

2.2. Consistency of BIC with fixed p

For the tuning parameter λ, it is desirable to have a data-driven method. Let the full graphical model be G_F, with the edge set $E_F = (e_{ij})_{1\le i<j\le p}$. Let an arbitrary graphical model be G with the edge set $E \subseteq E_F$, and the true model be G_T, with edge set $E_T = (e_{ij})_{(i,j):c_{ij,0}\neq 0, i<j}$, where $c_{ij,0}$ denotes the null value of the parameter. An over-fitted model G has edge set $E \supseteq E_T$ and $E \neq E_T$, and the collection of all over-fitted model is denoted by $\mathcal{G}_+$. An under-fitted model G has edge set $E \not\supseteq E_T$, and the collection of all over-fitted model is denoted by $\mathcal{G}_-$. Let d_T denote the number of true edges and d_G denote the number of edges in graph G.

In practice, as λ is unknown, we search for the optimal λ from the bounded interval $\Omega = [0, \lambda_{\max}]$, for some upper limit $\lambda_{\max}$. We further assume that the upper limit $\lambda_{\max} \to 0$, as $n \to \infty$. This implies that the search region shrinks to 0 as n tends to infinity. A similar assumption can be found in Wang, Li, and Tsai (2007). Given a tuning parameter λ, the penalized likelihood approach yields the estimated parameters $(\hat{c}_{ij,\lambda})_{1\le i\le j\le p}$. The resulting model is denoted by G_λ with edge set $E_\lambda = (e_{ij})_{(i,j):\hat{c}_{ij,\lambda}\neq 0}$. Given a λ, the associated BIC criterion is:

$$BIC_\lambda = -n \log |\hat{C}_\lambda| + n\mathrm{tr}(\hat{C}_\lambda \bar{A}) + \log(n) \sum_{1\le i<j\le p} I(\hat{c}_{ij,\lambda} \neq 0).$$

If we know the correct model G_T beforehand and obtain the maximum likelihood estimate $\hat{C}_{G_T}$, the associated BIC criterion is denoted as

$$BIC_{G_T} = -n \log |\hat{C}_{G_T}| + n\mathrm{tr}(\hat{C}_{G_T} \bar{A}) + \log(n) \sum_{1\le i<j\le p} I(c_{ij,0} \neq 0).$$

We focus the discussion on the SCAD penalty. We construct a working sequence of reference tuning parameters $\lambda_n = \log(n)/\sqrt{n}$ that satisfies the requirement that as $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$. Under such a working sequence of tuning parameters, according to Theorem 5.2 in Fan, Feng, and Wu (2009), with probability tending to one, the method not only identifies E_T, but also yields root-n consistent estimators for the nonzero c_{ij}'s.

Lemma 1. *For SCAD penalty,* $|BIC_{\lambda_n} - BIC_{G_T}| = O_p(1)$.

Next we establish the asymptotic order of the maximum difference of $\ell(\hat{C}_\lambda) - \ell(\hat{C}_{G_T})$ over $G_\lambda \in \mathcal{G}_-$.

Lemma 2. *There exists a constant* L_1 *such that*

$$\ell(C_\lambda) - \ell(\hat{C}_{G_T}) \le -L_1 n^{1/3}$$

with probability tending to 1 *uniformly for all the* $\lambda \in [0, \lambda_{\max}]$, *with* $G_\lambda \in \mathcal{G}_-$.

Because the penalty term is of order $\log n$, we have the following.

Theorem 1. *There exists a constant L_2 such that*

$$P(\inf_{G_\lambda \in \mathcal{G}_-} (BIC_\lambda - BIC_{G_T}) > L_2 n^{1/3}) \to 1, \text{ as } n \to \infty.$$

Theorem 2. *Under the regularity assumptions,*

$$P\Big\{\inf_{G_\lambda \in \mathcal{G}_-} BIC_\lambda > BIC_{\lambda_n}\Big\} \to 1, \text{ as } n \to \infty.$$

Next we consider an over-fitted model, in which some zero-valued parameters are included in the model to be estimated. They are regarded as nuisance parameters. So the property of the resulting BIC_λ can be derived under standard likelihood theory.

Theorem 3. *Under the regularity assumptions,* $Pr(\inf_{G_\lambda \in \mathcal{G}_+} BIC_\lambda > BIC_{\lambda_n}) \to 1$.

Theorems 2 and 3 imply that the λ's that fail to identify the true model have BIC larger than λ_n. Consequently, the λ value which minimizes the BIC criterion identifies the true model.

Theorem 4. *Under the regularity assumptions,* $Pr(G_{\hat{\lambda}_{BIC}} = G_T) \to 1$, *where* $\hat{\lambda}_{BIC}$ *is the tuning parameter that minimizes the BIC criterion with the SCAD penalty.*

It is worth pointing out that the penalty term of the BIC criterion is a step function of the smoothing parameter, so the minimum tuning parameter $\hat{\lambda}_{BIC}$ may not be unique. Nevertheless, the collection of tuning parameters $\hat{\lambda}_{BIC}$ that minimize the BIC correspond to the same correct model.

2.3. Consistency of modified BIC with $P_n \to \infty$

In the section above, the dimension of the covariance matrix p is fixed and the sample size n tends to infinity. In this section, we consider the situation in which p_n may depend on n, and p_n tends to infinity as n increases. Thus, in practice, researchers might include more variables as they increase the sample size. Under this high-dimensional setup, the penalized likelihood estimation of the covariance matrix has been investigated by Rothman et al. (2008) and Lam and Fan (2009).

To deal with high-dimensionality in generalized linear models, Chen and Chen (2008, 2012) proposed EBIC with an extra penalty on the size of the model space. For this setup, we propose to modify the BIC with an extra penalty term

of $4\log p_n$ on the dimension of the covariance matrix. Given a λ, the associated BIC criterion is:

$$BIC_\lambda = -n\log|\hat{C}_\lambda| + n\mathrm{tr}(\hat{C}_\lambda \bar{A}) + \{\log n + 4\log p_n\}\sum_{1\le i<j\le p} I(\hat{c}_{ij,\lambda}\neq 0).$$

If we know the correct model G_T beforehand, the associated BIC criterion is:

$$BIC_{G_T} = -n\log|\hat{C}_{G_T}| + n\mathrm{tr}(\hat{C}_{G_T}\bar{A}) + (\{\log n + 4\log p_n\}\sum_{1\le i<j\le p} I(c_{ij,0}\neq 0).$$

When p_n increases with n, the number of competing models increases. It is shown later in Lemmas 10 and 11 that the supreme difference between the likelihood of an over-fitting model and the true model is $O_p(4m\log p_n)$, given that the over-fitting model has m more parameters. To offset this difference, the penalty term adds an extra multiplying factor of $4\log p_n$ on the size of the model. Such a modification was also proposed in Foygel and Drton (2010), in which the likelihood of the EBIC was evaluated at the maximum likelihood estimate for all the sub-models.

We assume that there are constants τ_1 and τ_2 such that

$$0 < \tau_1 \le \lambda_{\min}(\Sigma_0) \le \lambda_{\max}(\Sigma_0) \le \tau_2 \le \infty, \text{ for all } n.$$

Such a condition uniformly bounds the eigenvalues of Σ_0 and allows a wide class of covariance matrices, as noted in Bickel and Levina (2008a,b). We further assume that d_T is bounded by a finite constant Q and $(p_n/n)(\log p_n)^k = O(1)$ for some $k > 1$. We search for the optimal λ from the bounded interval $\Omega = [0, \lambda_{\max}]$, for some upper limit $\lambda_{\max}$. We suppose $\lambda_{\max} \to 0$ and $\lambda_{\max} > \{(p_n/n)\log p_n\}^{1/2}$ as $n \to \infty$, and consider a working sequence of tuning parameters $\lambda_n = \{(p_n/n)\log p_n\}^{1/2}$. According to Theorems 1 and 2 in Lam and Fan (2009), there exists a local minimizer $\hat{C}_{\lambda_n}$ such that

$$||\hat{C}_{\lambda_n} - C_0||_F^2 = O_p\Big\{\frac{(p_n)\log p_n}{n}\Big\}.$$

Furthermore, with probability tending to 1, $\hat{c}_{ij,\lambda_n} = 0$, for all $c_{i,j} = 0$. This entails

$$\lim_{n\to\infty} P(G_{\lambda_n} = G_T) = 1.$$

Given any graph $G = (V, E)$, we can partition the concentration matrix into four sub-matrices, with the upper-left sub-matrix C^* the smallest sub-matrix containing all the non-zero off-diagonal entries, the upper-right C' and the lower-left C'' containing all zeros, and the lower-right C''' a diagonal matrix:

$$\begin{pmatrix} C^* & C' \\ C'' & C''' \end{pmatrix}.$$

When C is partitioned, the vector X is partitioned accordingly so that the sub-vector X^* has the covariance matrix C^*.

