Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation *

By Clifford Lam, Jianqing Fan
London School of Economics, Princeton university

This paper studies the sparsistency and rates of convergence for estimating sparse covariance and precision matrices based on penalized likelihood with non-concave penalty functions. Here, sparsistency refers to the property that all parameters that are zero are actually estimated as zero with probability tending to one. Depending on the case of applications, sparsity *priori* may occur on the covariance matrix, its inverse or its Cholesky decomposition. We study these three sparsity exploration problems under a unified framework with a general penalty function. We show that the rates of convergence for these problems under the Frobenius norm are of order $(s_n \log p_n/n)^{1/2}$, where s_n is the number of non-sparse elements, p_n is the size of the covariance matrix and n is the sample size. This explicitly spells out the contribution of high-dimensionality is merely of a logarithmic factor. The biases of the estimators using different penalty functions are explicitly obtained. As a result, for the L_1-penalty, to guarantee the sparsistency and optimal rate of convergence, the non-sparsity rates should be low: $s'_n = O(p_n)$ at most, among $O(p_n^2)$ parameters, for estimating sparse covariance or correlation matrix, sparse precision or inverse correlation matrix or sparse Cholesky factor, where s'_n is the number of the non-sparse elements on the off-diagonal entries. On the other hand, using the SCAD or hard-thresholding penalty functions, there is no such a restriction.

Short Title: Covariance Estimation with Penalization.

*Clifford Lam is Lecturer, Department of Statistics, London School of Economics and Political Science, London, WC2A 2AE (email: C.Lam2@lse.ac.uk); Jianqing Fan is Professor, Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (email: jqfan@princeton.edu). Financial support from the NSF grant DMS-0354223, DMS-0704337 and NIH grant R01-GM072611 is gratefully acknowledged.

AMS 2000 subject classifications. Primary 62F12; secondary 62J07.

Key words and phrases. Covariance matrix, high dimensionality, consistency, nonconcave penalized likelihood, sparsistency, asymptotic normality.

1 Introduction

Covariance matrix estimation is a common statistical problem that arises in many scientific applications. For example, in financial risk assessment or longitudinal study, an input of covariance matrix $\mathbf{\Sigma}$ is needed, whereas an inverse of the covariance matrix, the precision matrix $\mathbf{\Sigma}^{-1}$, is required for optimal portfolio selection, linear discriminant analysis or graphical network models. Yet, the number of parameters in the covariance matrix grows quickly with dimensionality. Depending on the applications, the sparsity of the covariance matrix or precision matrix are frequently imposed to strike a balance between biases and variances. For example, in longitudinal data analysis (see e.g. Diggle and Verbyla (1998), or Bickel and Levina (2008b)), it is reasonable to assume that remote data in time are weakly correlated, whereas in Gaussian graphical models, the sparsity of the precision matrix is a reasonable assumption (Dempster (1972)).

This initiates a series of researches focusing on the parsimony of a covariance matrix. Smith and Kohn (2002) used priors which admit zeros on the off-diagonal elements of the Cholesky factor of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, while Wong, Carter and Kohn (2003) used zero-admitting prior directly on the off-diagonal elements of $\mathbf{\Omega}$ to achieve parsimony. Wu and Pourahmadi (2003) used the Modified Cholesky Decomposition (MCD) to nonparametrically find a banded structure for $\mathbf{\Omega}$ for longitudinal data while preserving positive definiteness of the resulting estimator. Bickel and Levina (2008b) developed consistency theories on banding methods for longitudinal data, both for $\mathbf{\Sigma}$

and $\mathbf{\Omega}$.

Penalized likelihood methods are used by various authors to achieve parsimony on covariance selection. Fan and Peng (2004) has laid down a general framework for penalized likelihood with diverging dimensionality, with general conditions for oracle property stated and proved. However, it is not clear whether it is applicable to the specific case of covariance matrix estimation. In particular, they did not link the dimensionality p_n with the non-sparsity size s_n, which is the number of non-zero elements in the true covariance matrix $\mathbf{\Sigma}_0$, or precision matrix $\mathbf{\Omega}_0$. A direct application of their results to our setting can only handle a relatively small covariance matrix of size $p_n = o(n^{1/10})$, which behaves like a constant p_n.

Recently, there is a surge of interest on the estimation of sparse covariance matrix or precision matrix using penalized likelihood method. Huang, Liu, Pourahmadi and Liu (2006) used the LASSO on the off-diagonal elements of the Cholesky factor from MCD, while Meinshausen and Bühlmann (2006), d'Aspremont, Banerjee, and El Ghaoui (2008) and Yuan and Lin (2007) use different LASSO algorithms to select sparse elements in the precision matrix. A novel penalty called the nested Lasso was constructed in Levina, Rothman and Zhu (2008) to penalize on these off-diagonal elements. Thresholding the sample covariance matrix in high-dimensional setting was thoroughly studied by El Karoui (2007) and Bickel and Levina (2008a) with remarkable results for high dimensional applications. However, it is not directly applicable to estimating sparse precision matrix when the dimensionality p_n is greater than the sample size n. Wagaman and Levina (2007) proposed an Isomap method for discovering meaningful orderings of variables based on their correlations that result in block-diagonal or banded correlation structure, resulting in an Isoband estimator. A permutation invariant estimator, called SPICE, was proposed in Rothman, Bickel, Levina and Zhu

(2007) based on penalized likelihood with L_1-penalty on the off-diagonal elements for the precision matrix. They obtained remarkable results on the rates of convergence. The rate for estimating $\mathbf{\Omega}$ under the Frobenius norm is of order $(s_n \log p_n/n)^{1/2}$, with dimensionality cost only a logarithmic factor in the overall mean-square error, where $s_n = p_n + s_{n1}$, p_n is the number of the diagonal elements and s_{n1} is the number of the non-sparse off-diagonal entries. However, such rate of convergence does not address explicitly the sparsistency such as those in Fan and Li (2001) and Zhao and Yu (2006), the bias issues of the L_1-penalty nor the sampling distribution of nonsparse elements. These are the core issues of the study. By sparsistency, we mean the property that all parameters that are zero are actually estimated as zero with probability tending to one, a more loose definition than that of Ravikumar, Lafferty, Liu and Wasserman (2008).

In this paper, we investigate the aforementioned problems using penalized likelihood method. Assume a normal random sample $\{\mathbf{y}_i\}_{1\le i\le n}$ with mean zero and covariance matrix $\mathbf{\Sigma}_0$. The sparsity of the true precision matrix $\mathbf{\Omega}_0$ can be explored by minimizing the penalized negative normal likelihood:

$$q_1(\mathbf{\Omega}) = \mathrm{tr}(\mathbf{S\Omega}) - \log|\mathbf{\Omega}| + \sum_{i\neq j} p_{\lambda_{n1}}(|\omega_{ij}|), \tag{1.1}$$

where $\mathbf{S} = n^{-1}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^T$ is the sample covariance matrix, with $\mathbf{\Omega} = (\omega_{ij})$, and $p_{\lambda_{n1}}(\cdot)$ is a penalty function, depending on a regularization parameter λ_{n1}, which can be nonconvex. For instance, the L_1-penalty $p_\lambda(\theta) = \lambda|\theta|$ is convex, while the hard-thresholding penalty defined by $p_\lambda(\theta) = \lambda^2 - (|\theta| - \lambda)^2\mathbf{1}_{\{|\theta|<\lambda\}}$, and the SCAD penalty defined by

$$p'_\lambda(\theta) = \lambda\mathbf{1}_{\{\theta\le\lambda\}} + (a\lambda - \theta)_+\mathbf{1}_{\{\theta>\lambda\}}/(a-1), \text{ for some } a > 2, \tag{1.2}$$

are nonconvex. Nonconvex penalty is introduced to reduce bias when the true parameter has a relatively large magnitude. For example, the SCAD penalty remains constant when θ is large, while the L_1-penalty grows linearly with θ. See Fan and Li (2001) for a detailed account of this and other advantages of such a penalty function.

Similarly, the sparsity of the true covariance matrix $\mathbf{\Sigma}_0$ can be explored by minimizing

$$q_2(\mathbf{\Sigma}) = \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) + \log|\mathbf{\Sigma}| + \sum_{i\neq j} p_{\lambda_{n2}}(|\sigma_{ij}|), \tag{1.3}$$

where $\mathbf{\Sigma} = (\sigma_{ij})$. Note that we only penalize the off-diagonal elements of $\mathbf{\Sigma}$ or $\mathbf{\Omega}$ in the aforementioned two methods, since the diagonal elements of $\mathbf{\Sigma}_0$ and $\mathbf{\Omega}_0$ do not vanish.

The computation of the non-concave maximum likelihood problems can be solved by a sequence of L_1-penalized likelihood problems via local linear approximation (Zou and Li (2008)). For example, given the current estimate $\Omega_k = (\omega_{ij,k})$, by the local linear approximation to the penalty function,

$$\begin{aligned} q_1(\mathbf{\Omega}) \approx & \text{tr}(\mathbf{S}\mathbf{\Omega}) - \log|\mathbf{\Omega}| \\ & + \sum_{i\neq j}[p_{\lambda_{n1}}(|\omega_{ij,k}|) + p'_{\lambda_{n1}}(|\omega_{ij,k}|)(|\omega_{ij}| - |\omega_{ij,k}|)]. \end{aligned} \tag{1.4}$$

Hence, $\mathbf{\Omega}_{k+1}$ should be taken to maximize the right-hand side of (1.4):

$$\mathbf{\Omega}_{k+1} = \text{argmax}_{\mathbf{\Omega}} \left[\text{tr}(\mathbf{S}\mathbf{\Omega}) - \log|\mathbf{\Omega}| + \sum_{i\neq j} p'_{\lambda_{n1}}(|\omega_{ij,k}|)|\omega_{ij}| \right], \tag{1.5}$$

after ignoring the two constant terms. Problem (1.5) is the penalized L_1-likelihood. In particular, if we take the most primitive initial value $\mathbf{\Omega}_0 = \mathbf{0}$, then

$$\mathbf{\Omega}_1 = \text{argmax}_{\mathbf{\Omega}} \left[\text{tr}(\mathbf{S}\mathbf{\Omega}) - \log|\mathbf{\Omega}| + \lambda_{n1} \sum_{i\neq j} |\omega_{ij}| \right],$$

is already a good estimator. Iterations of (1.5) reduces the biases of the estimator. In fact, in a different setup, Zou and Li (2008) shows that one iteration of such a procedure suffices as long as the initial values are good enough. See Fan, Feng and Wu (2008) for detailed implementations on the estimation of precision matrices. See also Zhang (2007) for a general solution to the nonconvex penalized least-squares problem.

