# Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation [*]

By Clifford Lam and Jianqing Fan
Department of Operations Research and Financial Engineering
Princeton University, Princeton, NJ, 08544

This paper studies the sparsistency, rates of convergence, and asymptotic normality for estimating sparse covariance 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, or 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 nonsparse 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 obtain the sparsistency and optimal rate of convergence, the non-sparsity rates must be low: $s_n' = O(p_n^{1/2})$ among $O(p_n^2)$ parameters, for estimating sparse covariance matrix, or sparse precision matrix or sparse Cholesky factor and $s_n' = O(1)$ for estimating sparse correlation matrix or its inverse, 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 are no such a restriction.

*Short Title*: Covariance Estimation with Penalization.

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

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

[*] Clifford Lam is PhD student, Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (email: wlam@princeton.edu); 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.

# 1 Introduction

Covariance matrix estimation is a common statistical problem that arises in many science 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 grows quickly with dimensionality. Depending on the case of applications, the sparsity of the covariance matrix or precision matrix is frequently imposed to strike a balance between biases and variances. For example, in longitudinal data analysis (see e.g. [6], or [2]), it is reasonable to assume that remote data in time are weakly correlated, whereas in Gaussian graphical models, the sparsity of the precision is a reasonable assumption ([5]).

This initiates a series of research focusing on the parsimony estimation of covariance matrix. [18] used prior which admit zeros on the off-diagonal elements of the Cholesky factor of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, while [21] used zero-admitting prior directly on the off-diagonal elements of $\mathbf{\Omega}$ to achieve parsimony. [22] 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. [2] 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. [10] 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 handle with 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. [13] used the LASSO on the off-diagonal elements of the Cholesky factor from MCD, while [15], [4] and [23] use different LASSO algorithms to select sparse elements in the precision matrix. A novel penalty called the nested Lasso was constructed in [14] to penalize on these off-diagonal elements. Thresholding the sample covariance matrix in high-dimensional setting was thoroughly studied by [7] and [3]. [20] proposed an Isomap method for discovering meaningful orderings of variables based on their correlations that result in block-diagonal or banded correlation structure, resulting an ISoband estimator. A permutation invariant estimator, called SPICE, was proposed in [19] 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. In particular, when the precision matrix is estimated, $s_n = p_n + s_{n2}$, where $p_n$ is the number of the diagonal elements and $s_{n2}$ is the number of the non-sparse elements of the off-diagonal entries. When the inverse of correlation matrix is estimated, $s_n$ is merely $s_{n2}$, since the diagonal elements of correlation matrices are known to be one. However, such rate of convergence does not address explicitly the sparsistency such as those in [9] and [25], the sampling distribution of nonsparse elements, nor the bias issues of the $L_1$-penalty. 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.

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

$$q_1(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) + \log|\mathbf{\Sigma}| + \sum_{i\neq j} p_{\lambda_{n1}}(|\sigma_{ij}|), \tag{1.1}$$

where $\mathbf{S} = n^{-1}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^T$ is the sample covariance matrix, with $\mathbf{\Sigma} = (\sigma_{ij})$, and $p_{\lambda_{n1}}(\cdot)$ is a penalty function, depending on a regularization parameter $\lambda_{n1}$, 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

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

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 $\theta$ is large, while the $L_1$-penalty grows linearly with $\theta$. See [9] for a detailed account of this and other advantages of such a penalty function.

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

$$q_2(\mathbf{\Omega}) = \mathrm{tr}(\mathbf{S}\mathbf{\Omega}) - \log|\mathbf{\Omega}| + \sum_{i\neq j} p_{\lambda_{n2}}(|\omega_{ij}|), \tag{1.3}$$

where we use $\omega_{ij}$ to denote the $(i,j)$-th element of the precision matrix $\mathbf{\Omega}$. Note that we only penalize on the off-diagonal elements of $\mathbf{\Sigma}$ or $\mathbf{\Omega}$ in the aforementioned two menthods, since the diagonal elements of $\mathbf{\Sigma}_0$ and $\mathbf{\Omega}_0$ do not vanish.

The computation of the non-convcave maximum likelihood problems can be solved by a sequence of penalized $L_1$-likelihood problem via local linear approximation ([26]). In fact, [26] shows that one iteration of such a procedure suffices as long as the initial values are good enough. See [8] for detailed implementations on the estimation of precision matrices. See also [24] for a general solution to the nonconvex penalized least-squares problem.

In studying sparse covariance or precision matrix, it is important to distinguish the diagonal and off-diagonal elements, since the diagonal elements always 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 (2.1) 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 3.2. This is an extension of the work of [19] using the $L_1$-penalty. Such an extension is important since the non-concave penalized likelihood ameliorates the bias problem of the penalized $L_1$-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 $\lambda_{ni}$ in the problems (1.1) and (1.3) to satisfy both consistency and sparsistency properties. This is in fact one of the motivations for introducing nonconvex penalty functions in [9] and [10], but we state and prove the explicit rates in the current context. In particular, we demonstrate that penalized $L_1$-likelihood can achieve simultaneously the optimal rate and sparsistency for estimation of $\mathbf{\Sigma}_0$ or $\Omega_0$ only when the number of nonsparse elements in off-diagonal entries are no larger than $O(p_n^{1/2})$. On the other hand, using the nonconvex penalty like SCAD or hard-thresholding penalty, such an extra restriction is not needed.