Given G_T, let the corresponding sub-matrices be $C^{*,T}$, $C^{',T}$, $C^{'',T}$, $C^{''',T}$. The sub-vector $X^{*,T}$ has the covariance matrix $C^{*,T}$. If $G_{\lambda_n} = G_T$, then $\hat{C}^{',T}_{\lambda_n} = 0$, $\hat{C}^{'',T}_{\lambda_n} = 0$, and $\hat{C}^{''',T}_{\lambda_n} = \hat{C}^{''',T}_{G_T}$, which entails

$$\ell(\hat{C}_{\lambda_n}; X) - \ell(\hat{C}_{G_T}; X) = \ell(\hat{C}^{*,T}_{\lambda_n}; X^{*,T}) - \ell(\hat{C}^{*,T}_{G_T}; X^{*,T}),$$

while the latter is the difference of likelihood based on the penalized likelihood estimator and the mle estimator given the true model evaluated at the sub-vector $X^{*,T}$. For notational convenience, X and $X^{*,T}$ are omitted from the likelihood notation.

Lemma 3. *Under the regularity conditions above,*

$$|BIC_{\lambda_n} - BIC_{G_T}| = O_p(1).$$

Given any G, we can construct an extended graph $\dot{G} = G_T \cup G$ with the extended edge set $\dot{E} = E_T \cup E$. Based on $\dot{G}$, we partition the concentration matrix into $C^{*,e}$, $C^{',e}$, $C^{'',e}$, and $C^{''',e}$. The sub-vector of X whose concentration matrix corresponds to $C^{*,e}$ is denoted by $X^{*,e}$. Then the score function

$$U_n(C_0^{*,e}; X^{*,e}) = \frac{\partial \ell(C^{*,e}, X^{*,e})}{\partial C^{*,e}}|_{C_0^{*,e}} = n\{(C_0^{*,e})^{-1} - S_{X^{*,e}}\},$$

where $S_{X^{*,e}}$ is the sample covariance matrix for the subset data $X^{*,e}$. Because $d_T \leq Q$, we restrict our model space to $\mathcal{G} = \{G : \dim(C^{*,e}) \leq 2Q\}$.

In order to establish the asymptotic consistency of the modified BIC, we need a technical lemma.

Lemma 4. *Let Z_i, $i = 1, \ldots, n$, be independent and identically distributed random variables with zero mean and unit variance. Assume there exists a constant δ such that, for $|t| \leq \delta$, the absolute value of the third derivative of their cumulant generating function $|g^{(3)}(t)| \leq M$ for some constant M. If f_n is a sequence such that $f_n \to \infty$ as $n \to \infty$, then for any $m > 0$, we have*

$$P(\sum_{i=1}^{n} Z_n > \sqrt{2mn \log f_n}) = o(f_n^{-m}). \tag{2.3}$$

Now we establish the asymptotic order of the maximum score functions over all the possible models in the model space $\mathcal{G}$.

Lemma 5. *Under the regularity conditions,*

$$\max_{G \in \mathcal{G}} ||U_n(C_0^{*,e})|| = O_p(n^{1/2}(\log p_n)^{1/2}).$$

Lemma 6. *For all $C^{*,e}$ induced by $G \in \mathcal{G}$, and for some constant L_3,*

$$\sup\{\ell(C^{*,e}) - \ell(C_0^{*,e}) : ||C^{*,e} - C_0^{*,e}||_2 > n^{-1/3}\} \leq -L_3 n^{1/3}, \tag{2.4}$$

uniformly with probability tending to one.

Next we consider under-fitted models.

Lemma 7. *For all λ such that $\lambda \in [0, \lambda_{\max}]$, and $G_\lambda \in \mathcal{G}_-$, let $C^{*,e}$ be induced by $\dot{G} = G_\lambda \cup G_T$. Then*

$$\ell(\hat{C}_\lambda^{*,e}) - \ell(C_0^{*,e}) \leq -L_4 n^{1/3}$$

for some constant L_4 with probability tending to 1 uniformly.

Lemma 8. *For all λ such that $\lambda \in [0, \lambda_{\max}]$, and $G_\lambda \in \mathcal{G}_-$, let $C^{*,e}$ be induced by $\dot{G} = G_\lambda \cup G_T$. Let $\hat{C}_{G_T}^{*,e}$ be the maximum likelihood estimator under G_T for submatrix $C^{*,e}$. Then*

$$\sup_{G_\lambda \in \mathcal{G}_-} |\ell(\hat{C}_{G_T}^{*,e}) - \ell(C_0^{*,e})| = O_p(1). \tag{2.5}$$

Theorem 5. *For all λ such that $\lambda \in [0, \lambda_{\max}]$, and $G_\lambda \in \mathcal{G}_-$,*

$$P(\inf_{G_\lambda \in \mathcal{G}_-} (BIC_\lambda - BIC_{G_T}) > L_5 n^{1/2}) \to 1,$$

as $n \to \infty$, for some constant L_5.

Theorem 6. *For all λ such that $\lambda \in [0, \lambda_{\max}]$, and $G_\lambda \in \mathcal{G}_-$,*

$$P\big\{\inf_{G_\lambda \in \mathcal{G}_-} BIC_\lambda > BIC_{\lambda_n}\big\} \to 1, \text{ as } n \to \infty.$$

Let $U(C^{*,e})$ denote the column vector of the first derivatives $\partial\ell/\partial c_{ij}$ for all $c_{ij} \in C^{*,e}$. Let $U_G(C^{*,e})$ be the sub-vector of the first derivatives $\partial\ell/\partial c_{ij}$ for all $c_{ij} \in C^{*,e}$, and $(i,j) \in E_G$. The indices of the score vectors are reordered as r from 1 to its length. Let $H(C^{*,e})$ be the matrix of the second derivatives $\partial^2\ell/\partial C_{ij}\partial C_{i'j'}$ for all $c_{ij} \in C^{*,e}$, and $c_{i'j'} \in C^{*,e}$. Let $H_G(C^{*,e})$ be the sub-matrix of the second derivatives $\partial^2\ell/\partial c_{ij}\partial c_{i'j'}$ for all $c_{ij} \in C^{*,e}$, and $(i,j) \in E_G$, all $c_{i'j'} \in C^{*,e}$, and $(i',j') \in E_G$. The indices of the Hessian are reordered as r and t. Next we examine the maximum difference between the log-likelihood ratio statistic and the score test statistic over all graphs in the model space.