In studying sparse covariance or precision matrix, it is important to distinguish between the diagonal and off-diagonal elements, since the diagonal elements are always positive and contribute to the overall mean-squares errors. For example, the true correlation matrix, denoted by $\mathbf{\Gamma}_0$, has the same sparsity structure as $\mathbf{\Sigma}_0$ without the need to estimate its diagonal elements. In view of this fact, we introduce a revised method (3.2) to take this advantage. It turns out that the correlation matrix can be estimated with a faster rate of convergence, with rate $(s_{n1} \log p_n/n)^{1/2}$ instead of $((p_n + s_{n1}) \log p_n/n)^{1/2}$, where s_{n1} is the number of non-vanishing correlation coefficients. Similar advantages can be taken on the estimation of the true inverse correlation matrix, denoted by $\mathbf{\Psi}_0$. See Section 2.2. This is an extension of the work of Rothman *et al.* (2007) using the L_1-penalty. Such an extension is important since the non-concave penalized likelihood ameliorates the bias problem of the L_1-penalized likelihood.

The bias issues of the commonly used L_1-penalty, or LASSO, can be seen from our theoretical results. In fact, it is not always possible to choose the regularization parameters λ_{ni} in the problems (1.3) and (1.1) to satisfy both consistency and sparsistency properties. This is in fact one of the motivations for introducing nonconvex penalty functions in Fan and Li (2001) and Fan and Peng (2004), but we state and prove the explicit rates in the current context. In particular, we demonstrate that L_1-penalized likelihood can achieve simultaneously the optimal rate and sparsistency for estimation of $\mathbf{\Sigma}_0$ or $\mathbf{\Omega}_0$ only when the number of nonsparse elements in off-diagonal

entries are no larger than $O(p_n)$. On the other hand, using the nonconvex penalty like SCAD or hard-thresholding penalty, such an extra restriction is not needed.

In this paper, we also compare two different formulations of penalized likelihood using the modified Cholesky decomposition, exploring their respective rates of convergence and sparsity properties.

Throughout this paper, we use $\lambda_{\min}(A)$, $\lambda_{\max}(A)$, and $\mathrm{tr}(A)$ to denote the minimum eigenvalue, maximum eigenvalue, and trace of a symmetric matrix A, respectively. For a matrix B, we define the operator norm and the Frobenius norm, respectively, as $\|B\| = \lambda_{\max}^{1/2}(B^TB)$ and $\|B\|_F = \mathrm{tr}^{1/2}(B^TB)$.

2 Estimation of sparse precision matrix

In this section we present the analysis of (1.1) for estimating sparse precision matrix. Before stating and proving the rate of convergence and sparsistency of the resulting estimator, we introduce some notations and present regularity conditions concerning the penalty function $p_\lambda(\cdot)$ and the precision matrix $\mathbf{\Omega}_0$.

Let $S_1 = \{(i,j) : \omega_{ij}^0 \neq 0\}$, where $\mathbf{\Omega}_0 = (\omega_{ij}^0)$. Denote $s_{n1} = |S_1| - p_n$, which is the number of non-zero elements in the off-diagonal entries of $\mathbf{\Omega}_0$. Define

$$a_{n1} = \max_{(i,j)\in S_1} p'_{\lambda_{n1}}(|\omega_{ij}^0|), \quad b_{n1} = \max_{(i,j)\in S_1} p''_{\lambda_{n1}}(|\omega_{ij}^0|).$$

The term a_{n1} is related to the biases of the penalized likelihood estimate due to penalization. Note that for L_1-penalty, $a_{n1} = \lambda_n$ and $b_{n1} = 0$, whereas for SCAD, $a_{n1} = b_{n1} = 0$ for sufficiently large n under the last assumption of condition (B) below.

We assume the following regularity conditions:

(A) There exists constants τ_1 and τ_2 such that

$$0 < \tau_1 < \lambda_{\min}(\mathbf{\Sigma}_0) \leq \lambda_{\max}(\mathbf{\Sigma}_0) < \tau_2 < \infty \quad \text{for all } n.$$

(B) $a_{n1} = O(\{1 + p_n/(s_{n1}+1)\}(\log p_n/n)^{1/2})$, $b_{n1} = o(1)$, and $\min_{(i,j)\in S_1} |\omega_{ij}^0|/\lambda_{n1} \to \infty$ as $n \to \infty$.

(C) The penalty $p_\lambda(\cdot)$ is singular at the origin, with $\lim_{t\downarrow 0} p_\lambda(t)/(\lambda t) = k > 0$.

(D) There are constants C and D such that, when $\theta_1, \theta_2 > C\lambda_{n1}$, $|p'_{\lambda_{n1}}(\theta_1) - p'_{\lambda_{n1}}(\theta_2)| \leq D|\theta_1 - \theta_2|$.

Condition (A) bounds uniformly the eigenvalues of $\mathbf{\Sigma}_0$, which facilitates the proof of consistency. It also includes a wide class of covariance matrices as noted in Bickel and Levina (2008b). The rates a_{n1} and b_{n1} in condition (B) are also needed for proving consistency. If they are too large, the penalty term can dominate the likelihood term, resulting in poor estimates.

The last requirement in condition (B) states the rate at which the non-zero parameters can be distinguished from zero asymptotically. It is not explicitly needed in the proofs, but for asymptotically unbiased penalty functions, this is a necessary condition so that a_{n1} and b_{n1} are converging to zero fast enough as needed in the first part of condition (B). In particular, for the SCAD and hard-thresholding penalties, this condition implies that $a_{n1} = b_{n1} = 0$ exactly for sufficiently large n, thus allowing a flexible choice of λ_{n1}. For the SCAD penalty (1.2), the condition can be relaxed as $\min_{(i,j)\in S_1} |\omega_{ij}^0|/\lambda_{n1} > a$.

Singularity of the origin in condition (C) allows for sparse estimates (Fan and Li (2001)). Finally, condition (D) is a smoothing condition for the penalty function, and

is needed in proving asymptotic normality. The SCAD penalty, for instance, satisfies this condition by choosing the constant D, independent of n, to be large enough.

2.1 Properties of sparse precision matrix estimation

Minimizing (1.1) involves nonconvex minimization, and we need to prove that there exists a local minimizer $\hat{\mathbf{\Omega}}$ for the minimization problem. We give the rate of convergence under Frobenius norm. The proof is given in section 5. It is close to the one given in Rothman *et al.* (2007), but we now allow for a nonconvex penalty.

Theorem 1 *(Rate of convergence). Under regularity conditions (A)-(D), if* $(s_{n1} + 1)\log p_n/n = O(\lambda_{n1}^2)$ *and* $(p_n + s_{n1})\log p_n/n = o(1)$*, then there exists a local minimizer* $\hat{\mathbf{\Omega}}$ *such that* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_F^2 = O_P\{(p_n + s_{n1})\log p_n/n\}$.

Theorem 1 states explicitly how the non-sparsity size and dimensionality affect the rate of convergence. Since there are $(p_n + s_{n1})$ non-zero elements and each of them can be estimated at best with rate $O(n^{-1/2})$, the total square errors are at least of rate $(p_n + s_{n1})/n$. The price that we pay for high-dimensionality is merely a logarithmic factor $\log p_n$.

Theorem 1 is also applicable to the L_1-penalty function, where the condition for λ_{n1} can be relaxed to $\log p_n/n = O(\lambda_{n1}^2)$. In this case, the local minimizer becomes the global minimizer. The bias of the L_1-penalized estimate $a_{n1} \asymp \lambda_{n1}$ is controlled via condition (B), which entails an upper bound on $\lambda_{n1} = O((1 + p_n/(s_{n1} + 1))(\log p_n/n)^{1/2})$.

Next we show the sparsistency of the penalized covariance estimator (1.1). We use S^c to denote the complement of a set S.

Theorem 2 *(Sparsistency). Under regularity conditions (A), (C) and (D), for any local minimizer of (1.1) satisfying* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_F^2 = O_P\{(p_n + s_{n1})\log p_n/n\}$ *and* $\|\hat{\mathbf{\Omega}} -$

$\boldsymbol{\Omega}_0\|^2 = O_P(\eta_n)$ *for a sequence of* $\eta_n \to 0$, *if* $\log p_n/n + \eta_n = O(\lambda_{n1}^2)$, *then with probability tending to 1,* $\hat{\omega}_{ij} = 0$ *for all* $(i,j) \in S_1^c$.

First of all, since $\|M\|^2 \le \|M\|_F^2$ for any matrix M, we can always take $\eta_n = (p_n + s_{n1}) \log p_n/n$ in Theorem 2, but this will result in more stringent requirement on the number of sparse elements when L_1-penalty is used, as we now explain. The sparsistency requires a lower bound on the rate of the regularization parameter λ_{n1}. On the other hand, condition (B) imposes an upper bound on λ_{n1} when L_1-penalty is used in order to control the biases. Explicitly, we need, for L_1-penalized likelihood,

$$\log p_n/n + \eta_n = O(\lambda_{n1}^2) = (1 + p_n/(s_{n1} + 1))^2 \log p_n/n \tag{2.1}$$

for both consistency and sparsistency to be satisfied. We present two scenarios here for the two bounds to be compatible, making use of the inequalities $\|M\|_F^2/p_n \le \|M\|^2 \le \|M\|_F^2$ for a matrix M of size p_n.

1. We always have $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\| \le \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_F$. In the worst case scenario where they have the same order, then $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|^2 = O_P((p_n + s_{n1}) \log p_n/n)$ so that $\eta_n = (p_n + s_{n1}) \log p_n/n$. It is then easy to see from (2.1) that the two bounds are compatible only when $s_{n1} = O(p_n^{1/2})$.

2. We also have $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_F^2/p_n \le \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|^2$. In the optimistic scenario where they have the same order,

   $$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|^2 = O_P((1 + s_{n1}/p_n) \log p_n/n),$$

   where $1 + s_{n1}/p_n$ is the average number of non-zero elements in a row of the matrix $\boldsymbol{\Omega}_0$. Hence $\eta_n = (1 + s_{n1}/p_n) \log p_n/n$, and compatibility of the bounds requires $s_{n1} = O(p_n)$.