In this paper, apart from rates of convergence, we also develop the asymptotic normality of the resulting estimators, with rates for any regularization parameters specified. 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 $\lambda_{\min}(A)$, $\lambda_{\max}(A)$, and $\operatorname{tr}(A)$ 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 $\|B\| = \lambda_{\max}^{1/2}(B^T B)$ and $\|B\|_F = \operatorname{tr}^{1/2}(B^T B)$. We define the relation $a \succeq b$ if $b/a = O(1)$.

## 2 Estimation of sparse covariance matrix

We focus on analyzing the penalized likelihood method (1.1) for estimating sparse covariance 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 covariance matrix $\mathbf{\Sigma}_0$.

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

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

Note that for $L_1$-penalty, $a_{n1} = \max|\sigma_{i,j}^0|\lambda_n$ and $b_{n1} = 0$, whereas for SCAD, $a_{n1} = b_{n1} = 0$, for sufficiently large $n$ under the last assumption of condition (B).

We assume the following regularity conditions:

(A) There exists constants $\tau_1$ and $\tau_2$ 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 $\min_{(i,j)\in S_1} |\sigma_{ij}^0|/\lambda_{n1} \to \infty$ 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 [2]. 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 the first and second derivatives, $a_{n1}$ 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 means that $a_{n1} = b_{n1} = 0$ exactly for sufficiently large $n$, thus allowing a flexible choice of $\lambda_{n1}$. For the SCAD penalty (1.2), it can be relaxed as $\min_{(i,j)\in S_1} |\sigma_{ij}^0|/\lambda_{n1} > a$.

Singularity of the origin in condition (C) allows for sparse estimates ([9]). 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 covariance matrix estimation

Minimizing (1.1) involves nonconvex minimization, and we need to prove that there exists a local minimizer $\hat{\mathbf{\Sigma}}$ for the minimization problem. We give the rate of convergence under Frobenius norm. The proof is given in section 5.

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

Theorem 1 states explicitly how the non-sparsity size and dimensionality affect the rate of convergence. Since there are $(p_n + s_{n1})$ non-sparse 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 $\lambda_{n1}^2 \succeq (s_{n1}+1)\log p_n/n$ can be relaxed to $\lambda_{n1}^2 \succeq \log p_n/n$. In this case, the local minimizer becomes the global minimizer. The bias of the penalized $L_1$ 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), if* $(p_n + s_{n1}) \log p_n/n = o(1)$ *and* $\lambda_{n1}^2 \succeq (p_n + s_{n1}) \log p_n/n$*, then for any local minimizer of (1.1) satisfying* $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_F^2 = O_P\{(p_n + s_{n1}) \log p_n/n\}$*, with probability tending to 1,* $\hat{\sigma}_{ij} = 0$ *for all* $(i, j) \in S_1^c$*.*

The proof of the theorem is relegated to section 5. According to Theorem 2, the sparsistency requires a lower bound on the rate of the regularization parameter $\lambda_{n1}$. On the other hand, Condition (B) imposes an upper bound on $\lambda_{n1}$ in order to control the biases. For penalized $L_1$-likelihood, these two conditions are compatible only when $s_{n1} = O(p_n^{1/2})$. When this condition is violated, we can not guarantee simultaneously the rate of convergence specified in Theorem 1 and sparsistency.

On the other hand, if the penalty function used is unbiased, like the SCAD or the hard-thresholding penalties, we do not impose an extra upper bound for $\lambda_{n1}$ 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, $\lambda_{n1}$ is allowed to decay slower to zero than that for the $L_1$-penalty, allowing a larger order for $s_{n1}$ as long as we have $(p_n + s_{n1}) \log p_n/n = o(1)$.

We present the asymptotic normality of the estimators $\hat{\sigma}_{ij}$ with $(i, j) \in S_1$ only, since other elements are equal to 0 according to Theorem 2. Let $\Sigma_{\lambda_{n1}} = \text{diag}(p''_{\lambda_{n1}}(\text{vec}(\mathbf{\Sigma}_0)))$, $\mathbf{b}_1 = p'_{\lambda_{n1}}(|\text{vec}(\mathbf{\Sigma}_0)|)\text{sgn}(\text{vec}(\mathbf{\Sigma}_0))$, where $\text{vec}(A)$ vectorizes a matrix $A$ and $f(\text{vec}(A))$ represents a vector of elements $f(a_{ij})$. They are equal to zero for SCAD, when $n$ is large enough. For a column vector $\mathbf{a}$, a matrix $A$ and an index set $S$, we denote $[\mathbf{a}]_S$ the column vector $\mathbf{a}$ with rows having positions not in the index $S$ removed. Similarly, we denote $[A]_{S\times S}$ the submatrix $A$ with rows and columns having positions not in the index set $S$ removed. Finally, $\otimes$ denotes the Kronecker product operator and define $A^{\otimes 2} = A \otimes A$. We use $K$ to denote the commutation matrix (see e.g. Graybill (2001) for a definition).