Lemma 9.

$$\max_{G \in \mathcal{G}} |2\{\ell(\hat{C}_G^{*,e}) - \ell(C_0^{*,e})\} - U_G(C_0^{*,e})' H_G(C_0^{*,e})^{-1} U_G(C_0^{*,e})| = O_p(1).$$

Next we consider the over-fitted models. Let $\mathcal{G}_+(m) \subset \mathcal{G}_+$, with $\mathcal{G}_+(m) = \{G : G \in \mathcal{G}_+, d_G - d_T = m\}$, $m = 1, \ldots, Q - d_T$. Let $C^{*,e}$ be the upper left sub-matrix induced by the over-fitting model G. Let s_T and s_G denote the number of parameters within the sub-matrix $C^{*,e}$ according to graphs G_T and G, respectively. Let $D_G = (I_{s_T}, 0_{s_T, s_G - s_T})$, with I_{s_T} an identity matrix of dimension $s_T \times s_T$, and $0_{s_T, s_G - s_T}$ denoting a matrix of zeros with dimension $s_T \times (s_G - s_T)$. Let $M_{G/T}$ denote the difference matrix $(H_G(C_0^{*,e})^{-1} - D_G' H_T^{-1}(C_0^{*,e}) D_G)$. let $\lambda_{G[1]}, \ldots, \lambda_{G[m]}$ be the nonzero eigenvalues of $H_G(C_0^{*,e})^{1/2} M_{G/T}(H_G(C_0^{*,e})^{1/2}$, a projection matrix (Shao (2003)) with $\sum_{j=1}^{m} \lambda_{G[j]} = m$. Let

$$Q_{G/T} = U_G(C_0^{*,e})' M_{G/T} \mathbf{U}_G(C_0^{*,e}).$$

Lemma 10. *Let D_G, $\mathcal{G}_+(m)$, $M_{G/T}$, and $Q_{G/T}$ be as above. Then*

$$P(\sup_{G \in \mathcal{G}_+(m)} Q_{G/T} \geq 4m \log p_n) = o(1).$$

Lemma 11. *Under the regularity conditions,*

$$\sup_{G \in \mathcal{G}_+} |2\{\ell(\hat{C}_G^{*,e}) - \ell(\hat{C}_{G_T}^{*,e})\} - Q_{G/T}| = O_p(1). \tag{2.6}$$

Theorem 7. *For all λ such that $\lambda \in [0, \lambda_{\max}]$, and $G_\lambda \in \mathcal{G}_+$,*

$$P\big\{\inf_{G_\lambda \in \mathcal{G}_+} BIC_{G_\lambda} > BIC_{\lambda_n}\big\} \to 1, \text{ as } n \to \infty.$$

Theorem 6 and Theorem 7 imply that the λ's that fail to identify the true model have BIC larger than λ_n. Consequently, the λ value that minimizes the modified BIC criterion identifies the true model.

Theorem 8. *Under the regularity conditions, $Pr(G_{\hat{\lambda}_{BIC}} = G_T) \to 1$, where $\hat{\lambda}_{BIC}$, which may not be unique, is the tuning parameter that minimizes the modified BIC criterion with the SCAD penalty.*

3. Simulation Studies

Next we report on simulation studies to investigate the performance of BIC in penalized likelihood estimation of Gaussian graphical model. We compare its empirical performance with that of cross validation, which is another commonly used tuning parameter selection method. The K-fold cross-validation method partitions all the samples into K disjoint subsets with indices of subjects in k-fold as T_k, $k = 1, \ldots, K$. The K-fold cross-validation score is:

$$\mathrm{CV}(\lambda) = \sum_{k=1}^{K} n_k(-\log|\hat{C}_{\lambda,-k}| + \mathrm{tr}(\hat{C}_{\lambda,-k} S_k)),$$

where n_k is the size of the subset T_k, $\hat{C}_{\lambda,-k}$ is the estimated concentration matrix based on the sample $\cup_{j\neq k}T_j$, and S_k is the sample covariance matrix calculated on subset T_k. The optimum tuning parameter λ is selected to minimize CV. In our simulation, K was 5.

First we consider that p is fixed. We simulated two graphical model structures.

- Model 1. An AR(1) model with $p = 35$, $c_{ii} = 2$, and $c_{i,i-1} = c_{i-1,i} = 1$.
- Model 2. A full subset model with $p = 35$, $C_{ii} = 1.2$, and $C_{ij} = 1$ for all $1 \leq i \neq j \leq 10$, $C_{ij} = 0$, otherwise. The subset with vertices $1 \leq i \leq j \leq 10$ are all connected with edges.

Second we consider that p is large but all the edges are included in a bounded subset. The modified BIC is applied on the data sets. We simulated two graphical model structures.

- Model 1. An AR(1) model with $p = 250$, $c_{ii} = 3$, and $c_{i,i-1} = c_{i-1,i} = 0.8$, for $1 \leq i \leq 100$.
- Model 2. A full subset model with $p = 250$, $C_{ii} = 1.2$, and $C_{ij} = 1$ for all $1 \leq i \neq j \leq 20$, $C_{ij} = 0$, otherwise. The subset with vertices $1 \leq i \leq j \leq 20$ are all connected with edges.

For all the settings, we took sample size $N = 500$. All results are averaged from 100 simulated data set. For each model, we used penalized likelihood methods with SCAD and LASSO penalties. The tuning parameters for both penalties were selected through either the BIC criterion or the cross-validation criterion. To assess model selection performance, we evaluated the sensitivity, specificity, and Matthews correlation coefficient (MCC), fdr, and psr defined as follows:

$$\text{specificity} = \frac{\text{TN}}{TN + FP}, \ \text{sensitivity} = \frac{\text{TP}}{TP + FN},$$
$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$
$$\text{fdr} = \frac{\text{FP}}{TP + FP}, \ \text{psr} = \frac{\text{TP}}{TP + FN},$$

where TP, TN, FP, FN are the numbers of true positives, true negatives, false positives, and false negatives. Taking both true and false positives and negatives into account, MCC has been widely used to measure the quality of binary classifiers. Means and standard deviations of the above measures are provided in Tables 1-2.

Table 1. Results for Graphical Model with p=35 and N=500.

	LASSO				*SCAD*			
	model 1		model 2		model 1		model 2	
	bic	cv	bic	cv	bic	cv	bic	cv
fp	0.25	87.02	7.26	350.73	0.53	42.59	3.35	240.62
	(0.50)	(19.25)	(3.96)	(35.10)	(0.73)	(7.73)	(2.56)	(12.77)
fn	32.70	4.06	0.20	0.00	31.23	8.98	0.92	0.00
	(4.08)	(2.17)	(0.53)	(0.00)	(5.00)	(2.32)	(1.38)	(0.00)
tp	12.30	40.94	44.80	45.00	31.27	53.52	61.58	62.50
	(4.08)	(2.17)	(0.53)	(0.00)	(5.00)	(2.32)	(1.38)	(0.00)
tn	549.75	462.98	542.74	199.27	549.47	507.41	546.65	309.38
	(0.50)	(19.25)	(3.96)	(35.10)	(0.73)	(7.73)	(2.56)	(12.77)
spec	1.00	0.84	0.99	0.36	1.00	0.92	0.99	0.56
	(0.00)	(0.03)	(0.01)	(0.06)	(0.00)	(0.01)	(0.00)	(0.02)
sens	0.27	0.91	1.00	1.00	0.50	0.86	0.99	1.00
	(0.09)	(0.05)	(0.01)	(0.00)	(0.08)	(0.04)	(0.02)	(0.00)
mcc	0.50	0.49	0.92	0.20	0.68	0.65	0.96	0.34
	(0.08)	(0.04)	(0.04)	(0.03)	(0.05)	(0.03)	(0.02)	(0.01)
fdr	0.01	0.67	0.13	0.89	0.01	0.44	0.05	0.79
	(0.03)	(0.04)	(0.06)	(0.01)	(0.02)	(0.04)	(0.04)	(0.01)
psr	0.27	0.91	1.00	1.00	0.50	0.86	0.99	1.00
	(0.09)	(0.05)	(0.01)	(0.00)	(0.08)	(0.04)	(0.02)	(0.00)

SCAD:the SCAD penalty; LASSO: the L_1 penalty. Averages and standard errors are obtained from 100 simulated data sets. Averages are provided without parenthesis and standard errors are provided within parentheses.

Implementation was based on the GLASSO algorithm of Friedman, Hastie, and Tibshirani (2008) and the reiterative weighted GLASSO of Fan, Feng, and Wu (2009) We examined the empirical performance of the two penalty functions under the selection of optimal tuning parameter via BIC or cross-validation. Tables 1 and 2 provide the average specificity, sensitivity, Matthew's correlation coefficient, false discovery rate, and positive selection rate over 100 simulated data sets. Standard errors are provided in the parenthesis.