Hence even in the optimistic scenario, consistency and sparsistency are guaranteed only when $s_{n1} = O(p_n)$ if the L_1-penalty is used, i.e. the precision matrix has to be sparse enough.

However, if the penalty function used is unbiased, like the SCAD or the hard-thresholding penalties, we do not impose an extra upper bound for λ_{n1} since its first derivative $p'_{\lambda_{n1}}(|\theta|)$ goes to zero fast enough as $|\theta|$ increases (exactly equals zero for the SCAD and hard-thresholding penalties, when n is sufficiently large; see condition (B) and the explanation thereof). Thus, λ_{n1} is allowed to decay slower to zero than that for the L_1-penalty, allowing even the largest order $s_{n1} = O(p_n^2)$.

We remark that asymptotic normality for the estimators of the elements in S_1 have been established in a previous version of this paper. We omit it here for brevity.

2.2 Properties of sparse inverse correlation matrix estimation

The inverse correlation matrix $\mathbf{\Psi}_0$ retains the same sparse structure of $\mathbf{\Omega}_0$. Consistency and sparsity results can be achieved with p_n as large as $\log p_n = o(n)$, as long as $(s_{n1} + 1) \log p_n / n = o(1)$. We minimize, w.r.t. $\mathbf{\Psi} = (\psi_{ij})$,

$$\mathrm{tr}(\mathbf{\Psi}\hat{\mathbf{\Gamma}}_{\mathbf{S}}) - \log|\mathbf{\Psi}| + \sum_{i \neq j} p_{\nu_{n1}}(|\psi_{ij}|), \tag{2.2}$$

where $\hat{\mathbf{\Gamma}}_{\mathbf{S}} = \hat{\mathbf{W}}^{-1}\mathbf{S}\hat{\mathbf{W}}^{-1}$ is the sample correlation matrix, with $\hat{\mathbf{W}}^2 = \mathbf{D}_{\mathbf{S}}$ being the diagonal matrix with diagonal elements of $\mathbf{S}$, and ν_{n1} is a regularization parameter. After obtaining $\hat{\mathbf{\Psi}}$, $\mathbf{\Omega}_0$ can also be estimated by $\tilde{\mathbf{\Omega}} = \hat{\mathbf{W}}^{-1}\hat{\mathbf{\Psi}}\hat{\mathbf{W}}^{-1}$.

To present the rates of convergence for $\hat{\mathbf{\Psi}}$ and $\tilde{\mathbf{\Omega}}$, we define

$$c_{n1} = \max_{(i,j)\in S_1} p'_{\nu_{n1}}(|\psi^0_{ij}|), \quad d_{n1} = \max_{(i,j)\in S_1} p''_{\nu_{n1}}(|\psi^0_{ij}|),$$

where $\mathbf{\Psi}_0 = (\psi^0_{ij})$ and modify condition (D) to (D') with λ_{n1} there replaced by ν_{n1}, and impose

(B') $c_{n1} = O(\{\log p_n/n\}^{1/2})$, $d_{n1} = o(1)$. Also, $\min_{(i,j)\in S_1} |\psi^0_{ij}|/\nu_{n1} \to \infty$ as $n \to \infty$.

Theorem 3 *Under regularity conditions (A),(B'),(C) and (D'), if* $(s_{n1}+1)\log p_n/n = o(1)$ *and* $(s_{n1}+1)\log p_n/n = O(\nu^2_{n1})$, *then there exists a local minimizer* $\hat{\mathbf{\Psi}}$ *for (2.2) such that* $\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|^2_F = O_P(s_{n1}\log p_n/n)$ *and* $\|\tilde{\mathbf{\Omega}} - \mathbf{\Omega}_0\|^2 = O_P((s_{n1}+1)\log p_n/n)$ *under the operator norm.*

The proof is sketched in section 5. Note that an order of $\{p_n \log p_n/n\}^{1/2}$ is removed by estimating the inverse correlation rather than the precision matrix, which is somewhat surprising since inverse correlation matrix, unlike correlation matrix, does not have known diagonal elements that contribute no errors to the estimation. This can be explained and proved as follows. If $s_{n1} = O(p_n)$, the result is obvious. When $s_{n1} = o(p_n)$, most of off-diagonal elements are zero. Indeed, there are at most $O(s_{n1})$ columns of the inverse correlation matrix contain at least one non-zero elements. The rest of the columns that have all zero off-diagonal elements must have diagonal entries 1. These columns represent variables that are actually uncorrelated from the rest. Now, it is easy to see from (2.2), that these diagonal elements, which are one, are all estimated exactly as one with no estimation error. Hence an order of $(p_n \log p_n/n)^{1/2}$ is not present even in the case of estimating the inverse correlation matrix.

For the L_1-penalty, our result reduces to that given in Rothman *et al.* (2007), and the condition for ν_{n1} can be relaxed to $\log p_n/n = O(\nu^2_{n1})$. We offer the sparsistency result as follows.

Theorem 4 *(Sparsistency) Under the conditions given in Theorem 3, for any local minimizer of (2.2) satisfying* $\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|^2_F = O_P(s_{n1}\log p_n/n)$ *and* $\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|^2 = O_P(\eta_n)$

for some $\eta_n \to 0$, *if* $\log p_n/n + \eta_n = O(\nu_{n1}^2)$, *then with probability tending to 1,* $\hat{\psi}_{ij} = 0$ *for all* $(i,j) \in S_1^c$.

The proof follows exactly the same as that for Theorem 2 in section 2.1, and is thus omitted.

For the L_1-penalty, control of biases and sparsistency requires ν_{n1} to satisfy bounds like (2.1):

$$\log p_n/n + \eta_n = O(\nu_{n1}^2) = \log p_n/n. \tag{2.3}$$

This leads to two scenarios:

1. The worst case scenario has

   $$\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|^2 = \|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|_F^2 = O_P(s_{n1} \log p_n/n),$$

   meaning $\eta_n = s_{n1} \log p_n/n$. Then compatibility of the bounds in (2.3) requires $s_{n1} = O(1)$.

2. The optimistic scenario has

   $$\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|^2 = \|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|_F^2/p_n = O_P(s_{n1}/p_n \cdot \log p_n/n),$$

   meaning $\eta_n = s_{n1}/p_n \cdot \log p_n/n$. Then compatibility of the bounds in (2.3) requires $s_{n1} = O(p_n)$.

On the other hand, for penalties like the SCAD or the hard-thresholding penalty, we do not need an upper bound for s_{n1}. Hence there is no restriction on the order of s_{n1} as long as $s_{n1} \log p_n/n = o(1)$. It is clear that SCAD results in better sampling properties than the L_1-penalized estimator in precision or inverse correlation matrix estimation.

3 Estimation of sparse covariance matrix

In this section, we analyze the sparse covariance estimation using penalized likelihood (1.3). Then it is modified to estimate the correlation matrix, which improves the rate of convergence.

3.1 Properties of sparse covariance matrix estimation

Let $S_2 = \{(i,j) : \sigma_{ij}^0 \neq 0\}$, where $\mathbf{\Sigma}_0 = (\sigma_{ij}^0)$. Denote by $s_{n2} = |S_2| - p_n$, so that s_{n2} is the non-sparsity size for $\mathbf{\Sigma}_0$ on the off-diagonal entries. Put

$$a_{n2} = \max_{(i,j)\in S_2} p'_{\lambda_{n2}}(|\sigma_{ij}^0|), \quad b_{n2} = \max_{(i,j)\in S_2} p''_{\lambda_{n2}}(|\sigma_{ij}^0|).$$

Technical conditions in section 2 need some revision. In particular, condition (D) now becomes condition (D2) with λ_{n1} there replaced by λ_{n2}. Condition (B) should now be

(B2) $a_{n2} = O(\{1 + p_n/(s_{n2}+1)\}(\log p_n/n)^{1/2})$, $b_{n2} = o(1)$, and
$\min_{(i,j)\in S_2} |\sigma_{ij}^0|/\lambda_{n2} \to \infty$ as $n \to \infty$.

Theorem 5 *(Rate of convergence). Under regularity conditions (A), (B2), (C) and (D2), if* $(p_n + s_{n2}) \log p_n/n = o(1)$ *and* $(s_{n2}+1)\log p_n/n = O(\lambda_{n2}^2)$*, then there exists a local minimizer* $\hat{\mathbf{\Sigma}}$ *such that* $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F^2 = O_P\{(p_n + s_{n2})\log p_n/n\}$.

The proof is given in section 5. When the L_1-penalty is used, condition for λ_{n2} is relaxed to $\log p_n/n = O(\lambda_{n2}^2)$. Like the case for precision matrix estimation, the control of the bias term a_{n2} imposes, for the L_1-penalty, $\lambda_{n2} = O((1 + p_n/(s_{n2} + 1))^2(\log p_n/n)^{1/2})$.

Theorem 6 *(Sparsistency). Under conditions in Theorem 5, for any local minimizer* $\hat{\mathbf{\Sigma}}$ *of (1.3) satisfying* $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F^2 = O_P((p_n + s_{n2}) \log p_n/n)$ *and* $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|^2 = O_P(\eta_n)$ *for some* $\eta_n \to 0$*, if* $\log p_n/n + \eta_n = O(\lambda_{n2}^2)$*, then with probability tending to 1,* $\hat{\sigma}_{ij} = 0$ *for all* $(i, j) \in S_2^c$.

The proof is sketched in section 5. For the L_1-penalized likelihood, controlling of bias for consistency together with sparsistency requires

$$\log p_n/n + \eta_n = O(\lambda_{n2}^2) = (1 + p_n/(s_{n2} + 1))^2 \log p_n/n. \tag{3.1}$$

This is the same condition as (2.1), and hence in the worst case scenario where

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|^2 = \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F^2 = O_P((p_n + s_{n2}) \log p_n/n),$$

we need $s_{n2} = O(p_n^{1/2})$. In the optimistic scenario where

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|^2 = \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F^2/p_n,$$

we need $s_{n2} = O(p_n)$. In both cases, the matrix $\mathbf{\Sigma}_0$ has to be very sparse, but the former is much sparser.

On the other hand, if unbiased penalty functions like the SCAD or hard-thresholding penalties are used, we do not need an upper bound on λ_{n2} since the bias $a_{n2} = 0$ for sufficiently large n. This allows for more flexibility on the order of s_{n2}.

Similar to section 2, asymptotic normality for the estimators of the elements in S_2 can be proved under certain assumptions.