**Theorem 3** *(Asymptotic normality) Under conditions in Theorem 1 and* $(p_n + s_{n1})^2/n =$

$o(1)$, *for a unit vector* $\boldsymbol{\alpha}$ *of length* $s_{n1}+p_n$, *we have for* $\hat{\boldsymbol{\Sigma}}$ *in Theorem 1,*

$$n^{1/2}\boldsymbol{\alpha}^T[(I_{p_n^2}+K)\boldsymbol{\Omega}_0^{\otimes 2}]^{-1/2}_{S_1\times S_1} \cdot \{[\Sigma_{\lambda_{n1}}+\boldsymbol{\Omega}_0^{\otimes 2}]_{S_1\times S_1}[vec(\hat{\boldsymbol{\Sigma}})-vec(\boldsymbol{\Sigma}_0)]_{S_1}+[\mathbf{b}_1]_{S_1}\} \xrightarrow{\mathcal{D}} N(0,1).$$

For SCAD or the hard-thresholding penalty, under the last condition in Condition (B), we have for sufficiently large $n$,

$$(\Sigma_{\lambda_{n1}})_{S_1\times S_1}=0 \quad \text{and} \quad [\mathrm{vec}(\boldsymbol{\Sigma}_0)]_{S_1}=0.$$

## 2.2 Properties of sparse correlation matrix estimation

The correlation matrix $\boldsymbol{\Gamma}_0$ retains the same sparse structure of $\boldsymbol{\Sigma}_0$ with known diagonal elements. This special structure allows us to estimate $\boldsymbol{\Gamma}_0$ more accurately. To take the advantage of the known diagonal elements, the sparse correlation matrix $\boldsymbol{\Gamma}_0$ is estimated by minimizing w.r.t. $\boldsymbol{\Gamma}=(\gamma_{ij})$,

$$\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\hat{\boldsymbol{\Gamma}}_{\mathbf{S}})+\log|\boldsymbol{\Gamma}|+\sum_{i\neq j}p_{\nu_{n1}}(|\gamma_{ij}|), \tag{2.1}$$

where $\hat{\boldsymbol{\Gamma}}_{\mathbf{S}}=\hat{\mathbf{W}}^{-1}\mathbf{S}\hat{\mathbf{W}}^{-1}$ 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 $\nu_{n1}$ is a regularization parameter. After obtaining $\hat{\boldsymbol{\Gamma}}$, $\boldsymbol{\Sigma}_0$ can also be estimated by $\tilde{\boldsymbol{\Sigma}}=\hat{\mathbf{W}}\hat{\boldsymbol{\Gamma}}\hat{\mathbf{W}}$.

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

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

where $\boldsymbol{\Gamma}_0=(\gamma^0_{ij})$. We adapt the condition (D) to (D') with $\lambda_{n1}$ there replaced by $\nu_{n1}$, and (B) to (B') as follows:

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

**Theorem 4** *Under regularity conditions (A),(B'),(C) and (D'), if* $p_n/n=o(1)$, $s_{n1}\log p_n/n=o(1)$ *and* $\nu_{n1}^2\succeq(s_{n1}+1)\log p_n/n$, *then there exists a local minimizer* $\hat{\boldsymbol{\Gamma}}$ *for (2.1) such that*

$$\|\hat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}_0\|_F^2=O_P(s_{n1}\log p_n/n).$$

*In addition, for the operator norm, we have*

$$\|\tilde{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}_0\|^2=O_P\{(s_{n1}+1)\log p_n/n\}.$$

The proof of this theorem is similar to that of Theorem 1 and is sketched in section 5. The condition $\nu_{n1}^2 \succeq (s_{n1}+1)\log p_n/n$ can be relaxed to $\nu_{n1}^2 \succeq \log p_n/n$ when the $L_1$-penalty is used. This theorem indeed shows that the correlation matrix can be estimated more accurately, without the errors from estimating the diagonal elements. It spells clearly the contribution due to dimensionality is merely a factor of $\log p_n$. The following theorem gives the condition under which sparsistency holds.

**Theorem 5** *Under the conditions of Theorem 4, the local minimizer $\hat{\mathbf{\Gamma}}$ in Theorem 4 must satisfy $\hat{\gamma}_{ij}=0$ for all $(i,j)\in S_1^c$ with probability tending to one.*

Like Theorem 2, Theorem 5 holds for any local minimizer with the property given in Theorem 4. The proof follows similarly to that of Theorem 2, with an application of Theorem 4 and establishing $\|\hat{\mathbf{W}}^{-1}-\mathbf{W}_0^{-1}\| = O_P(\{\log p_n/n\}^{1/2})$ and $\max_{i,j}|(\hat{\mathbf{\Gamma}}_{\mathbf{S}})_{ij}-\gamma_{ij}^0| = O_P(\{\log p_n/n\}^{1/2})$, where $\mathbf{W}_0 = \mathbf{D}_{\mathbf{W}}^{1/2}$.

Theorems 4 and 5 are applicable to the $L_1$-penalty. In this case, the local minimizer becomes the global one. In order to have the optimal rate of convergence and sparsitency simultaneously, we need the conditions on the biases in Theorem 4 and on the variance in Theorem 5 compatible. It is easy to calculate that the compatibility requires $s_{n1} = O(1)$, finite number of non-sparse correlation. This is too much a restrictive in many applications.

However, like the case in estimating $\hat{\mathbf{\Sigma}}$, if the penalty function is flat at tails such as the SCAD or the hard-thresholding penalties, no upper bound for $\nu_{n1}$ is needed in order to control the bias term and hence no restriction on $s_{n1}$ is imposed, as long as the conditions in Theorem 2 hold. It is clear that SCAD results in better sampling properties than the penalized $L_1$ estimator.

# 3 Estimation of sparse precision matrix

In this section, we analyze the sparse precision matrix estimation using penalized likelihood (1.3). The method is modified to estimate the inverse correlation matrix, which improves the rate of convergence.

### 3.1 Properties of sparse precision matrix estimation

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

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

Technical conditions in section 2 need some revision. In particular, condition (D) now becomes condition (D2) with $\lambda_{n1}$ there replaced by $\lambda_{n2}$. 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
$\min_{(i,j)\in S_2} |\omega_{ij}^0|/\lambda_{n2} \to \infty$ as $n \to \infty$.