Across all different settings, the SCAD penalty consistently yielded better performance than the LASSO penalty. With p fixed, for AR(1) the average MCC for SCAD was 0.68 versus 0.50 for LASSO; for the full subset model, the average MCC for SCAD was 0.96 versus 0.92 for LASSO. With p large and subset dimension bounded, for AR(1) the average MCC for SCAD was 0.99 versus 0.93 for LASSO; for the full subset model, the average MCC for SCAD was 0.78 versus 0.61 for LASSO.

Another comparison was conducted between BIC and cross validation. It is noted that cross validation consistently gave higher fdr, and lower MCC than did BIC in the settings investigated in our study. For instance, when $p = 35$,

Table 2. Results for Graphical Model with $p = 250$ and $N = 500$.

	LASSO				*SCAD*			
	model 1		model 2		model 1		model 2	
	bic	cv	bic	cv	bic	cv	bic	cv
fp	16.30	566.24	10.22	952.57	2.94	1981.43	7.99	90.46
	(5.92)	(28.12)	(5.40)	(68.63)	(1.72)	(47.17)	(4.22)	(13.37)
fn	0.12	0.00	110.01	23.48	0.36	0.00	112.80	76.93
	(0.38)	(0.00)	(8.49)	(3.95)	(0.72)	(0.00)	(7.58)	(4.82)
tp	98.88	99.00	79.99	166.52	223.64	224.00	202.20	238.07
	(0.38)	(0.00)	(8.49)	(3.95)	(0.72)	(0.00)	(7.58)	(4.82)
tn	31010	30460	30925	29982	31023	29045	30927	30845
	(5.92)	(28.12)	(5.40)	(68.63)	(1.72)	(47.17)	(4.22)	(13.37)
spec	1.00	0.98	1.00	0.97	1.00	0.94	1.00	1.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
sens	1.00	1.00	0.42	0.88	1.00	1.00	0.64	0.76
	(0.00)	(0.00)	(0.04)	(0.02)	(0.00)	(0.00)	(0.02)	(0.02)
mcc	0.93	0.38	0.61	0.35	0.99	0.31	0.78	0.74
	(0.02)	(0.01)	(0.03)	(0.01)	(0.00)	(0.00)	(0.01)	(0.02)
fdr	0.14	0.85	0.11	0.85	0.01	0.90	0.04	0.27
	(0.04)	(0.01)	(0.05)	(0.01)	(0.01)	(0.00)	(0.02)	(0.03)
psr	1.00	1.00	0.42	0.88	1.00	1.00	0.64	0.76
	(0.00)	(0.00)	(0.04)	(0.02)	(0.00)	(0.00)	(0.02)	(0.02)

SCAD:the SCAD penalty; LASSO: the L_1 penalty. Averages and standard errors are obtained from 100 simulated data sets. Averages are provided without parenthesis and standard errors are provided within parentheses.

in AR(1) model, when SCAD penalty was used, the fdr rate of cross validation was as high as 0.44 versus a low fdr rate of 0.01 for BIC. The MCC of cv was 0.65 versus a higher MCC of 0.68 of BIC. For full subset model, the fdr rate of cross validation was as high as 0.79 versus a low fdr rate of 0.05 for BIC. The MCC of cv was as low as 0.34 versus a higher MCC of 0.96 of BIC. With $p = 250$ in the AR(1) model, the fdr rate of cross validation was as high as 0.90 versus a low fdr rate of 0.01 for BIC. The MCC of cv was 0.31 versus a higher MCC of 0.99 of BIC. For full subset model,the fdr rate of cross validation was 0.27 versus a low fdr rate of 0.04 for BIC. The MCC of cv was 0.74 versus a higher MCC of 0.78 of BIC. Overall, when the graph was sparse with the set of all edges bounded, cross validation method had a higher false discovery rate than did BIC. Although BIC enjoyed a better control of false discovery rate, its positive selection rate or sensitivity was lower than that of cross validation. Taking sensitivity and selectivity into consideration, BIC enjoyed a better performance as reflected by the higher Matthew correlation coefficient. Computationally, BIC is more convenient to use than cross validation.

For the high-dimensional inverse covariance matrix case, the modified BIC with SCAD penalty retained satisfactory performance. For example, for the AR(1) model with $p = 250$ and the number of true edges $d_T = 224$, the modified BIC with SCAD had a fdr of 0.01, a psr of 1.00, and a MCC of 0.99. For the full subset model with $p = 250$ and the number of true edges $d_T = 315$, the modified BIC with SCAD had a fdr of 0.04, a psr of 0.64 and a MCC of 0.78. This empirical performance supports the asymptotic consistency result of the modified BIC with SCAD when p_n is large and the number of true edges is bounded.

4. Conclusion

We have investigated tuning parameter selection for penalized likelihood estimation of the inverse covariance matrix. We establish the consistency of the BIC criterion for selecting the true graphical model using the SCAD penalty, when p is fixed. A modified BIC with an extra penalty on the dimension of the inverse covariance matrix is shown to be selection consistent when p tends to infinity with the sample sizes and the number of true edges is bounded.

Acknowledgement

This research is supported by the Canadian National Science and Engineering Research Council grants to Gao and Wu.

Appendix

In proofs, the notation $\hat{C} - C$ stands for a column vector stacked from the difference matrix between $\hat{C}$ and C.

Proof of Lemma 1. According to Theorem 5.2 in Fan, Feng, and Wu (2009), under the reference sequence of tuning parameters, we have

$$\lim_{n\to\infty} P\big(\sum_{i<j} I(\hat{c}_{ij,\lambda_n} \neq 0) = \sum_{i<j} I(c_{ij,0} \neq 0)\big) = 1.$$

Furthermore, both $\hat{C}_{\lambda_n}$ and $\hat{C}_{G_T}$ are root-n consistent for C_0, the null value. This entails $|\ell(\hat{C}_{\lambda_n}) - \ell(\hat{C}_{G_T})| = O_p(1)$. The result follows.

Proof of Lemma 2. Denote the score vector $\partial\ell/\partial C$ by U_n. For any value of C such that $||C - C_0||_2 \leq n^{-1/3}$, we have

$$\ell(C) - \ell(C_0) = (C - C_0)U_n(C_0) - \frac{1}{2}(C - C_0)' \frac{\partial^2 \ell}{\partial C \partial C}(\tilde{C})(C - C_0),$$

for some $\tilde{C}$ between C_0 and C. Let $\otimes$ denote the Kronecker product of two matrices. As

$$\frac{\partial^2 \ell}{\partial C \partial C} = (C)^{-1} \otimes (C)^{-1},$$

for any $\epsilon > 0$, there exists a constant $\delta > 0$, such that when n is large enough,

$$(1-\epsilon)||\frac{\partial^2 \ell}{\partial C \partial C}(C_0)|| \leq ||\frac{\partial^2 \ell}{\partial C \partial C}(\tilde{C})|| \leq (1+\epsilon)||\frac{\partial^2 \ell}{\partial C \partial C}(C_0)||$$

for all $||\tilde{C} - C_0|| \leq \delta$. This is due to fact that the eigenvalues of C are bounded from 0 and ∞, and the matrix inverse C^{-1} is continuous in C. Therefore,

$$(C - C_0)' \frac{\partial^2 \ell}{\partial C \partial C}(\tilde{C})(C - C_0) \geq nM(1-\epsilon)||C - C_0||_2^2$$

for some constant M. Thus for any $||C - C_0||_2 = n^{-1/3}$, we have

$$\ell(C) - \ell(C_0) \leq n^{-1/3}||U_n(C_0)|| - \frac{M}{2} n^{1/3}(1-\epsilon).$$

Because $||U_n(C_0)|| = O_p(n^{1/2})$, we have

$$\ell(C) - \ell(C_0) \leq n^{1/6} - M\frac{n^{1/3}}{2} \leq -L_1 n^{1/3}$$

for all C such that $||C - C_0||_2 = n^{-1/3}$ for a constant L_1, with probability tending to 1.