3.2 Properties of sparse correlation matrix estimation

The correlation matrix $\mathbf{\Gamma}_0$ retains the same sparse structure of $\mathbf{\Sigma}_0$ with known diagonal elements. This special structure allows us to estimate $\mathbf{\Gamma}_0$ more accurately. To take

the advantage of the known diagonal elements, the sparse correlation matrix $\mathbf{\Gamma}_0$ is estimated by minimizing w.r.t. $\mathbf{\Gamma} = (\gamma_{ij})$,

$$\mathrm{tr}(\mathbf{\Gamma}^{-1}\hat{\mathbf{\Gamma}}_{\mathbf{S}}) + \log|\mathbf{\Gamma}| + \sum_{i \neq j} p_{\nu_{n2}}(|\gamma_{ij}|), \tag{3.2}$$

where ν_{n2} is a regularization parameter. After obtaining $\hat{\mathbf{\Gamma}}$, $\mathbf{\Sigma}_0$ can be estimated by $\tilde{\mathbf{\Sigma}} = \hat{\mathbf{W}}\hat{\mathbf{\Gamma}}\hat{\mathbf{W}}$.

To present the rates of convergence for $\hat{\mathbf{\Gamma}}$ and $\tilde{\mathbf{\Sigma}}$, we define

$$c_{n2} = \max_{(i,j)\in S_2} p'_{\nu_{n2}}(|\gamma^0_{ij}|), \quad d_{n2} = \max_{(i,j)\in S_2} p''_{\nu_{n2}}(|\gamma^0_{ij}|),$$

where $\mathbf{\Gamma}_0 = (\gamma^0_{ij})$. We adapt the condition (D) to (D2') with λ_{n2} there replaced by ν_{n2}, and (B) to (B2') as follows:

(B2') $c_{n2} = O(\{\log p_n/n\}^{1/2})$, $d_{n2} = o(1)$, and $\min_{(i,j)\in S_2} |\gamma^0_{ij}|/\nu_{n2} \to \infty$ as $n \to \infty$.

Theorem 7 *Under regularity conditions (A),(B2'),(C) and (D2'), if* $s_{n2}\log p_n/n = o(1)$ *and* $(s_{n2}+1)\log p_n/n = O(\nu^2_{n2})$, *then there exists a local minimizer* $\hat{\mathbf{\Gamma}}$ *for (3.2) such that*

$$\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|_F^2 = O_P(s_{n2}\log p_n/n).$$

In addition, for the operator norm, we have

$$\|\tilde{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|^2 = O_P\{(s_{n2}+1)\log p_n/n\}.$$

The proof is sketched in section 5. The condition $(s_{n2}+1)\log p_n/n = O(\nu^2_{n2})$ can be relaxed to $\log p_n/n = O(\nu^2_{n2})$ when the L_1-penalty is used. This theorem shows that the correlation matrix, like the inverse correlation matrix, can be estimated more accurately, since diagonal elements are known to be one.

Theorem 8 *(Sparsistency). Under conditions in Theorem 7, for any local minimizer* $\hat{\mathbf{\Gamma}}$ *of (3.2) satisfying* $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|_F^2 = O_P(s_{n2} \log p_n/n)$ *and* $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|^2 = O_P(\eta_n)$ *for some* $\eta_n \to 0$ *, if* $\log p_n/n + \eta_n = O(\nu_{n2}^2)$*, then with probability tending to 1,* $\hat{\gamma}_{ij} = 0$ *for all* $(i, j) \in S_2^c$.

The proof follows exactly the same as that for Theorem 6 in section 5, and is omitted. For the L_1-penalized likelihood, controlling of bias and sparsistency requires

$$\log p_n/n + \eta_n = O(\nu_{n2}^2) = \log p_n/n. \tag{3.3}$$

This is the same condition as (2.3), hence in the worst scenario where

$$\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|^2 = \|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|_F^2 = O_P(s_{n2} \log p_n/n),$$

we need $s_{n2} = O(1)$. In the optimistic scenario where

$$\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|^2 = \|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}_0\|_F^2/p_n = O_P(s_{n2}/p_n \cdot \log p_n/n),$$

we need $s_{n2} = O(p_n)$.

The use of unbiased penalties like the SCAD or hard-thresholding penalties, like results in previous sections, does not impose an upper bound on the regularization parameter since bias $c_{n2} = 0$ for sufficiently large n. This gives more flexibility to the order of s_{n2} allowed.

4 Extension to sparse Cholesky decomposition

Pourahmadi (1999) proposed the modified Cholesky decomposition (MCD) which facilitates the sparse estimation of $\mathbf{\Omega}$ through penalization. The idea is to represent

zero-mean data $\mathbf{y} = (y_1, \cdots, y_{p_n})^T$ using autoregressive models:

$$y_i = \sum_{j=1}^{i-1} \phi_{ij} y_j + \epsilon_i, \text{ and } \mathbf{T}\mathbf{\Sigma}\mathbf{T}^T = \mathbf{D}, \tag{4.1}$$

where $\mathbf{T}$ is the unique unit lower triangular matrix with ones on its diagonal and $(i,j)^{\text{th}}$ element $-\phi_{ij}$ for $j < i$, and $\mathbf{D}$ is diagonal with i^{th} element $\sigma_i^2 = \text{var}(\epsilon_i)$. The optimization problem is unconstrained (since the ϕ_{ij}'s are free variables), and the estimate for $\mathbf{\Omega}$ is always positive-definite.

Huang *et al.* (2006) and Levina *et al.* (2008) both used the MCD for estimation of $\mathbf{\Omega}_0$. The former maximized the log-likelihood (ML) over $\mathbf{T}$ and $\mathbf{D}$ simultaneously, while the latter suggested also a least square version (LS), with $\mathbf{D}$ being first set to the identity matrix and then minimizing over $\mathbf{T}$ to obtain $\hat{\mathbf{T}}$. The latter corresponds to the original Cholesky decomposition. The sparse Cholesky factor can be estimated through

$$(ML) : q_3(\mathbf{T}, \mathbf{D}) = \text{tr}(\mathbf{T}^T\mathbf{D}^{-1}\mathbf{T}\mathbf{S}) + \log|\mathbf{D}| + 2\sum_{i<j} p_{\lambda_{n3}}(|t_{ij}|). \tag{4.2}$$

This is indeed the same as (1.1) with the substitution of $\mathbf{\Omega} = \mathbf{T}^T\mathbf{D}^{-1}\mathbf{T}$ and penalization parameter λ_{n3}. Noticing that (4.1) can be written as $\mathbf{T}\mathbf{y} = \boldsymbol{\varepsilon}$, the least square version is to minimize $\text{tr}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \text{tr}(\mathbf{T}^T\mathbf{T}\mathbf{y}\mathbf{y}^T)$ in the matrix notation. Aggregating n observations and adding sparsity penalties, the least-square criterion is to minimize

$$(LS) : \qquad q_4(\mathbf{T}) = \text{tr}(\mathbf{T}^T\mathbf{T}\mathbf{S}) + 2\sum_{i<j} p_{\lambda_{n4}}(|t_{ij}|). \tag{4.3}$$

In view of the results in sections 2.2 and 3.2, we can also write the covariance in (4.2) as $\mathbf{S} = \hat{\mathbf{W}}\hat{\mathbf{\Gamma}}_{\mathbf{S}}\hat{\mathbf{W}}$ and then replace $\mathbf{D}^{-1/2}\mathbf{T}\hat{\mathbf{W}}$ by $\mathbf{T}$, resulting in the normalized (NL) version as follows:

$$(NL) : \qquad q_5(\mathbf{T}) = \text{tr}(\mathbf{T}^T\mathbf{T}\hat{\mathbf{\Gamma}}_{\mathbf{S}}) - 2\log|\mathbf{T}| + 2\sum_{i<j} p_{\lambda_{n5}}(|t_{ij}|). \tag{4.4}$$

4.1 Properties of sparse Cholesky factor estimation

Since all the $\mathbf{T}$'s introduced in the three models above have the same sparsity structure, let S and s_{n3} be the non-sparsity set and non-sparsity size associated with each $\mathbf{T}$ above. Define

$$a_{n3} = \max_{(i,j)\in S} p'_{\lambda_{n3}}(|t^0_{ij}|), \quad b_{n3} = \max_{(i,j)\in S} p''_{\lambda_{n3}}(|t^0_{ij}|).$$

For (ML), condition (D) is adapted to (D3) with λ_{n1} there replaced by λ_{n3}. Condition (B) is modified as

(B3) $a_{n3} = O(\{1 + p_n/(s_{n3}+1)\}(\log p_n/n)^{1/2})$, $b_{n3} = o(1)$ and
$\min_{(i,j)\in S} |\phi^0_{ij}|/\lambda_{n3} \to \infty$ as $n \to \infty$.

After obtaining $\hat{\mathbf{T}}$ and $\hat{\mathbf{D}}$ from minimizing (ML), we set $\hat{\mathbf{\Omega}} = \hat{\mathbf{T}}^T\hat{\mathbf{D}}^{-1}\hat{\mathbf{T}}$.

Theorem 9 *Under regularity conditions (A),(B3),(C),(D3), if* $(p_n + s_{n3})\log p_n/n = o(1)$ *and* $(s_{n3}+1)\log p_n/n = O(\lambda^2_{n3})$, *then there exists a local minimizer* $\hat{\mathbf{T}}$ *and* $\hat{\mathbf{D}}$ *for (ML) such that* $\|\hat{\mathbf{T}} - \mathbf{T}_0\|^2_F = O_P(s_{n3}\log p_n/n)$, $\|\hat{\mathbf{D}} - \mathbf{D}_0\|^2_F = O_P(p_n \log p_n/n)$ *and* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|^2_F = O_P\{(p_n + s_{n3})\log p_n/n\}$.

The proof is similar to those of Theorems 5 and 7 and is omitted. The Cholesky factor $\mathbf{T}$ has ones on its main diagonal without the need for estimation. Hence, the rate of convergence is faster than $\hat{\mathbf{\Omega}}$. If the L_1-penalty is used, condition for λ_{n3} can be relaxed to $\log p_n/n = O(\lambda^2_{n3})$.