Note that the condition $\min_{(i,j)\in S_2} |\omega_{ij}^0|/\lambda_{n2} > a$ suffices for SCAD penalty defined in (1.2).

**Theorem 6** *(Rate of convergence). Under regularity conditions (A), (B2), (C) and (D2), if* $(p_n + s_{n2}) \log p_n/n = o(1)$ *and*

$$\lambda_{n2}^2 \succeq (s_{n2}+1) \log p_n/n,$$

*then there exists a local minimizer* $\hat{\mathbf{\Omega}}$ *such that* $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_F^2 = O_P\{(p_n + s_{n2}) \log p_n/n\}$.

This theorem is applicable to the penalized $L_1$-likelihood, which was studied thoroughly by [19]. The condition for $\lambda_{n2}$ can then be relaxed to $\lambda_{n2}^2 \succeq \log p_n/n$. In this case, the local minimizer becomes the global one. Hence, our result is an extension of the remarkable result of [19], giving an important understanding how the bias $a_{n2}$ and variance are controlled via the choice of $\lambda_{n2}$ and what role the nonsparse size $s_{n2}$ plays. The proof of the theorem is similar to that of Theorem 1 and is omitted.

**Theorem 7** *(Sparsistency). Under the conditions given in Theorem 6, if* $\lambda_{n2}^2 \succeq (p_n + s_{n2}) \log p_n/n$*, then with probability tending to 1, the local minimizer given in Theorem 6 must satisfy* $\hat{\omega}_{ij} = 0$ *for all* $(i,j) \in S_2^c$.

The proof of this theorem is sketched in section 5. Similar to what we have discussed before, the sparsistency property requires a lower bound on $\lambda_{n2}$, while the rate of convergence imposes an upper bound $\lambda_{n2}$ in order to reduce the bias. To achieve simultaneously the optimal rate of convergence and sparsistency, we need these two conditions compatible. This entails $s_{n2} = O(p_n^{1/2})$. This limitation is due to the biases of the penalized $L_1$-estimator. On the other hand, for the penalty function like SCAD or hard-thresholding, the bias term $a_{n2} = 0$ for sufficiently large $n$ and hence there is no upper bound on the choice of $\lambda_{n2}$. As a result, it does not induce an extra upper bound on $s_{n2}$. Of course, the condition $(p_n + s_{n2}) \log p_n/n = o(1)$ is needed for Theorem 6.

We now establish the asymptotic normality for the estimators $\hat{\omega}_{ij}$ with $(i, j) \in S_2$, since the sparsity property holds following Theorem 7. Let $\Omega_{\lambda_{n2}} = \mathrm{diag}(p''_{\lambda_{n2}}(\mathrm{vec}(\mathbf{\Omega}_0)))$, $\mathbf{b}_2 = p'_{\lambda_{n2}}(|\mathrm{vec}(\mathbf{\Omega}_0)|)\mathrm{sgn}(\mathrm{vec}(\mathbf{\Omega}_0))$. The proof is omitted since it is similar to Theorem 3.

**Theorem 8** *(Asymptotic normality) Under conditions in Theorem 6 and $(p_n+s_{n2})^2/n = o(1)$, for a unit vector $\boldsymbol{\alpha}$ of length $s_{n2} + p_n$, we have for $\hat{\mathbf{\Omega}}$ in Theorem 6,*

$$n^{1/2}\boldsymbol{\alpha}^T[(I_{p_n^2} + K)\mathbf{\Sigma}_0^{\otimes 2}]^{-1/2}_{S_2\times S_2} \cdot \{[\Omega_{\lambda_{n2}} + \mathbf{\Sigma}_0^{\otimes 2}]_{S_2\times S_2}[vec(\hat{\mathbf{\Omega}}) - vec(\mathbf{\Omega}_0)]_{S_2} + [\mathbf{b}_2]_{S_2}\} \xrightarrow{\mathcal{D}} N(0, 1).$$

## 3.2 Properties of sparse inverse correlation matrix estimation

In this section, we show that the inverse correlation matrix can be estimated more accurately via a simple modification of (1.3). More specifically, the inverse correlation matrix $\mathbf{\Psi}_0 = \mathbf{W}_0\mathbf{\Omega}_0\mathbf{W}_0$ can be estimated by minimizing w.r.t. $\mathbf{\Psi} = (\psi_{ij})$,

$$\mathrm{tr}(\mathbf{\Psi}\hat{\mathbf{\Gamma}}_{\mathbf{S}}) - \log|\mathbf{\Psi}| + \sum_{i\neq j} p_{\nu_{n2}}(|\psi_{ij}|), \tag{3.1}$$

where $\hat{\mathbf{\Gamma}}_{\mathbf{S}}$ is defined in (2.1), and $\nu_{n2}$ 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

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

where $\mathbf{\Psi}_0 = (\psi^0_{ij})$ and modify condition (D) to (D2') with $\lambda_{n1}$ there replaced by $\nu_{n2}$, and impose

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

**Theorem 9** *Under regularity conditions (A),(B2'),(C) and (D2'), if $(s_{n2}+1)\log p_n/n = o(1)$ and $\nu^2_{n2} \succeq (s_{n2} + 1)\log p_n/n$, then there exists a local minimizer $\hat{\mathbf{\Psi}}$ for (3.1) such that $\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|^2_F = O_P((s_{n2}+1)\log p_n/n)$ and $\|\tilde{\mathbf{\Omega}} - \mathbf{\Omega}_0\|^2 = O_P((s_{n2}+1)\log p_n/n)$ under the operator norm.*

The proof is similar to that of Theorem 4 and is omitted. Yet, this theorem is very different from Theorem 4 and other consistency theorems, in that the condition $p_n/n = o(1)$ is not needed. Indeed $p_n$ can be as large as $o(\exp(n))$ here as long as $(s_{n2}+1)\log p_n/n = o(1)$ is satisfied.