Because $\ell(C)$ is concave in C, the above result implies that the maximum of $\ell(C)$ is attained inside $||C - C_0||_2 < n^{-1/3}$. Concavity also implies that

$$\begin{aligned} &\sup\{\ell(C) - \ell(C_0) : ||C - C_0||_2 > n^{-1/3}\} \\ &\quad \leq \sup\{\ell(C) - \ell(C_0) : ||C - C_0||_2 = n^{-1/3}\} \leq -L_1 n^{1/3}. \end{aligned} \tag{A.1}$$

Because $G_\lambda \in \mathcal{G}_-$, there is at least one edge (i,j) that $\hat{C}_{ij,\lambda} = 0$, while $C_{ij,0} \neq 0$. For this non-zero $|C_{ij,0}|$, there is a lower bound depending on neither n or C. Thus $||\hat{C}_\lambda - C_0||_2 \geq |C_{ij,0}| > n^{-1/3}$, and this leads to

$$\ell(\hat{C}_\lambda) - \ell(C_0) \leq -L_1 n^{1/3}$$

with probability tending to 1 uniformly for $G_\lambda \in \mathcal{G}_-$.

Furthermore we have

$$\begin{aligned} &\ell(\hat{C}_\lambda) - \ell(\hat{C}_{G_T}) \\ &\quad = \ell(\hat{C}_\lambda) - \ell(C_0) + \ell(C_0) - \ell(\hat{C}_{G_T}) \\ &\quad = \ell(\hat{C}_\lambda) - \ell(C_0) + O_p(1). \end{aligned} \tag{A.2}$$

In view of Lemma 1, the result follows.

Proof of Theorem 3. Let the maximum likelihood estimator under the true model and under the over-fitted model be $\hat{C}_{G_T}$, and $\hat{C}_{G_\lambda}$. Note that $\hat{C}_{G_\lambda}$ is

different from $\hat{C}_\lambda$. According to standard asymptotic theory for the loglikelihood ratio statistic, we have $2(\ell(\hat{C}_{G_\lambda}) - \ell(\hat{C}_{G_T})) \sim \chi^2_{d_{G_\lambda}-d_T} = O_p(1)$. Furthermore, from Theorem 5.2 in Fan, Feng, and Wu (2009), $|\hat{C}_\lambda - \hat{C}_{G_\lambda}| = O_p(n^{-\frac{1}{2}})$. This implies $\ell(\hat{C}_\lambda) = \ell(\hat{C}_{G_\lambda}) + O_p(1)$. Combining, we have

$$\begin{aligned}(BIC_\lambda - BIC_{G_T}) &= -2\ell(\hat{C}_\lambda) + 2\ell(\hat{C}_{G_T}) + \log n(d_{G_\lambda} - d_T) \\ &= -2\ell(\hat{C}_{G_\lambda}) + 2\ell(\hat{C}_{G_T}) + \log n(d_{G_\lambda} - d_T) + O_p(1) \\ &= \log n(d_{G_\lambda} - d_T) + O_p(1).\end{aligned} \tag{A.3}$$

In view of Lemma 1, the result follows.

Proof of Lemma 3. It suffices to show that $|\ell(\hat{C}^{*,T}_{\lambda_n}) - \ell(\hat{C}^{*,T}_{G_T})| = O_p(1)$. Focusing on the sub-matrix $C^{*,T}$, as $d_T \le Q$, the problem reduces to a problem of estimation with finite dimensions. The working sequence λ_n satisfies the condition that $\lambda_n \to 0$, and $\sqrt{n}\lambda_n \to \infty$. By Theorem 5.2 of Fan, Feng, and Wu (2009), $\hat{C}^{*,T}_{\lambda_n}$ is a $\sqrt{n}-$consistent estimator of $C^{*,T}_0$. By mle's property, $\hat{C}^{*,T}_{G_T}$ is also an $\sqrt{n}-$consistent estimator of $C^{*,T}_0$. Thus $|\hat{C}^{*,T}_{\lambda_n} - \hat{C}^{*,T}_{G_T}| = O_p(n^{-1/2})$. Because $\partial^2\ell(C^{*,T})/(\partial C^{*,T}\partial C^{*,T}) = (C^{*,T})^{-1} \otimes (C^{*,T})^{-1}$, the eigenvalues of $C^{*,T}$ are uniformly bounded for all n and, since matrix inverse is continuous over non-singular matrix, we have that for n large enough, $\partial\ell^2(C^{*,T})/\partial C^{*,T}\partial C^{*,T}\big|_{\tilde{C}^{*,T}} = O_p(n)$ for any $\tilde{C}^{*,T}$ between $\hat{C}^{*,T}_{\lambda_n}$ and $\hat{C}^{*,T}_{G_T}$. This leads to $|\ell(\hat{C}^{*,T}_{\lambda_n}) - \ell(\hat{C}^{*,T}_{G_T})| = O_p(1)$.

Proof of Lemma 4. By Taylor expansion, for $|t| \le \delta$, the cumulant generating function

$$g(t) = \frac{t^2}{2} + g^{(3)}(t^*)\frac{t^3}{6},$$

for some $0 \le |t^*| \le |t| \le \delta$. For any $|t|/\sqrt{n} \le \delta$, the moment generating function of $n^{-1/2}\sum_{i=1}^n Z_i$ is equal to

$$\phi_n(t) = \exp\{\frac{t^2}{2} + \frac{g^{(3)}(t^*/\sqrt{n})t^3}{6\sqrt{n}}\}.$$

For convenience, let $q_n = \sqrt{2m\log f_n}$. It can be shown that

$$I(n^{-1/2}\sum_{i=1}^n Z_i > q_n) \le \exp\{t[n^{-1/2}\sum_{i=1}^n Z_i - q_n]\},$$

for any $t > 0$. Then

$$\begin{aligned} P(n^{-1/2}\sum_{i=1}^{n} Z_i > q_n) &\le E[\exp\{t[n^{-1/2}\sum_{i=1}^{n} Z_i - q_n]\}] \\ &= \exp\{\frac{t^2}{2} + \frac{g^{(3)}(t^*/\sqrt{n})t^3}{6\sqrt{n}} - q_n t\} \\ &= \exp\{\frac{t^2}{2}(1+o(1)) - q_n t\}. \end{aligned}$$

With $t = q_n$, we have

$$P(\sum_{i=1}^{n} Z_i > \sqrt{2mn\log f_n}) \le \exp\{-\frac{1}{2}q_n^2(1+o(1))\} = o(f_n^{-m}).$$

Proof of Lemma 5. It can be shown that $\mathrm{Var}\,(U_n(C_0^{*,e})/n) = (C^{*,e})^{-1} \otimes (C^{*,e})^{-1}$. As

$$0 < \tau_1 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le \tau_2 < \infty,$$

we have

$$0 < \frac{1}{\tau_2} \le \lambda_{\min}(C) \le \lambda_{\max}(C) \le \frac{1}{\tau_1} < \infty.$$

Because $C^{*,e}$ is a sub-matrix of C with $C^{'},^{e}$ and $C^{''},^{e}$ being 0, and $C^{'''},^{e}$ being diagonal, the eigenvalues of $C^{*,e}$ are a subset of that of C's. Therefore,

$$0 < \tau_1^2 \le \lambda_{\min}(\mathrm{Var}\,(U_n\frac{C_0^{*,e}}{n})) \le \lambda_{\max}(\mathrm{Var}\,(U_n\frac{C_0^{*,e}}{n})) \le \tau_2^2 < \infty,$$

uniformly for all the $G \in \mathcal{G}$. Let $U_n(C_0^*)_{[ij]}$ denote the element of the score vector corresponding to differentiation with respect to the parameter c_{ij}. We have

$$\begin{aligned} U_n(C_0^*)_{[ij]} &= n\sigma_{ij} - \sum_{l=1}^{n}(X_l^{(i)} - \overline{X}^{(i)})(X_l^{(j)} - \overline{X}^{(j)}) \\ &= \sum_{l=1}^{n}(\sigma_{ij} - X_l^{(i)}X_l^{(j)}) + n\overline{X}^{(i)}\overline{X}^{(j)}. \end{aligned} \tag{A.4}$$

Because $X_l^{(i)}$ is normal, by Lemma 4 we have $\max_i |\overline{X}^{(i)}| = O_p(n^{-1/2}\log p_n^{1/2})$. It then suffices to show that $\max_{(i,j)} |\sum_{l=1}^{n}(\sigma_{ij} - X_l^{(i)}X_l^{(j)})| = O_p(n^{1/2}(\log p_n)^{1/2})$.