Theorem 10 *(Sparsistency). Under the conditions in Theorem 9, for any local minimizer* $\hat{\mathbf{T}}$, $\hat{\mathbf{D}}$ *of (4.2) satisfying* $\|\hat{\mathbf{T}} - \mathbf{T}_0\|^2_F = O_P(s_{n3}\log p_n/n)$ *and* $\|\hat{\mathbf{D}} - \mathbf{D}_0\|^2_F = O_P(p_n \log p_n/n)$, *if* $\log p_n/n + \eta_n + \zeta_n = O(\lambda^2_{n3})$, *then sparsistency holds for* $\hat{\mathbf{T}}$, *provided that* $\|\hat{\mathbf{T}} - \mathbf{T}_0\|^2 = O_P(\eta_n)$ *and* $\|\hat{\mathbf{D}} - \mathbf{D}_0\|^2 = O_P(\zeta_n)$, *for some* $\eta_n, \zeta_n \to 0$.

The proof is in section 5. For the L_1-penalized likelihood, control of bias and sparsistency impose the following:

$$\log p_n/n + \eta_n + \zeta_n = O(\lambda_{n3}^2) = (1 + p_n/(s_{n3}+1))^2 \log p_n/n. \tag{4.5}$$

The worst scenario corresponds to $\eta_n = s_{n3} \log p_n/n$ and $\zeta_n = p_n \log p_n/n$, so that we need $s_{n3} = O(p_n^{1/2})$. The optimistic scenario corresponds to $\eta_n = s_{n3}/p_n \cdot \log p_n/n$ and $\zeta_n = \log p_n/n$, so that we need $s_{n3} = O(p_n)$.

On the other hand, such a restriction is not needed for unbiased penalties like SCAD or hard-thresholding, which gives more flexibility on the order of s_{n3}.

4.2 Properties of sparse normalized Cholesky factor estimation

We now turn to analyzing the normalized penalized likelihood (4.4). With $\mathbf{T} = (t_{ij})$ in (NL) which is lower triangular, define

$$a_{n5} = \max_{(i,j)\in S} p'_{\lambda_{n5}}(|t_{ij}^0|), \quad b_{n5} = \max_{(i,j)\in S} p''_{\lambda_{n5}}(|t_{ij}^0|).$$

Condition (D) is now changed to (D5) with λ_{n1} there replaced by λ_{n5}. Condition (B) is now substituted by

(B5) $a_{n5}^2 = O(\log p_n/n)$, $b_{n5} = o(1)$, $\min_{(i,j)\in S} |t_{ij}^0|/\lambda_{n5} \to \infty$ as $n \to \infty$.

Theorem 11 *(Rate of convergence) Under regularity conditions (A),(B5),(C) and (D5), if* $s_{n3} \log p_n/n = o(1)$ *and* $(s_{n3}+1) \log p_n/n = O(\lambda_{n5}^2)$*, then there exists a local minimizer* $\hat{\mathbf{T}}$ *for (NL) such that* $\|\hat{\mathbf{T}} - \mathbf{T}_0\|_F^2 = O_P(s_{n3} \log p_n/n)$ *and rate of convergence in the Frobenius norm*

$$\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_F^2 = O_P\{(p_n + s_{n3}) \log p_n/n\},$$

and in the operator norm, it is improved to

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|^2 = O_P\{(s_{n3}+1)\log p_n/n)\}.$$

The proof is similar to that of Theorems 5 and 7 and is omitted. The condition for λ_{n3} can be relaxed to $\log p_n/n = O(\lambda_{n3}^2)$ when the L_1-penalty is used. Similar to Theorem 3, $\log p_n$ can also be as large as $o(n)$, as long as $s_{n3}\log p_n/n = o(1)$. It is evident that normalizing with $\hat{\mathbf{W}}$ results in an improvement in the rate of convergence in operator norm.

Theorem 12 *(Sparsistency). Under the conditions in Theorem 11, for any local minimizer* $\hat{\mathbf{T}}$ *of (4.4) satisfying* $\|\hat{\mathbf{T}} - \mathbf{T}_0\|_F^2 = O_P(s_{n3}\log p_n/n)$ *if* $\log p_n/n + \eta_n = O(\lambda_{n5}^2)$*, then sparsistency holds for* $\hat{\mathbf{T}}$*, provided that* $\|\hat{\mathbf{T}} - \mathbf{T}_0\|^2 = O(\eta_n)$ *for some* $\eta_n \to 0$*.*

Proof is omitted since it goes exactly the same as that of Theorem 10. The above results apply also to the L_1-penalty. For simultaneous persistency and optimal rate of convergence using L_1-penalty, the biases inherent in L_1-penalty induce the restriction $s_{n3} = O(1)$ in the worst scenario where $\eta_n^2 = s_{n3}\log p_n/n$, and $s_{n3} = O(p_n)$ in the optimistic scenario where $\eta_n^2 = s_{n3}/p_n \cdot \log p_n/n$. This restriction does not apply to the SCAD and other asymptotically unbiased penalty functions.

5 Proofs

We first prove two lemmas. The first one concerns with inequalities involving operator and Frobenius norms. The other one concerns with order estimation for elements in a matrix of the form $\mathbf{A}(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}$, which is useful in proving results concerning sparsistency.

Lemma 1 *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be real matrices such that the product* $\mathbf{AB}$ *is defined. Then, defining* $\|\mathbf{A}\|_{\min}^2 = \lambda_{\min}(\mathbf{A}^T\mathbf{A})$*, we have*

$$\|\mathbf{A}\|_{\min}\|\mathbf{B}\|_F \leq \|\mathbf{AB}\|_F \leq \|\mathbf{A}\|\|\mathbf{B}\|_F. \tag{5.1}$$

In particular, if $\mathbf{A} = (a_{ij})$*, then* $|a_{ij}| \leq \|\mathbf{A}\|$ *for each* i, j*.*

Proof of Lemma 1. Write $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_q)$, where $\mathbf{b}_i$ is the i-th column vector in $\mathbf{B}$. Then

$$\begin{aligned}\|\mathbf{AB}\|_F^2 = \mathrm{tr}(\mathbf{B}^T\mathbf{A}^T\mathbf{AB}) = \sum_{i=1}^{q} \mathbf{b}_i^T\mathbf{A}^T\mathbf{A}\mathbf{b}_i &\leq \lambda_{\max}(\mathbf{A}^T\mathbf{A})\sum_{i=1}^{q}\|\mathbf{b}_i\|^2 \\ &= \|\mathbf{A}\|^2\|\mathbf{B}\|_F^2.\end{aligned}$$

Similarly,

$$\begin{aligned}\|\mathbf{AB}\|_F^2 = \sum_{i=1}^{q} \mathbf{b}_i^T\mathbf{A}^T\mathbf{A}\mathbf{b}_i &\geq \lambda_{\min}(\mathbf{A}^T\mathbf{A})\sum_{i=1}^{q}\|\mathbf{b}_i\|^2 \\ &= \|\mathbf{A}\|_{\min}^2\|\mathbf{B}\|_F^2,\end{aligned}$$

which completes the proof of (5.1). To prove $|a_{ij}| \leq \|\mathbf{A}\|$, note that $a_{ij} = \mathbf{e}_i^T\mathbf{A}\mathbf{e}_j$, where $\mathbf{e}_i$ is the unit column vector with one at the i-th position, and zero elsewhere. Hence using (5.1),

$$|a_{ij}| = |\mathbf{e}_i^T\mathbf{A}\mathbf{e}_j| \leq \|\mathbf{A}\mathbf{e}_j\|_F \leq \|\mathbf{A}\| \cdot \|\mathbf{e}_j\|_F = \|\mathbf{A}\|,$$

and this completes the proof of the lemma. □

Lemma 2 *Let* $\mathbf{S}$ *be a sample covariance matrix of a random sample* $\{\mathbf{y}_i\}_{1\leq i\leq n}$ *with* $\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$*. Assume* $p_n/n = o(1)$*,* $\mathbf{\Sigma}_0$ *has eigenvalues uniformly bounded above as* $n \to \infty$*, and* $\mathbf{A} = \mathbf{A}_0 + \Delta_1$*,* $\mathbf{B} = \mathbf{B}_0 + \Delta_2$ *are matrices such that the constant matrices* $\|\mathbf{A}_0\| = O(1)$ *and* $\|\mathbf{B}_0\| = O(1)$ *independent of the data, with* $\|\Delta_1\|, \|\Delta_2\| = o_P(1)$*. Then* $\max_{i,j}|(\mathbf{A}(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B})_{ij}| = O_P(\{\log p_n/n\}^{1/2})$*.*

Proof of Lemma 2. We first prove the lemma with $\mathbf{A}$ and $\mathbf{B}$ independent of the data. Let $\mathbf{x}_i = \mathbf{A}\mathbf{y}_i$ and $\mathbf{w}_i = \mathbf{B}^T\mathbf{y}_i$. Define $\mathbf{u}_i = (\mathbf{x}_i^T, \mathbf{w}_i^T)^T$, with covariance matrix

$$\mathbf{\Sigma}_\mathbf{u} = \mathrm{var}(\mathbf{u}_i) = \begin{pmatrix} \mathbf{A}\mathbf{\Sigma}_0\mathbf{A}^T & \mathbf{A}\mathbf{\Sigma}_0\mathbf{B} \\ \mathbf{B}^T\mathbf{\Sigma}_0\mathbf{A}^T & \mathbf{B}^T\mathbf{\Sigma}_0\mathbf{B} \end{pmatrix}.$$

Since $\|(\mathbf{A}^T \quad \mathbf{B})^T\| \leq (\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2)^{1/2} = O(1)$ and $\|\mathbf{\Sigma}_0\| = O(1)$ uniformly, we have $\|\mathbf{\Sigma}_\mathbf{u}\| = O(1)$ uniformly. Then, with $\mathbf{S}_\mathbf{u} = n^{-1}\sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^T$, which is the sample covariance matrix for the random sample $\{\mathbf{u}_i\}_{1\leq i\leq n}$, by lemma 3 of , we have

$$\max_{i,j} |(\mathbf{S}_\mathbf{u} - \mathbf{\Sigma}_\mathbf{u})_{ij}| = O_P(\{\log p_n/n\}^{1/2}).$$

In particular, it means that

$$\max_{i,j} |(\mathbf{A}(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B})_{ij}| = \left(n^{-1}\sum_{r=1}^n \mathbf{x}_r\mathbf{w}_r^T - \mathbf{A}\mathbf{\Sigma}_0\mathbf{B}\right)_{ij} = O_P(\{\log p_n/n\}^{1/2}),$$

which completes the proof for $\mathbf{A}$ and $\mathbf{B}$ independent of the data.