Comparing to the proof of Theorem 4, on top of removing an order of $\{p_n \log p_n/n\}^{1/2}$ by estimating the inverse correlation rather than the inverse covariance, there is no need to estimate the minimum eigenvalue of term like (5.2), which involves the sample covariance matrix $\hat{\mathbf{\Gamma}}_{\mathbf{S}}$ in the integrand (5.3) and behaves badly when $p_n$ is comparable to $n$ or larger than $n$. We give more details on this in the next subsection.

It is somewhat surprising the convergence rate for $\hat{\mathbf{\Psi}}$ under the Frobenius norm is of order $O_p((s_{n2}+1)\log p_n/n)$, since unlike the correlation matrix, the inverse correlation matrix does not have known diagonal elements. Thus, it seems that an order of $(p_n \log p_n/n)^{1/2}$ contributed from estimating the diagonal elements cannot be eliminated. This can be explained and proved as follows. If $s_{n2} \succeq p_n$, the result is obvious. When $s_{n2} = o(p_n)$, most of off-diagonal elements are zero. Indeed, there are at most $O(s_{n2})$ columns of the inverse correlation matrix contain at least one non-sparse elements. The rest columns that have all zero off-diagonal elements must have diagonal entries of 1. These columns represent variables that are actually uncorrelated from the rest. Now, it is easy to see from (3.1), that these diagonal elements, which are one, are all estimated exactly. 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 [19], and the condition for $\nu_{n2}$ can be relaxed to $\nu_{n2}^2 \succeq \log p_n/n$. We offer the sparsistency result as follows.

**Theorem 10** *(Sparsistency) Under the conditions given in Theorem 9, with probability tending to 1, the local minimizer $\hat{\mathbf{\Psi}}$ given in Theorem 9 satisfies $\hat{\psi}_{ij} = 0$ for all $(i,j) \in S_2^c$.*

The proof follows similarly to that of Theorem 7 and is omitted. This theorem says that sparsity structure for $\mathbf{\Omega}_0$ can be estimated more accurately than the covariance matrix $\mathbf{\Sigma}_0$, in the sense that $p_n$ allowed here is much larger than in Theorem 5.

Similarly to what we remarked before, if we use the $L_1$-penalty, for the resulting inverse correlation matrix estimator to have both optimal rate of convergence and sparsistency, we need to impose $s_{n2} = O(1)$. On the other hand, for penalty functions like the SCAD or the hard-thresholding penalties, we do not need an upper bound on $s_{n2}$, as long as $(s_{n2}+1)\log p_n/n = o(1)$ is satisfied.

### 3.3 Remarks

As discussed briefly in the previous subsection, the condition $p_n/n = o(1)$ is not needed for the consistency of the local minimizer for (3.1). In the proof of Theorem 9, a lower bound on the minimum eigenvalue is needed for

$$\int_0^1 \mathbf{\Psi}_v^{-1} \otimes \mathbf{\Psi}_v^{-1}(1-v)dv,$$

with $\mathbf{\Psi}_v = \mathbf{\Psi}_0 + v\Delta_U$ and $\Delta_U = \alpha_n U$, where $\alpha_n^2 = s_{n2}\log p_n/n$, $U$ is a symmetric matrix with $\|U\|_F = C$ for some constant $C$. Even if $p_n > n$, asymptotically $\mathbf{\Psi}_v^{-1} \otimes \mathbf{\Psi}_v^{-1} = \mathbf{\Gamma}_0 \otimes \mathbf{\Gamma}_0$. Thus a quadratic form involving the integral above is positive by condition (A), and the proof is still valid as long as $\alpha_n = o(1)$. Hence $p_n$ can be as large as $o(\exp(n))$.

On the other hand, a similar integral appears in the proof of Theorem 4 which requires a lower bound on the minimum eigenvalue of (compare to (5.2) and (5.3))

$$\int_0^1 h(v, \mathbf{\Gamma}_v)(1-v)dv, \tag{3.2}$$

where

$$\mathbf{\Gamma}_v = \mathbf{\Gamma}_0 + v\Delta_U, \quad \Delta_U = \alpha_n U,$$
$$h(v, \mathbf{\Sigma}_v) = \mathbf{\Gamma}_v^{-1} \otimes \mathbf{\Gamma}_v^{-1}\hat{\mathbf{\Gamma}}_{\mathbf{S}}\mathbf{\Gamma}_v^{-1} + \mathbf{\Gamma}_v^{-1}\hat{\mathbf{\Gamma}}_{\mathbf{S}}\mathbf{\Gamma}_v^{-1} \otimes \mathbf{\Gamma}_v^{-1} - \mathbf{\Gamma}_v^{-1} \otimes \mathbf{\Gamma}_v^{-1}.$$

with $\alpha_n^2 = s_{n1}\log p_n/n$ and $U$ a symmetric matrix with zeros on its main diagonal such that $\|U\|_F = C$ for some constant $C$. If $p_n$ is larger than $n$ for instance, then $\hat{\mathbf{\Gamma}}_{\mathbf{S}}$ is singular. Thus the minimum eigenvalue of the integral above can be negative, since the term $-\mathbf{\Gamma}_v^{-1} \otimes \mathbf{\Gamma}_v^{-1}$ asymptotically equals $-\mathbf{\Gamma}_0^{-1} \otimes \mathbf{\Gamma}_0^{-1}$. Going through the proof, we then cannot guarantee the positivity of a quadratic form involving the integral (3.2). Even $p_n/n$ approaches a constant c with $0 < c < 1$, the positivity needed is still not guaranteed.