Let

$$Z_l = \frac{\sigma_{ij} - X_l^{(i)}X_l^{(j)}}{\sqrt{\mathrm{Var}\,(X_l^{(i)}X_l^{(j)})}}.$$

As $\max_{(i,j)} \mathrm{Var}\,(X_l^{(i)} X_l^{(j)})$ is uniformly bounded away from zero and infinity, it can be shown that $|\sum_{l=1}^n (\sigma_{ij} - X_l^{(i)} X_l^{(j)})| \le M|\sum_{l=1}^n Z_l|$, uniformly for $G \in \mathcal{G}$ for some constant M.

To apply Lemma 4 to the sum of Z_ls, we need to check the uniform boundedness of the third derivative of the cumulant generating function over the model space.

By Lancaster (1954), the cumulant generating function $g(t)$ for $X_l^{(i)} X_l^{(j)}$ is $-1/2 \log|I - 2\Sigma^{*,e} K t|$, where $K = k_i k_j'$, and k_i is a vector of zeros except the ith entry is 1. The third derivative $g^{(3)}(t) = 8\mathrm{tr}[\{\Sigma^{*,e} K (I - 2\Sigma^{*,e} K t)^{-1}\}^3]$. Let $J = (I - 2\Sigma^{*,e} K t)$, then $g^{(3)}(t) = (k_j' J \Sigma^{*,e} k_i)^3 = \{(J\Sigma^{*,e})_{[ij]}\}^3$. By choosing δ small enough, we can make the $|(J\Sigma^{*,e})_{[ij]}|$ uniformly bounded for $t \le \delta$, and all (i,j). According to Lemma 4, we have

$$P(|U_n(C_0^{*,e})_{[ij]}| > \sqrt{2mn \log p_n}) = o(p_n^{-m}),$$

and hence

$$P(||U_n(C_0^{*,e})|| > \sqrt{2mn \log p_n}) = o(\frac{m(m+1)}{2} p_n^{-m}),$$

where m is the dimension of $C^{*,e}$. There are in total at most p_n^m possible $C^{*,e}$ induced by $G \in \mathcal{G}$. As m is bounded, according to Bonferroni's inequality we have

$$\max_{G \in \mathcal{G}} P(||U_n(C_0^{*,e})|| > \sqrt{2mn \log p_n}) = o(\frac{m(m+1)}{2} p_n^{-m}) \times p_n^m = o(1).$$

Proof of Lemma 6. For any $C^{*,e}$ such that $||C^{*,e} - C_0^{*,e}|| \le n^{-1/3}$, we have

$$\ell(C^{*,e}) - \ell(C_0^{*,e}) = (C^{*,e} - C_0^{*,e}) U_n(C_0^{*,e}) - \frac{1}{2}(C^{*,e} - C_0^{*,e})' \frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}}(\tilde{C}^{*,e})(C^{*,e} - C_0^{*,e})$$

for some $\tilde{C}^{*,e}$ between $C_0^{*,e}$ and $C^{*,e}$. Since

$$\frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}} = (C^{*,e})^{-1} \otimes (C^{*,e})^{-1},$$

we have

$$0 < \tau_1^2 \le \lambda_{\min}(n^{-1} \frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}}(C_0^{*,e})) \le \lambda_{\max}(n^{-1} \frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}}(C_0^{*,e})) \le \tau_2^2 < \infty.$$

For any $\epsilon > 0$, there exist a constant $\delta > 0$, such that when n is large enough,

$$(1 - \epsilon)||\frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}}(C_0^{*,e})|| \le ||\frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}}(\tilde{C}^*)|| \le (1 + \epsilon)||\frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}}(C_0^{*,e})||,$$

for all $C^{*,e}$ induced by all $G \in \mathcal{G}$ and all $||\tilde{C}^{*,e} - C_0^{*,e}|| \leq \delta$. This is due to fact that all the eigenvalues of C are bounded from 0 and ∞, the dimension of $C^{*,e}$ is bounded, and matrix inverse is continuous. Therefore,

$$(C^{*,e} - C_0^{*,e})' \frac{\partial^2 \ell}{\partial C^{*,e} \partial C^{*,e}} (\tilde{C}^{*,e})(C^{*,e} - C_0^{*,e}) \geq Mn(1-\epsilon)||C^{*,e} - C_0^{*,e}||_2^2$$

for some constant M. Thus for any $||C^{*,e} - C_0^{*,e}||_2 = n^{-1/3}$, we have

$$\ell(C^{*,e}) - \ell(C_0^{*,e}) \leq n^{-1/3}||U_n(C_0^{*,e})|| - \frac{M}{2} n^{1/3}(1-\epsilon).$$

Because $\max_{G \in \mathcal{G}} ||U_n(C_0^{*,e})|| = O_p(n^{1/2}(\log p_n)^{1/2})$, we have

$$\ell(C^{*,e}) - \ell(C_0^{*,e}) \leq n^{1/6}(\log p_n)^{1/2} - M(1-\epsilon)\frac{n^{1/3}}{2} \leq -L_3 n^{1/3},$$

uniformly over $G \in \mathcal{G}$ for a generic constant L_3.

Because $\ell(C^{*,e})$ is concave in $C^{*,e}$, the above result implies that the maximum of $\ell(C^{*,e})$ is attained inside $||C^{*,e} - C_0^{*,e}|| < n^{-1/3}$. The concavity also implies that, uniformly over $G \in \mathcal{G}$ with probability tending to one,

$$\begin{aligned} &\sup\{\ell(C^{*,e}) - \ell(C_0^{*,e}) : ||C^{*,e} - C_0^{*,e}|| > n^{-1/3}\} \\ &\quad \leq \sup\{\ell(C^{*,e}) - \ell(C_0^{*,e}) : ||C^{*,e} - C_0^{*,e}|| = n^{-1/3}\} \leq -L_3 n^{1/3}. \end{aligned} \tag{A.5}$$

Proof of Lemma 7. Given a λ with $G_\lambda \in \mathcal{G}-$, there is at least one edge (i,j) with $\hat{C}_{ij,\lambda}^{*,e} = 0$, while $C_{ij,0}^{*,e} \neq 0$. Thus $||\hat{C}_\lambda^{*,e} - C_0^{*,e}||_2 \geq |C_{ij,0}^{*,e}| > n^{-1/3}$, because $|C_{ij,0}^{*,e}|$ has a lower bound not depending on n. According to Lemma 6,

$$\ell(\hat{C}_\lambda^{*,e}) - \ell(C_0^{*,e}) \leq -L_4 n^{1/3}$$

with probability tending to 1, uniformly for all λ such that $G_\lambda \in \mathcal{G} - .$

Proof of Lemma 8. Let $C^{*,T}$ be induced by G_T. Then $C^{*,T}$ is a submatrix of $C^{*,e}$. Let $A = C^{*,e}/C^{*,T}$ represent the complement of $C^{*,T}$ within $C^{*,e}$. On A the estimates of $\hat{C}_{G_T}$ are either zero for off-diagonal entries or the inverse of sample standard deviation for diagonal entries. According to Theorem 5.10 (Bai and Silverstein (2006)), $\sup_{G_\lambda \in \mathcal{G}-} |\ell(\hat{A}_{G_T}) - \ell(A_0)| = O_p(1)$. Then

$$\begin{aligned} &\sup_{G_\lambda \in \mathcal{G}-} |\ell(\hat{C}_{G_T}^{*,e}) - \ell(C_0^{*,e})| \\ &\quad \leq \sup_{G_\lambda \in \mathcal{G}-} |\ell(\hat{A}_{G_T}) - \ell(A_0)| + |\ell(\hat{C}_{G_T}^{*,T}) - \ell(C_0^{*,T})| \\ &\quad = O_p(1). \end{aligned} \tag{A.6}$$

Proof of Theorem 5. Combining results from Lemmas 7 and 8, We have

$$\begin{aligned}&\ell(\hat{C}^{*,e}_{\lambda}) - \ell(\hat{C}^{*,e}_{G_T})\\ &\quad= \ell(\hat{C}^{*,e}_{\lambda}) - \ell(C^{*,e}_0) + \ell(C^{*,e}_0) - \ell(\hat{C}^{*,e}_{G_T})\\ &\quad\le -Mn^{1/3}\end{aligned} \tag{A.7}$$

for some generic constant M, with probability tending to 1 uniformly for $G_\lambda \in \mathcal{G} -$. Because the difference in penalty terms is of order $\log n$, the result of the theorem follows.