Now consider $\mathbf{A} = \mathbf{A}_0 + \Delta_1$, $\mathbf{B} = \mathbf{B}_0 + \Delta_2$ as in the statement of the lemma. Then

$$\mathbf{A}(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B} = K_1 + K_2 + K_3 + K_4, \tag{5.2}$$

where $K_1 = \mathbf{A}_0(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0$, $K_2 = \Delta_1(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0$, $K_3 = \mathbf{A}_0(\mathbf{S} - \mathbf{\Sigma}_0)\Delta_2$ and $K_4 = \Delta_1(\mathbf{S} - \mathbf{\Sigma}_0)\Delta_2$. Now $\max_{i,j} |(K_1)_{ij}| = O_P(\{\log p_n/n\}^{1/2})$ as proved before. Consider K_2. Suppose the maximum element of the matrix is at the (i,j)-th position. Then we can set

$$\Delta_1 = c_n\mathbf{B}_0^T(\mathbf{S} - \mathbf{\Sigma}_0)^T = c_n\mathbf{B}_0^T(\mathbf{S} - \mathbf{\Sigma}_0),$$

where $c_n^2 = o(n/p_n)$ to keep $\|\Delta_1\| = o_P(1)$ since $\|\mathbf{B}_0\| = O(1)$ and $\|\mathbf{S} - \mathbf{\Sigma}_0\|^2 = O_P(p_n/n)$ (see chapter 2 of Bai and Silverstein (2006)), so the maximum element is now on the diagonal, with

$$\max_{i,j} |(\Delta_1(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0)_{ij}| \leq c_n \max_k |(\mathbf{B}_0^T(\mathbf{S} - \mathbf{\Sigma}_0)^2\mathbf{B}_0)_{kk}|. \tag{5.3}$$

Since $\mathbf{S} - \mathbf{\Sigma}_0$ is symmetric, we can find a rotation matrix $\mathbf{Q}$ (i.e. $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = I$) so that

$$\mathbf{S} - \mathbf{\Sigma}_0 = \mathbf{Q}\Lambda\mathbf{Q}^T,$$

where Λ is a diagonal matrix with real entries. Then since $c_n\|\Lambda\|^2 = o_P(\{p_n/n\}^{1/2})$ but $\|\Lambda\| = O_P(\{p_n/n\}^{1/2})$, we have

$$\begin{aligned} c_n \max_k |(\mathbf{B}_0^T(\mathbf{S} - \mathbf{\Sigma}_0)^2\mathbf{B}_0)_{kk}| &= \max_k |(\mathbf{B}_0^T\mathbf{Q}c_n\Lambda^2\mathbf{Q}^T\mathbf{B}_0)_{kk}| \\ &\le \max_k |(\mathbf{B}_0^T\mathbf{Q}\Lambda\mathbf{Q}^T\mathbf{B}_0)_{kk}| \\ &= \max_k |(\mathbf{B}_0^T(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0)_{kk}| = O_P(\{\log p_n/n\}^{1/2}), \end{aligned}$$

where the last line used the previous proof for constant matrix $\mathbf{B}_0$. Then combining with (5.3), we have $\max_{i,j} |(K_2)_{ij}| = O_P(\{\log p_n/n\}^{1/2})$. Similar arguments goes for K_3. For K_4, similar arguments hold and we will end up setting

$$\Delta_1 = c_n(\mathbf{S} - \mathbf{\Sigma}_0), \;\; \Delta_2 = d_n(\mathbf{S} - \mathbf{\Sigma}_0)^2,$$

where $c_n^2 = o(n/p_n)$ and $d_n = o(n/p_n)$ to keep $\|\Delta_1\|, \|\Delta_2\| = o_P(1)$. Then we have $c_n d_n\|\Lambda\|^4 = o_P(\{p_n/n\}^{1/2})$, and

$$\begin{aligned} \max_{i,j} |(K_4)_{ij}| \le c_n d_n \max_k |[(\mathbf{S} - \mathbf{\Sigma}_0)^4]_k| &= \max_k |[\mathbf{Q}c_n d_n\Lambda^4\mathbf{Q}^T]_k| \\ &\le \max_k |(\mathbf{Q}\Lambda\mathbf{Q}^T)_k| = \max_k |(\mathbf{S} - \mathbf{\Sigma}_0)_k| \\ &= O_P(\{\log p_n/n\}^{1/2}). \end{aligned}$$

This completes the proof of the lemma. □

Proof of Theorem 1. Let U be a symmetric matrix of size p_n, $\mathbf{D}_U$ be its diagonal matrix and $\mathbf{R}_U = U - \mathbf{D}_U$ be its off-diagonal matrix. Set $\Delta_U = \alpha_n\mathbf{R}_U + \beta_n\mathbf{D}_U$. We

would like to show that, for $\alpha_n = (s_{n1} \log p_n/n)^{1/2}$ and $\beta_n = (p_n \log p_n/n)^{1/2}$, and for a set $\mathcal{A}$ defined as $\mathcal{A} = \{U : \|\mathbf{R}_U\|_F = C_1, \|\mathbf{D}_U\|_F = C_2\}$,

$$P\Big(\inf_{U \in \mathcal{A}} q_1(\mathbf{\Omega}_0 + \Delta_U) > q_1(\mathbf{\Omega}_0) \Big) \to 1,$$

for sufficiently large constants C_1 and C_2. This implies that there is a local minimizer in $\{\mathbf{\Omega}_0 + \Delta_U : \|\mathbf{R}_U\|_F = C_1, \|\mathbf{D}_U\|_F = C_2\}$ such that $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_F = O_P(\alpha_n + \beta_n)$.

Consider, for $\mathbf{\Sigma} = \mathbf{\Sigma}_0 + \Delta_U$, the difference

$$q_1(\mathbf{\Omega}) - q_1(\mathbf{\Omega}_0) = I_1 + I_2 + I_3,$$

where

$$\begin{aligned}
I_1 &= \mathrm{tr}(\mathbf{S}\mathbf{\Omega}) - \log|\mathbf{\Omega}| - (\mathrm{tr}(\mathbf{S}\mathbf{\Omega}_0) - \log|\mathbf{\Omega}_0|), \\
I_2 &= \sum_{(i,j) \in S_1^c} (p_{\lambda_{n1}}(|\omega_{ij}|) - p_{\lambda_{n1}}(|\omega_{ij}^0|)), \\
I_3 &= \sum_{(i,j) \in S_1, i \neq j} (p_{\lambda_{n1}}(|\omega_{ij}|) - p_{\lambda_{n1}}(|\omega_{ij}^0|)).
\end{aligned}$$

It suffice to show that the difference is positive asymptotically with probability tending to 1. Using Taylor's expansion with the integral remainder, we have $I_1 = K_1 + K_2$, where

$$\begin{aligned}
K_1 &= \mathrm{tr}((\mathbf{S} - \mathbf{\Sigma}_0)\Delta_U), \\
K_2 &= \mathrm{vec}(\Delta_U)^T \Big\{ \int_0^1 g(v, \mathbf{\Omega}_v)(1-v)dv \Big\} \mathrm{vec}(\Delta_U),
\end{aligned} \tag{5.4}$$

with the definitions $\mathbf{\Omega}_v = \mathbf{\Omega}_0 + v\Delta_U$, and $g(v, \mathbf{\Omega}_v) = \mathbf{\Omega}_v^{-1} \otimes \mathbf{\Omega}_v^{-1}$. Now

$$\begin{aligned}
K_2 &\geq \int_0^1 (1-v) \min_{0 \leq v \leq 1} \lambda_{\min}(\mathbf{\Omega}_v^{-1} \otimes \mathbf{\Omega}_v^{-1}) dv \cdot \|\mathrm{vec}(\Delta_U)\|^2 \\
&= \|\mathrm{vec}(\Delta_U)\|^2/2 \cdot \min_{0 \leq v \leq 1} \lambda_{\max}^{-2}(\mathbf{\Omega}_v) \geq \|\mathrm{vec}(\Delta_U)\|^2/2 \cdot (\|\mathbf{\Omega}_0\| + \|\Delta_U\|)^{-2} \\
&\geq (C_1^2 \alpha_n^2 + C_2^2 \beta_n^2)/2 \cdot (\tau_1^{-1} + o(1))^{-2},
\end{aligned}$$

where we used $\|\Delta_U\| = o(1)$.

Consider K_1. It is clear that $|K_1| \leq L_1 + L_2$, where

$$L_1 = \Big| \sum_{(i,j)\in S_1} (\mathbf{S} - \mathbf{\Sigma}_0)_{ij} (\Delta_U)_{ij} \Big|,$$
$$L_2 = \Big| \sum_{(i,j)\in S_1^c} (\mathbf{S} - \mathbf{\Sigma}_0)_{ij} (\Delta_U)_{ij} \Big|.$$

Using Lemma 1 and 2, we have

$$\begin{aligned} L_1 &\leq (s_{n1} + p_n)^{1/2} \max_{i,j} |(\mathbf{S} - \mathbf{\Sigma}_0)_{ij}| \cdot \|\Delta_U\|_F \\ &\leq O_P(\alpha_n + \beta_n) \cdot \|\Delta_U\|_F \\ &= O_P(C_1 \alpha_n^2 + C_2 \beta_n^2), \end{aligned}$$

This is dominated by K_2 when C_1 and C_2 are sufficiently large.

Now, consider $I_2 - L_2$. Since we assumed $(s_{n1}+1)\log p_n/n = O(\lambda_{n1}^2)$, by condition (C), when n is sufficiently large, we have $\alpha_n = O(\lambda_{n1})$ and $p_{\lambda_{n1}}(|\alpha_n u_{ij}|) \geq \lambda_{n1} k_1 |\alpha_n u_{ij}|$ for some positive constant k_1. Using $p_{\lambda_{n1}}(0) = 0$, we then have

$$I_2 = \sum_{(i,j)\in S_1^c} p_{\lambda_{n1}}(|\alpha_n u_{ij}|) \geq k_1 \lambda_{n1} \alpha_n \sum_{(i,j)\in S_1^c} |u_{ij}|.$$

Hence

$$\begin{aligned} I_2 - L_2 &\geq \sum_{(i,j)\in S_1^c} \big\{ \lambda_{n1} k_1 |\alpha_n u_{ij}| - |(\mathbf{S} - \mathbf{\Sigma}_0)_{ij}| \cdot |\alpha_n u_{ij}| \big\} \\ &\geq \sum_{(i,j)\in S_1^c} \big[\lambda_{n1} k_1 - O_P(\{\log p_n/n\}^{1/2}) \big] \cdot |\alpha_n u_{ij}| \\ &= \lambda_{n1} \alpha_n \sum_{(i,j)\in S_1^c} \big[k_1 - O_P(\lambda_{n1}^{-1} \{\log p_n/n\}^{1/2}) \big] \cdot |u_{ij}|. \end{aligned}$$

With the assumption that $(s_{n1}+1)\log p_n/n = O(\lambda_{n1}^2)$, we see from the above that $I_2 - L_2 \geq 0$ since $O_P(\lambda_{n1}^{-1}\{\log p_n/n\}^{1/2}) = o_P(1)$.