Hence, estimating the correlation matrix is more difficult than estimating its inverse when $p_n$ is large comparing with $n$.

## 4 Extension to sparse Cholesky decomposition

[17] 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:

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

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 $i^{\text{th}}$ element $\sigma_i^2 = \text{var}(\epsilon_i)$. The optimization problem is unconstrained (since the $\phi_{ij}$'s are free variables), and the estimate for $\mathbf{\Omega}$ is always positive-definite.

[13] and [14] 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

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

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

In view of the results in sections 2.2 and 3.2, we can also replace the sample covariance in (4.3) by the sample correlation, 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

$$a_{n3} = \max_{(i,j)\in S} p'_{\lambda_{n3}}(|t_{ij}^0|), \quad b_{n3} = \max_{(i,j)\in S} p''_{\lambda_{n3}}(|t_{ij}^0|).$$

For (ML), condition (D) is adapted to (D3) with $\lambda_{n1}$ there replaced by $\lambda_{n3}$. 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 $\min_{(i,j)\in S} |\phi_{ij}^0|/\lambda_{n3} \to \infty$ 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 11** *Under regularity conditions (A),(B3),(C),(D3), if* $(p_n + s_{n3}) \log p_n/n = o(1)$ *and* $\lambda_{n3}^2 \succeq (s_{n3} + 1) \log p_n/n$*, then there exists a local minimizer* $\hat{\mathbf{T}}$ *for (ML) such that* $\|\hat{\mathbf{T}} - \mathbf{T}_0\|_F^2 = O_P(s_{n3} \log p_n/n)$ *and* $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_F^2 = O_P\{(p_n + s_{n3}) \log p_n/n\}$.

The proof is similar to those of Theorem 1 and 4 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 $\hat{\boldsymbol{\Omega}}$. If the $L_1$-penalty is used, condition for $\lambda_{n3}$ can be relaxed to $\lambda_{n3}^2 \succeq \log p_n/n$.

**Theorem 12** *(Sparsistency). Under the conditions in Theorem 11, if* $\lambda_{n3}^2 \succeq (p_n + s_{n3}) \log p_n/n$*, then the sparsistency holds for* $\hat{\mathbf{T}}$.

The proof is similar to that of Theorem 2 and is omitted. Similar to what remarked before, in order to have simultaneous optimal rate of convergence and sparsistency, the condition $s_{n3} = O(p_n^{1/2})$ is needed when $L_1$-penalty is used. On the other hand, such a restriction is not needed for unbiased penalties like SCAD or hard-thresholding.

## 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 $\lambda_{n1}$ there replaced by $\lambda_{n5}$. Condition (B) is now substituted by

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

**Theorem 13** *(Rate of convergence) Under regularity conditions (A),(B5),(C) and (D5), if* $(s_{n3}+1) \log p_n/n = o(1)$ *and* $\lambda_{n3}^2 \succeq (s_{n3}+1) \log p_n/n$*, 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*

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

The proof is similar to that of Theorems 1 and 4 and is omitted. The condition for $\lambda_{n3}$ can be relaxed to $\lambda_{n3}^2 \succeq \log p_n/n$ when the $L_1$-penalty function is used. Similar to Theorem 9, $p_n$ can also be as large as $o(\exp(n))$, 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 14** *(Sparsistency). Under the condition of Theorem 13, the sparsistent property holds for* $\hat{\mathbf{T}}$.

Proof is omitted since it is similar to that of Theorem 2. 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 $s_{n3} = O(1)$. 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,

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

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* $\mathbf{\Sigma}_0$ *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

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

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 [2], 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

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

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. Find a matrix $\mathbf{C}_0$ of the same size as $\Delta_1$ and $\mathbf{D}_0$ of the same size as $\Delta_2$ for each $n$, with the property that $\|\mathbf{C}_0\| = O(1) = \|\mathbf{D}_0\|$, and each element in $\mathbf{C}_0$ or $\mathbf{D}_0$ are larger in magnitude than the corresponding elements in $\Delta_1$ and $\Delta_2$ respectively. With probability going to 1, $\mathbf{C}_0$ and $\mathbf{D}_0$ can be found satisfying the conditions just described for each $n$, since $\|\Delta_1\|, \|\Delta_2\| = o_P(1)$, meaning that, by Lemma 1, each element in $\Delta_1$ and $\Delta_2$ are of order $o_P(1)$.