Proof of Lemma 9. Let $\ell^{(1)}_r$ be the first derivative of the log-likelihood with respect to the rth parameter, and $\ell^{(2)}_{rt}$ be the second partial derivative of the log-likelihood with respect to the rth and tth parameter. A Taylor expansion of $\ell^{(1)}_r(\hat{C}^{*,e}_G) = 0$ around $C^{*,e}_0$ gives the system of equations:

$$\begin{aligned}0 = \ell^{(1)}_r(\hat{C}^{*,e}_G) = \ell^{(1)}_r(C^{*,e}_0) &+ \sum_t \ell^{(2)}_{rt}(C^{*,e}_0)(\hat{C}^{*,e}_G - C^{*,e}_0)_{[t]}\\ &+ \sum_{tu} \frac{1}{2}\ell^{(3)}_{rtu}(\tilde{C}^{*,e})(\hat{C}^{*,e}_G - C^{*,e}_0)_{[t]}(\hat{C}^{*,e}_G - C^{*,e}_0)_{[u]},\end{aligned} \tag{A.8}$$

for some $\tilde{C}^{*,e}$ between $\hat{C}^{*,e}_G$ and $C^{*,e}_0$. For notational brevity, if no argument is specified in such as $\ell^{(1)}_r$ and $\ell^{(2)}_{rt}$, it is assumed to be evaluated at $C^{*,e}_0$. After taking matrix inversion on both sides of (A.8), we have

$$(C^{*,e}_0 - \hat{C}^{*,e}_G)_{[r]} = \sum_t \big\{\ell^{rt}\ell^{(1)}_t + \frac{1}{2}\ell^{rt}\sum_{uv}(\hat{C}^{*,e}_G - C^{*,e}_0)_{[u]}(\hat{C}^{*,e}_G - C^{*,e}_0)_{[v]}\ell^{(3)}_{tuv}(\tilde{C}^{*,e})\big\}, \tag{A.9}$$

with ℓ^{rt} denoting the $(r,t)th$ entry of the inverse of the matrix $\ell^{(2)} = (\ell^{(2)}_{rt})$. Let $M_{rt} = E(\ell^{(2)}_{rt}) = \ell^{(2)}_{rt}$ and $M^{rt} = E(\ell^{rt}) = \ell^{rt}$. That Σ has all eigenvalues bounded away from zero implies

$$\max_{G\in\mathcal{G}} |M^{rt}(C^{*,e}_0)| \le \frac{L_6}{n} \tag{A.10}$$

for some generic constant L_6. By Lemma 5,

$$\max_{G\in\mathcal{G}} |\ell^{(1)}_r| = O_p(n^{1/2}(\log p_n)^{1/2}).$$

Now we rewrite (A.9) as

$$(C^{*,e}_0 - \hat{C}^{*,e}_G)_{[r]} = \sum_t M^{rt}\ell^{(1)}_t + R_n,$$

with the error term

$$R_n = \sum_t \{\frac{1}{2}\ell^{rt} \sum_{uv} (\hat{C}_G^{*,e} - C_0^{*,e})_{[u]} (\hat{C}_G^{*,e} - C_0^{*,e})_{[v]} \ell^{(3)}_{tuv}(\tilde{C}^{*,e})\}.$$

On the other hand, in light of Lemma 6, we have

$$\lim_{n\to\infty} P(\max_{G\in\mathcal{G}} ||\hat{C}_G^{*,e} - C_0^{*,e}|| \le n^{-1/3}) \to 1.$$

As the eigenvalues of Σ are all bounded and matrix inverse is continuous, we have

$$\lim_{n\to\infty} P(\max_{G\in\mathcal{G}} ||\ell^{(3)}(\tilde{C}^{*,e})|| \le L_7 N) \to 1$$

for some constant L_7. Combining, we can show that

$$\max_{G\in\mathcal{G}} ||\hat{C}_G^{*,e} - C_0^{*,e} + H_G(C_0^{*,e})^{-1} U_G(C_0^{*,e})|| = O_p(n^{-2/3}). \tag{A.11}$$

Next, Taylor expansion for the log-likelihood leads to

$$\begin{aligned}&\ell(\hat{C}_G^{*,e} - \ell(C_0^{*,e}) \\ &= U_G(C_0^{*,e})'(\hat{C}_G^{*,e} - C_0^{*,e}) + \frac{1}{2}\sum_{rt} (\hat{C}_G^{*,e} - C_0^{*,e})_{[r]} (\hat{C}_G^{*,e} - C_0^{*,e})_{[t]} M_{rt} + \tilde{R}_n,\end{aligned} \tag{A.12}$$

where the error term is

$$\tilde{R}_n = \frac{1}{6}\sum_{rtu} (\hat{C}_G^{*,e} - C_0^{*,e})_{[r]} (\hat{C}_G^{*,e} - C_0^{*,e})_{[t]} (\hat{C}_G^{*,e} - C_0^{*,e})_{[u]} \ell^{(3)}_{rtu}(\tilde{C}^{*,e}),$$

with $\max_{G\in\mathcal{G}} |\tilde{R}_n| = O_p(1)$ based on the results above. This implies

$$\max_{G\in\mathcal{G}} |2\{\ell(\hat{C}_G^{*,e}) - \ell(C_0^{*,e})\} - U_G(C_0^{*,e})' H_G(C_0^{*,e})^{-1} U_G(C_0^{*,e})| = O_p(1).$$

Proof of Theorem 10. Let $Q_{G/T} = z'_{G/T} z_{G/T}$, where $z_{G/T} = M_{G/T}^{1/2} U_G(C_0^{*,e})$. Let v be any unit vector of length s_G. Then

$$\begin{aligned}\frac{\sqrt{n} v' z_{G/T}}{\mathrm{Var}\,(v' z_{G/T})} &= \frac{\sum_r \{a_{G,r} \sum_{i=1}^n U_G(C_0^{*,e}, Y_i^{*,e})_r\}}{\mathrm{Var}\,(v' z_{G/T})} \\ &= \sum_{i=1}^n Y_{G,i},\end{aligned}$$

with $U_G(C_0^{*,e}, X_i^{*,e})$ denoting the score vector on data point $X_i^{*,e}$, $a_{G,r}$ denoting the rth element in the vector $a_G = v'(nM_{G/T})^{1/2}$, and

$$Y_{G,i} = \sum_r \frac{\{a_{G,r} U_G(C_0^{*,e}, X_i^{*,e})_r\}}{\mathrm{Var}\,(v' z_{G/T})}.$$

According to the assumptions, $\sup_{G\in\mathcal{G}_+(m)}||a_G||$, $\sup_{G\in\mathcal{G}_+(m)}\mathrm{Var}\,(U_G(C_0^{*,e},X_i^{*,e}))$ and $\sup_{G\in\mathcal{G}_+(m)}\{\mathrm{Var}\,(v'z_{G/T})\}^{-1}$ are all bounded. This entails that the third derivatives of the cumulant generating function $g(t)$ of $Y_{g,i}$ is bounded, i.e., $|g^{(3)}(t)|\le M$ for some constant M for all $G\in\mathcal{G}_+(m)$, and $0\le|t|\le\delta$. Furthermore, the variables $Y_{G,i}$, $i=1,\ldots,n$ are independent and identically distributed with zero mean and unit variance. Thus according to Lemma 4,

$$P\Big(\frac{\sqrt{n}v'z_{G/T}}{\mathrm{Var}\,(v'z_{G/T})}\ge\sqrt{2mn\log(P_n^2)}\Big)=o(p_n^{-2m}).$$