Now, with L_1 dominated by K_2 and $I_2 - L_2 \geq 0$, the proof completes if we can show that I_3 is also dominated by K_2, since we have proved that $K_2 > 0$. Using Taylor's expansion, we can arrive at

$$|I_3| \leq C_1 \alpha_n s_{n1}^{1/2} a_{n1} + C_1^2 b_{n1} \alpha_n^2 / 2 \cdot (1 + o(1)).$$

By condition (B), we have

$$|I_3| = C \cdot O(\alpha_n^2 + \beta_n^2) + C^2 \cdot o(\alpha_n^2),$$

which is dominated by K_2 with large enough constants C_1 and C_2. This completes the proof of the theorem. □

Proof of Theorem 2. For $\mathbf{\Omega}$ a minimizer of (1.1), the derivative for $q_1(\mathbf{\Omega})$ w.r.t. ω_{ij} for $(i, j) \in S_2^c$ is

$$\frac{\partial q_1(\mathbf{\Omega})}{\partial \omega_{ij}} = 2(s_{ij} - \sigma_{ij} + p'_{\lambda_{n1}}(|\omega_{ij}|)\mathrm{sgn}(\omega_{ij})),$$

where $\mathrm{sgn}(a)$ denotes the sign of a. We need to estimate the order of $s_{ij} - \sigma_{ij}$ independent of i and j.

Decompose $s_{ij} - \sigma_{ij} = I_1 + I_2$, where

$$I_1 = s_{ij} - \sigma_{ij}^0, \;\; I_2 = \sigma_{ij}^0 - \sigma_{ij}.$$

By Lemma 2 or Lemma 3 of Bickel and Levina (2006), it follows that $\max_{i,j} |I_1| = O_P(\{\log p_n / n\}^{1/2})$. It remains to estimate the order of I_2.

By Lemma 1, $|\sigma_{ij} - \sigma_{ij}^0| \leq \|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|$, which has order

$$\begin{aligned} \|\mathbf{\Sigma} - \mathbf{\Sigma}_0\| &= \|\mathbf{\Sigma}(\mathbf{\Omega} - \mathbf{\Omega}_0)\mathbf{\Sigma}_0\| \\ &\leq \|\mathbf{\Sigma}\| \cdot \|\mathbf{\Omega} - \mathbf{\Omega}_0\| \cdot \|\mathbf{\Sigma}_0\| \\ &= O(\|\mathbf{\Omega} - \mathbf{\Omega}_0\|), \end{aligned}$$

where we used Condition (A) to get $\|\mathbf{\Sigma}_0\| = O(1)$, and using $\eta_n \to 0$ so that $\lambda_{\min}(\mathbf{\Omega} - \mathbf{\Omega}_0) = o(1)$ for $\|\mathbf{\Omega} - \mathbf{\Omega}_0\| = O(\eta_n^{1/2})$,

$$\begin{aligned}\|\mathbf{\Sigma}\| = \lambda_{\min}^{-1}(\mathbf{\Omega}) &\leq (\lambda_{\min}(\mathbf{\Omega}_0) + \lambda_{\min}(\mathbf{\Omega} - \mathbf{\Omega}_0))^{-1} \\ &= (O(1) + o(1))^{-1} = O(1).\end{aligned}$$

Hence $\|\mathbf{\Omega} - \mathbf{\Omega}_0\| = O(\eta_n^{1/2})$ implies $|I_2| = O(\eta_n^{1/2})$.

Combining the last two results yields that

$$\begin{aligned}\max_{i,j} |s_{ij} - \sigma_{ij}| &= O_P(|s_{ij} - \sigma_{ij}^0| + \eta_n^{1/2}) \\ &= O_P(\{\log p_n/n\}^{1/2} + \eta_n^{1/2}).\end{aligned}$$

By conditions (C) and (D), we have

$$p'_{\lambda_{n1}}(|\omega_{ij}|) = C_3\lambda_{n1}$$

for ω_{ij} in a small neighborhood of 0 (excluding 0 itself) and some positive constant C_3. Hence if ω_{ij} lies in a small neighborhood of 0, we need to have $\log p_n/n + \eta_n = O(\lambda_{n1}^2)$ in order to have the sign of $\partial q_1(\mathbf{\Omega})/\partial\omega_{ij}$ depends on $\text{sgn}(\omega_{ij})$ only with probability tending to 1. The proof of the theorem is completed. □

Proof of Theorem 3. Because of the similarity between equations (2.2) and (1.1), the Frobenius norm result has nearly identical proof as Theorem 1, except that we now set $\Delta_U = \alpha_n U$. For the operator norm result, we refer readers to the proof of Theorem 2 of Rothman *et al.* (2007). □

Proof of Theorem 5. The proof is similar to that of Theorem 1. We only sketch briefly the proof, pointing out the important differences.

Let $\alpha_n = (s_{n2}\log p_n/n)^{1/2}$ and $\beta_n = (p_n \log p_n/n)^{1/2}$, and define $\mathcal{A} = \{U : \|\mathbf{R}_U\|_F = C_1, \|\mathbf{D}_U\|_F = C_2\}$. Want to show

$$P\Big(\inf_{U\in\mathcal{A}} q_2(\mathbf{\Sigma}_0 + \Delta_U) > q_2(\mathbf{\Sigma}_0)\Big) \to 1,$$

for sufficiently large constants C_1 and C_2.

For $\mathbf{\Sigma} = \mathbf{\Sigma}_0 + \Delta_U$, the difference

$$q_2(\mathbf{\Sigma}) - q_2(\mathbf{\Sigma}_0) = I_1 + I_2 + I_3,$$

where

$$\begin{aligned}
I_1 &= \mathrm{tr}(S\mathbf{\Omega}) + \log|\mathbf{\Sigma}| - (\mathrm{tr}(S\mathbf{\Omega}_0) + \log|\mathbf{\Sigma}_0|),\\
I_2 &= \sum_{(i,j)\in S_2^c} (p_{\lambda_{n2}}(|\sigma_{ij}|) - p_{\lambda_{n2}}(|\sigma_{ij}^0|)),\\
I_3 &= \sum_{(i,j)\in S_2, i\neq j} (p_{\lambda_{n2}}(|\sigma_{ij}|) - p_{\lambda_{n2}}(|\sigma_{ij}^0|)),
\end{aligned}$$

with $I_1 = K_1 + K_2$, where

$$\begin{aligned}
K_1 &= -\mathrm{tr}((\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{\Omega}_0\Delta_U\mathbf{\Omega}_0) = -\mathrm{tr}((\mathbf{S}_{\mathbf{\Omega}_0} - \mathbf{\Omega}_0)\Delta_U),\\
K_2 &= \mathrm{vec}(\Delta_U)^T\Big\{\int_0^1 g(v,\mathbf{\Sigma}_v)(1-v)dv,\Big\}\mathrm{vec}(\Delta_U),
\end{aligned} \tag{5.5}$$

and $\mathbf{\Sigma}_v = \mathbf{\Sigma}_0 + v\Delta_U$, $\mathbf{S}_{\mathbf{\Omega}_0}$ is the sample covariance matrix of a random sample $\{\mathbf{x}_i\}_{1\le i\le n}$ having $\mathbf{x}_i \sim N(\mathbf{0}, \mathbf{\Omega}_0)$. Also,

$$g(v,\mathbf{\Sigma}_v) = \mathbf{\Sigma}_v^{-1}\otimes\mathbf{\Sigma}_v^{-1}\mathbf{S}\mathbf{\Sigma}_v^{-1} + \mathbf{\Sigma}_v^{-1}\mathbf{S}\mathbf{\Sigma}_v^{-1}\otimes\mathbf{\Sigma}_v^{-1} - \mathbf{\Sigma}_v^{-1}\otimes\mathbf{\Sigma}_v^{-1}. \tag{5.6}$$

The treatment of K_2 is different from that in Theorem 1. By condition (A), we have

$$\|v\Delta_U\Omega_0\| \le \|\Delta_U\|\|\mathbf{\Omega}_0\| \le \tau_1^{-1}(C_1\alpha_n + C_2\beta_n) = o(1).$$

Thus, we can use the Neumann series expansion to arrive at

$$\mathbf{\Sigma}_v^{-1} = \mathbf{\Omega}_0(I + v\Delta_U\mathbf{\Omega}_0)^{-1} = \mathbf{\Omega}_0(I - v\Delta_U\mathbf{\Omega}_0 + o(1)).$$

That is, $\mathbf{\Sigma}_v^{-1} = \mathbf{\Omega}_0 + O_P(\alpha_n + \beta_n)$, and $\|\mathbf{\Sigma}_v^{-1}\| = \tau_1^{-1} + O_P(\alpha_n + \beta_n)$. With $\mathbf{S}_I$ defined as the sample covariance matrix formed from a random sample $\{\mathbf{x}_i\}_{1\le i\le n}$ having $\mathbf{x}_i \sim N(\mathbf{0}, I)$,

$$\|\mathbf{S} - \mathbf{\Sigma}_0\| = O_P(\|\mathbf{S}_I - I\|) = O_P(\{p_n/n\}^{1/2})$$

(see e.g. chapter 2 of Bai and Silverstein (2006)). These entail

$$\begin{aligned}\mathbf{S}\mathbf{\Sigma}_v^{-1} &= (\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{\Sigma}_v^{-1} + \mathbf{\Sigma}_0\mathbf{\Sigma}_v^{-1} \\ &= O_P(\{p_n/n\}^{1/2}) + I + O_P(\alpha_n + \beta_n) \\ &= I + o_P(1).\end{aligned}$$

Combining these results, we have

$$g(v, \mathbf{\Sigma}_v) = \mathbf{\Omega}_0 \otimes \mathbf{\Omega}_0 + O_P(\alpha_n + \beta_n).$$

Consequently,

$$\begin{aligned}K_2 &= \mathrm{vec}(\Delta_U)^T\Big\{\int_0^1 \mathbf{\Omega}_0 \otimes \mathbf{\Omega}_0(1 + o_P(1))(1 - v)dv\Big\}\mathrm{vec}(\Delta_U) \\ &\ge \lambda_{\min}(\mathbf{\Omega}_0 \otimes \mathbf{\Omega}_0)\|\mathrm{vec}(\Delta_U)\|^2/2 \cdot (1 + o_P(1)) \\ &= \tau_1^{-2}(C_1^2\alpha_n^2 + C_2^2\beta_n^2)/2 \cdot (1 + o_P(1)).\end{aligned}$$

All other terms are dealt with similarly as in the proof of Theorem 1, and hence we omit them. □

Proof of Theorem 6. The proof is similar to that of Theorem 2. We only show the main differences.