With these constructions, with probability going to 1,

$$\max_{i,j} |(\mathbf{C}_0(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0)_{ij}| > \max_{i,j} |(\Delta_1(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0)_{ij}|,$$

and the former has order $O_P(\{\log p_n/n\}^{1/2})$ by the previous proof. Similarly, $\max_{i,j} |(\mathbf{A}_0(\mathbf{S} - \mathbf{\Sigma}_0)\Delta_2)_{ij}|$ and $\max_{i,j} |(\Delta_1(\mathbf{S} - \mathbf{\Sigma}_0)\Delta_2)_{ij}|$ are all of order $O_P(\{\log p_n/n\}^{1/2})$. 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 $\mathcal{A} = \{U : \|\mathbf{R}_U\|_F = C_1, \|\mathbf{D}_U\|_F = C_2\}$,

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

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

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

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

where

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

It suffice to show that the difference is positive asymptotically with probability tending to 1. Using Taylor's expansion with the integral remainder (details not shown), we have $I_1 = K_1 + K_2$, where

$$\begin{aligned} K_1 &= -\mathrm{tr}((\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{\Omega}_0\Delta_U\mathbf{\Omega}_0) = -\mathrm{tr}((\mathbf{S}_{\mathbf{\Omega}_0} - \mathbf{\Omega}_0)\Delta_U), \\ K_2 &= \mathrm{vec}(\Delta_U)^T\Big\{\int_0^1 g(v, \mathbf{\Sigma}_v)(1-v)dv\Big\}\mathrm{vec}(\Delta_U), \qquad (5.2) \end{aligned}$$

with the definitions $\mathbf{\Sigma}_v = \mathbf{\Sigma}_0 + v\Delta_U$, and $\mathbf{S}_{\mathbf{\Omega}_0}$ 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,

$$g(v, \mathbf{\Sigma}_v) = \mathbf{\Sigma}_v^{-1} \otimes \mathbf{\Sigma}_v^{-1}\mathbf{S}\mathbf{\Sigma}_v^{-1} + \mathbf{\Sigma}_v^{-1}\mathbf{S}\mathbf{\Sigma}_v^{-1} \otimes \mathbf{\Sigma}_v^{-1} - \mathbf{\Sigma}_v^{-1} \otimes \mathbf{\Sigma}_v^{-1}. \tag{5.3}$$

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 $\|\mathbf{\Sigma}_v^{-1}\| = \tau_1^{-1} + O_P(\alpha_n + \beta_n)$. By the semi-circular law, 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 [1]). 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,

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

where $C_1 = \|\mathbf{R}_U\|_F$, $C_2 = \|\mathbf{D}_U\|_F$.

Next, we deal with $K_1$. It is clear that $|K_1| \le L_1 + L_2$, where

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

Using Lemma 1 and 2, we have

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

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

Now, consider $I_2 - L_2$. Since we assumed $\lambda_{n1}^2 \succeq (s_{n1}+1)\log p_n/n$, by condition (C), when $n$ is sufficiently large, we have $\lambda_{n1} \succeq \alpha_n$ 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

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

With the assumption that $\lambda_{n1}^2 \succeq (s_{n1}+1)\log p_n/n$, 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**. It suffice to show that, for each $(i, j) \in S_1^c$, for $\sigma_{ij}$ in a sufficiently small neighborhood around 0, the derivative $\partial q_1(\mathbf{\Sigma})/\partial \sigma_{ij}$ has the same sign as $\sigma_{ij}$ with probability tending to 1. It is easy to show

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

where $\mathrm{sgn}(a)$ denotes the sign of $a$. 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 $j$.

Now consider $\mathbf{\Sigma}$ in the neighborhood $\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|_F^2 = O_p\{(p_n + s_{n1}) \log p_n/n\}$, the rate specified in the theorem. 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

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

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

Applying the above result for $I_2$, using Lemma 1, we have

$$\begin{aligned} \max_{i,j} |(I_2)_{ij}| &\leq \|\mathbf{\Omega}\| \cdot \|\mathbf{\Sigma} - \mathbf{\Sigma}_0\| \cdot \|\mathbf{\Omega}\| \\ &= O_P(\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|). \end{aligned}$$

Hence we have

$$\begin{aligned} \max_{i,j} |(-\mathbf{\Omega S \Omega} + \mathbf{\Omega})_{ij}| &= O_P(\{\log p_n/n\}^{1/2} + \|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|) \\ &= O_P(\{(p_n + s_{n1}) \log p_n/n\}^{1/2}). \end{aligned}$$

Note that by Conditions (C) and (D), we have

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

for $\sigma_{ij}$ in a small neighborhood of 0 (excluding 0 itself) and some positive constant $C_3$. Hence if $\lambda_{n1}^2 \succeq (p_n + s_{n1}) \log p_n / n$, the term $p'_{\lambda_{n1}}(|\sigma_{ij}|)\text{sgn}(\sigma_{ij})$ dominates over $-(\mathbf{\Omega S \Omega})_{ij} + \omega_{ij}$ with probability tending to 1, making the sign of the derivative $\partial q_1(\mathbf{\Sigma})/\partial \sigma_{ij}$ depends on $\text{sgn}(\sigma_{ij})$ only. Hence, we obtain the result. $\square$.