On the other hand, given any finite set of unit vectors $\mathcal{V}$, it can be shown that $\max_{v\in\mathcal{V}}\mathrm{Var}\,(v'z_{G/T})\le 1$. This is because $H_G^{1/2}M_{G/T}H_G^{1/2}$ is a projection matrix of rank m, and then $M_{G/T}^{1/2}H_G^{1/2}$ can be represented by $A\Gamma A'$, where Γ is a diagonal matrix of m nonzero diagonal entries and A is an ortho-normal matrix. Then $z_{G/t}$ can be represented as $\mathcal{A}\Gamma A'H_G^{-1/2}U_G(C_0^{*,e})$. Thus $\mathrm{Var}\,(v'z_{G/T})=v'\,\mathrm{Cov}\,(z_{G,t})v=v'A\Gamma A'H_G^{-1/2}H_GH_G^{-1/2}A\Gamma A'v=v'A\Gamma A'v\le 1$, for any unit vector v. Therefore,

$$\begin{aligned}&P(v'z_{G/T}\ge\sqrt{4m\log p_n})\\&\quad\le P(\frac{v'z_{G/T}}{\mathrm{Var}\,(v'z_{G/T})}\ge\sqrt{4m\log p_n})=o(p_n^{-2m}).\end{aligned}$$

Combining, $P(v'z_{G/T}\ge\sqrt{4m\log p_n}=o(p_n^{-2m})$. By Lemma 2 in Chen and Chen (2012),

$$P(z_{G/T}'z_{G/T}\ge 2l\log n)\le\sum_{v\in\mathcal{V}}P(v'z_{G/T}\ge\sqrt{2l\log n})$$

for any constant $l>0$, where $\mathcal{V}$ is a finite set of unit vectors independent of n. Therefore, $P(z_{G/T}'z_{G/T}\ge 4m\log p_n)=o(p_n^{-2m})$. As there are p_n^{2m} different over-fitting model with dimension $s_G=m+s_T$, by the Bonferroni inequality,

$$P(\sup_{G\in\mathcal{G}_+(m)}z_{G/T}'z_{G/T}\ge 4m\log p_n)=o(1).$$

Proof of Lemma 11. From Lemma 10, we have

$$\sup_{G\in\mathcal{G}+}|2\{\ell(\hat{C}_G^{*,e})-\ell(C_0^{*,e})\}-U_G(C_0^{*,e})'H_G(C_0^{*,e})^{-1}U_G(C_0^{*,e})|=O_p(1).$$

Furthermore,

$$\sup_{G\in\mathcal{G}+}|2\{\ell(\hat{C}_{G_T}^{*,e})-\ell(C_0^{*,e})\}-U_T(C_0^{*,e})'H_T(C_0^{*,e})^{-1}U_T(C_0^{*,e})|=O_p(1),$$

where $C^{*,e}$ is induced by G and $\hat{C}_{G_T}^{*,e}$ denote the maximum likelihood estimate for the submatrix $C^{*,e}$. Note that $U_T(C_0^{*,e})=D_GU_G(C_0^{*,e})$, so

$$\sup_{G\in\mathcal{G}_+}|2\{\ell(\hat{C}_G^{*,e})-\ell(\hat{C}_{G_T}^{*,e})\}-Q_{G/T}|=O_p(1).\tag{A.13}$$

Proof of Theorem 7. Let

$$R_\lambda = 2\{\ell(\hat{C}^{*,e}_{G_\lambda}) - \ell(\hat{C}^{*,e}_{G_T})\} - Q_{G_\lambda/T}.$$

Given any λ such that $G_\lambda \in \mathcal{G}_+(m)$, we consider the following three events: $B_{G_\lambda} = \{Q_{G_\lambda/T} \le 4m\log p_n\}$, $J_{G_\lambda}(M) = \{|R_\lambda| \le M\}$, and $F_\lambda = \{\mathrm{BIC}_\lambda > \mathrm{BIC}_{G_T}\}$. According to Lemma 10, we have $P(\cap_{G_\lambda\in\mathcal{G}_+(m)}B_{G_\lambda}) \ge 1-o(1)$. This entails that given any $\epsilon > 0$, for n large enough, $P(\cap_{G_\lambda\in\mathcal{G}_+(m)}B_{G_\lambda}) \ge 1-\epsilon/2$. In light of Lemma 11, there exists M_ϵ and, for n large enough, $P(\cap_{G_\lambda\in\mathcal{G}_+(m)}J_{G_\lambda}(M_\epsilon)) \ge 1-\epsilon/2$. Furthermore, for n large enough, if B_{G_λ} and $J_{G_\lambda}(M_\epsilon)$ both hold, then F_λ holds because

$$\begin{aligned}
\mathrm{BIC}_\lambda - \mathrm{BIC}_{G_T} =& -2\left\{\ell\left(\hat{C}^{*,e}_{\lambda}\right) - \ell\left(\hat{C}^{*,e}_{G_T}\right)\right\} + m(\log n + 4\log p_n)\\
\ge& -2\left\{\ell\left(\hat{C}^{*,e}_{G_\lambda}\right) - \ell\left(\hat{C}^{*,e}_{G_T}\right)\right\} + m(\log n + 4\log p_n)\\
\ge& -4m\log(p_n) - 2M_\epsilon + m(\log n + 4\log p_n)\\
\ge& m\log(n) - 2M_\epsilon\\
>& 0,
\end{aligned}$$

with $\hat{C}^{*,e}_{G_\lambda}$ denoting the maximum likelihood estimator given the model G_λ. This implies that $P(\cap_{G_\lambda\in\mathcal{G}_+(m)}F_\lambda) \ge P(\cap_{G_\lambda\in\mathcal{G}_+(m)}\{B_{G_\lambda}\cap J_{G_\lambda}(M_\epsilon)\}) \ge 1-\epsilon$, for n large enough. Therefore, we have

$$P\Big\{\inf_{G_\lambda\in\mathcal{G}_+(m)}\mathrm{BIC}_\lambda > \mathrm{BIC}_{G_T}\Big\} \to 1, \text{ as } n\to\infty.$$

As there are only finite number of subsets $\mathcal{G}_+(m)$ within $\mathcal{G}_+$, in light of Lemma 3 the result follows.

References

Bai, Z. and Silverstein, J. W. (2006). *Spectral Analysis of Large Dimensional Random Matrices.* Science Press, Beijing.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.

Banerjee, O., Ghaoui, L. E. and D'Aspremont, A. (2007). Model selection through sparse maximum likelihood estimation. *J. Machine Learning Res.* **9**, 485-516.

Chen, J. H. and Chen, Z. H. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

Chen, J. H. and Chen, Z. H. (2012). Extended BIC for Small-n-large-P Sparse GLM. *Statist. Sinica*, **22**, 555-574.

Dempster, A. P. (1972). Covariance selection. *Biometrika* **32**, 95-108.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-60.

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Statist.* **3**, 521-541.

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Information Processing Systems* **23**, 2020-2028.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254-4278.

Lancaster, H. O. (1954). Traces and cumulants of quadratic forms in normal variables. *J. Roy. Statist. Soc. Ser. B* **16**, 247-254.

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

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs with the Lasso. *Ann. Statist.* **34**, 1436-62.

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic J. Statist.* **2**, 494-515.

Shao, J. (2003). *Mathematical Statistics.* Springer, New York.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267-288.

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. **94**, 553-68.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1533.

Department of Mathematics and Statistics, York University, Toronto, Canada.

E-mail: xingao@mathstat.yorku.ca

Department of Mathematics and Statistics, York University, Toronto, Canada.

E-mail: puq@mathstat.yorku.ca

Department of Mathematics and Statistics, York University, Toronto, Canada.

E-mail: wuyh@mathstat.yorku.ca

Department of Mathematics and Statistics, York University, Toronto, Canada.

E-mail: hongxu@mathstat.yorku.ca

(Received September 2009; accepted October 2011)