It is easy to show

$$\frac{\partial q_2(\mathbf{\Sigma})}{\partial \sigma_{ij}} = 2(-(\mathbf{\Omega S \Omega})_{ij} + \omega_{ij} + p'_{\lambda_n}(|\sigma_{ij}|)\mathrm{sgn}(\sigma_{ij})).$$

Our aim is to estimate the order of $|(-\mathbf{\Omega S \Omega} + \mathbf{\Omega})_{ij}|$, finding an upper bound which is independent of both i and j.

Write

$$-\mathbf{\Omega S \Omega} + \mathbf{\Omega} = I_1 + I_2,$$

where $I_1 = -\mathbf{\Omega}(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{\Omega}$ and $I_2 = \mathbf{\Omega}(\mathbf{\Sigma} - \mathbf{\Sigma}_0)\mathbf{\Omega}$. Since

$$\begin{aligned} \|\mathbf{\Omega}\| = \lambda_{\min}^{-1}(\mathbf{\Sigma}) &\leq (\lambda_{\min}(\mathbf{\Sigma}_0) + \lambda_{\min}(\mathbf{\Sigma} - \mathbf{\Sigma}_0))^{-1} \\ &= \tau_1^{-1} + o(1), \end{aligned}$$

we have

$$\mathbf{\Omega} = \mathbf{\Omega}_0 + (\mathbf{\Omega} - \mathbf{\Omega}_0) = \mathbf{\Omega}_0 - \mathbf{\Omega}(\mathbf{\Sigma} - \mathbf{\Sigma}_0)\mathbf{\Omega}_0 = \mathbf{\Omega}_0 + \Delta,$$

where $\|\Delta\| \leq \|\mathbf{\Omega}\| \cdot \|\mathbf{\Sigma} - \mathbf{\Sigma}_0\| \cdot \|\mathbf{\Omega}_0\| = O(\eta_n^{1/2}) = o(1)$ by Lemma 1, with $\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|^2 = O(\eta_n)$. Hence we can apply Lemma 2 and conclude that $\max_{i,j} |(I_1)_{ij}| = O_P(\{\log p_n/n\}^{1/2})$.

For I_2, we have

$$\max_{i,j} |(I_2)_{ij}| \leq \|\mathbf{\Omega}\| \cdot \|\mathbf{\Sigma} - \mathbf{\Sigma}_0\| \cdot \|\mathbf{\Omega}\| = O(\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|) = O(\eta_n^{1/2}).$$

Hence we have

$$\max_{i,j} |(-\mathbf{\Omega S \Omega} + \mathbf{\Omega})_{ij}| = O(\{\log p_n/n\}^{1/2} + \eta_n^{1/2}).$$

The rest goes similar to the proof of Theorem 2, and is omitted. □

Proof of Theorem 7. The proof is nearly identical to that of Theorem 5, except that we now set $\Delta_U = \alpha_n U$. The fact that $(\hat{\mathbf{\Gamma}}_{\mathbf{S}})_{ii} = 1 = \gamma^0_{ii}$ has no estimation error eliminates an order $(p_n \log p_n/n)^{1/2}$ that contributes from estimating $\mathrm{tr}((\hat{\mathbf{\Gamma}}_{\mathbf{S}} - \mathbf{\Gamma}_0)\mathbf{\Psi}_0\Delta_U\mathbf{\Psi}_0)$ for (3.2). This is why we can estimate more accurately for the sparse correlation.

For the operator norm result, we refer readers to the proof of Theorem 2 of Rothman *et al.* (2007). □

Proof of Theorem 10. For ($\mathbf{T}$,$\mathbf{D}$) a minimizer of (4.2), the derivative for $q_3(\mathbf{T}, \mathbf{D})$ w.r.t. t_{ij} for $(i, j) \in S_3^c$ is

$$\frac{\partial q_3(\mathbf{T}, \mathbf{D})}{\partial t_{ij}} = 2((\mathbf{S}\mathbf{T}^T\mathbf{D}^{-1})_{ji} + p'_{\lambda_{n3}}(|t_{ij}|)\mathrm{sgn}(t_{ij})).$$

Now $\mathbf{S}\mathbf{T}^T\mathbf{D}^{-1} = I_1 + I_2 + I_3 + I_4$, where

$$I_1 = (\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{T}^T\mathbf{D}^{-1}, \qquad I_2 = \mathbf{\Sigma}_0(\mathbf{T} - \mathbf{T}_0)^T\mathbf{D}^{-1},$$
$$I_3 = \mathbf{\Sigma}_0\mathbf{T}_0^T(\mathbf{D}^{-1} - \mathbf{D}_0^{-1}), \quad I_4 = \mathbf{\Sigma}_0\mathbf{T}_0^T\mathbf{D}_0^{-1}.$$

By the MCD (4.1), $I_4 = \mathbf{T}_0^{-1}$. Since $i > j$ for $(i, j) \in S_3^c$, we must have $(\mathbf{T}_0^{-1})_{ji} = 0$. Hence we can ignore I_4.

Since $\|\mathbf{T} - \mathbf{T}_0\|^2 = O(\eta_n)$ and $\|\mathbf{D} - \mathbf{D}_0\|^2 = O(\zeta_n)$ with $\eta_n, \zeta_n = o(1)$, and by condition (A) we can easily show $\|\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\| = O(\|\mathbf{D} - \mathbf{D}_0\|) = O(\zeta_n^{1/2})$. Then we can apply Lemma 2 to show that $\max_{ij} |(I_1)_{ij}| = (\log p_n/n)^{1/2}$.

For I_2, we have $\max_{ij} |(I_2)_{ij}| \le \|\mathbf{\Sigma}_0\| \cdot \|\mathbf{T} - \mathbf{T}_0\| \cdot \|\mathbf{D}^{-1}\| = O(\eta_n^{1/2})$. And finally, $\max_{ij} |(I_3)_{ij}| \le \|\mathbf{\Sigma}_0\| \cdot \|\mathbf{T}_0\| \cdot \|\mathbf{D}^{-1} - \mathbf{D}_0^{-1}\| = O(\zeta_n^{1/2})$.

With all these, we have $\max_{(i,j)\in S_3^c} |(\mathbf{S}\mathbf{T}^T\mathbf{D}^{-1})_{ji}|^2 = \log p_n/n + \eta_n + \zeta_n$. The rest of the proof goes like that of Theorem 2 or 6. □

References

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

[2] Bickel, P.J. and Levina, E. (2008), Covariance Regularization by Thresholding, to appear in *Ann. Statist.*

[3] Bickel, P.J. and Levina, E. (2008), Regularized Estimation of Large Covariance Matrices, *Ann. Statist.*, **36(1)**, 199–227.

[4] d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008), First-order Methods For Sparse Covariance Selection, *SIAM. J. Matrix Anal. and Appl.*, **30(1)**, 56–66.

[5] Dempster, A.P. (1972), Covariance Selection, *Biometrics*, **28**, 157–175.

[6] Diggle, P. and Verbyla, A. (1998), Nonparametric Estimation of Covariance Structure in Longitudinal Data, *Biometrics*, **54(2)**, 401–415.

[7] El Karoui, N. (2007). Operator Norm Consistent Estimation of a Large Dimensional Sparse Covariance Matrices. *Technical report 734*, Department of Statistics, UC-Berkeley.

[8] Fan, J., Feng, Y. and Wu, Y. (2008). Network Exploration via the Adaptive LASSO and SCAD Penalties. *Manuscript.*

[9] Fan, J. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

[10] Fan, J. and Peng, H. (2004), Nonconcave Penalized Likelihood With a Diverging Number of Parameters, *Ann. Statist.*, **32**, 928–961.

[11] Huang, J., Liu, N., Pourahmadi, M. and Liu, L. (2006), Covariance Matrix Selection and Estimation via Penalised Normal Likelihood, *Biometrika*, **93(1)**, 85–98.

[12] Levina, E., Rothman, A.J. and Zhu, J. (2008), Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty, *Ann. Applied Statist.*, **2(1)**, 245–263.

[13] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.

[14] Pourahmadi, M. (1999), Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation, *Biometrika*, **86**, 677–690.

[15] Smith, M. and Kohn, R. (2002), Parsimonious Covariance Matrix Estimation for Longitudinal Data, *J. Amer. Statist. Assoc.*, **97(460)**, 1141–1153.

[16] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2008). *Sparse additive models. Manuscript.*

[17] Rothman, A.J., Bickel, P.J., Levina, E. and Zhu, J. (2007), Sparse Permutation Invariant Covariance Estimation, Technical report No. 467, Dept. of Statistics, Univ. of Michigan.

[18] Wagaman, A.S. and Levina, E. (2007). Discovering sparse covariance structures with the Isomap, to appear in the Journal of Computational and Graphical Statistics.

[19] Wong, F., Carter, C. and Kohn, R. (2003). Efficient Estimation of Covariance Selection Models, *Biometrika*, **90**, 809–830.

[20] Wu, W.B. and Pourahmadi, M. (2003), Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data, *Biometrika*, **94**, 1–17.

[21] Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model, *Biometrika*, **90**, 831–844.

[22] Zhang, C.H. (2007). Penalized Linear Unbiased Selection. *Manuscript*.

[23] Zhao, P. and Yu, B. (2006), On Model Selection Consistency of Lasso, Technical Report, Statistics Department, UC-Berkeley.

[24] Zou, H. and Li, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models (With Discussion). *Ann. Statist.*, **36(4)**, 1509–1533.