**Proof of Theorem 3**.

Following the sparsity property in Theorem 2, let $Q_1(\mathbf{\Sigma}) = q_1((\sigma_{ij}\mathbf{1}_{(i,j)\in S_1}))$, and we solve $\partial Q_1(\hat{\mathbf{\Sigma}})/\partial \sigma_{ij} = 0$ for $(i,j) \in S_1$, where $\hat{\mathbf{\Sigma}}$ is a local minimizer in Theorem 1. That is, we solve

$$[\nabla Q_1(\hat{\mathbf{\Sigma}})]_{S_1} = [\partial Q_1(\hat{\mathbf{\Sigma}})/\partial \text{vec}(\mathbf{\Sigma})]_{S_1} = \mathbf{0}.$$

By Taylor's expansion, we have

$$[-\text{vec}(\mathbf{\Omega}_0 \mathbf{S} \mathbf{\Omega}_0) + \text{vec}(\mathbf{\Omega}_0) + \mathbf{b}_1]_{S_1} \\ + [g(c, \mathbf{\Sigma}_c) + \nabla^2 P_{\lambda_{n1}}]_{S_1 \times S_1} [\text{vec}(\hat{\mathbf{\Sigma}}) - \text{vec}(\mathbf{\Sigma}_0)]_{S_1} = \mathbf{0},$$

where $\nabla^2 P_{\lambda_{n1}} = \text{diag}(p''_{\lambda_{n1}}(\text{vec}(\mathbf{\Sigma}_c)))$, with $\mathbf{\Sigma}_c = \mathbf{\Sigma}_0 + c\Delta$ and $\Delta = \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0$ where $c$ is a constant, and $g(c, \mathbf{\Sigma}_c)$ is as defined in the proof of Theorem 1.

Since $\|\mathbf{\Omega}_0 \Delta \mathbf{\Omega}_0\| = O_P(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|) = o_P(1)$, from an argument in the proof of Theorem 1, we have

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

Also by regularity condition (D), we have $\nabla^2 P_{\lambda_{n1}} = \Sigma_{\lambda_{n1}} + o_P(1)$. Therefore,

$$\begin{aligned} & [\mathbf{\Omega}_0^{\otimes 2} + \Sigma_{\lambda_{n1}}]_{S_1 \times S_1} [\text{vec}(\hat{\mathbf{\Sigma}}) - \text{vec}(\mathbf{\Sigma}_0)]_{S_1} (1 + o_P(1)) + [\mathbf{b}_1]_{S_1} \\ = \ & -[\text{vec}(\mathbf{\Omega}_0 \mathbf{S} \mathbf{\Omega}_0) - \text{vec}(\mathbf{\Omega}_0)]_{S_1}. \end{aligned}$$

Hence, for a unit vector $\boldsymbol{\alpha}$ of length $s_{n1} + p_n$,

$$\begin{aligned} & n^{1/2} \boldsymbol{\alpha}^T [(I_{p_n^2} + K)\mathbf{\Omega}_0^{\otimes 2}]_{S_1 \times S_1}^{-1/2} \\ & \quad \cdot \left\{ [\mathbf{\Omega}_0^{\otimes 2} + \Sigma_{\lambda_{n1}}]_{S_1 \times S_1} [\text{vec}(\hat{\mathbf{\Sigma}}) - \text{vec}(\mathbf{\Sigma}_0)]_{S_1} (1 + o_P(1)) + [\mathbf{b}_1]_{S_1} \right\} \\ & = n^{-1/2} \sum_{i=1}^n w_i, \end{aligned}$$

where $w_i = -\boldsymbol{\alpha}^T [(I_{p_n^2} + K)\mathbf{\Omega}_0^{\otimes 2}]_{S_1 \times S_1}^{-1/2} [\text{vec}(\mathbf{\Omega}_0(\mathbf{y}_i \mathbf{y}_i^T - \mathbf{\Sigma}_0)\mathbf{\Omega}_0)]_{S_1}$ defines a sequence of i.i.d. random variables with mean zero and variance 1, since

$$\text{var}([\text{vec}(\mathbf{\Omega}_0 \mathbf{y}_i \mathbf{y}_i^T \mathbf{\Omega}_0)]_{S_1}) = [(I_{p_n^2} + K)\mathbf{\Omega}_0^{\otimes 2}]_{S_1 \times S_1}$$

(see e.g. Theorem 9.2.2 of [11] and p90 of [16]). Therefore, it remains to check the Lindeberg condition.

Firstly, $\sum_{i=1}^n E(n^{-1}w_i^2) = \text{var}(w_1) = 1$. Also, for an $\epsilon > 0$,

$$\begin{aligned}\sum_{i=1}^n E(n^{-1}w_i^2\mathbf{1}_{\{|w_i|\geq \epsilon n^{1/2}\}}) &= E(w_1^2\mathbf{1}_{\{|w_1|\geq \epsilon n^{1/2}\}})\\ &\leq \{E(w_1^4)P(w_1 > \epsilon n^{1/2})\}^{1/2}.\end{aligned}$$

For the latter probability, by the Markov inequality,

$$P(|w_1| > \epsilon n^{1/2}) \leq 1/(\epsilon^2 n) = O(n^{-1}). \quad (5.4)$$

For the former expectation, write

$$\mathbf{a}^T = \boldsymbol{\alpha}^T[(I_{p_n^2} + K)\boldsymbol{\Omega}_0^{\otimes 2}]_{S_1\times S_1}^{-1/2}, \quad \mathbf{x} = [\text{vec}(\boldsymbol{\Omega}_0(\mathbf{y}_i\mathbf{y}_i^T - \boldsymbol{\Sigma}_0)\boldsymbol{\Omega}_0)]_{S_1}.$$

Then, since elements of $\mathbf{x}$ are elements of a Wishart matrix centered to zero, it has a finite fourth moment. Consequently,

$$\begin{aligned}E(w_1^4) &= \text{tr}\{E(\mathbf{x}\mathbf{x}^T \otimes \mathbf{x}\mathbf{x}^T)\cdot \mathbf{a}\mathbf{a}^T \otimes \mathbf{a}\mathbf{a}^T\}\\ &= O((p_n + s_{n1})^2),\end{aligned} \quad (5.5)$$

With (5.4) and (5.5), we have

$$\sum_{i=1}^n E(n^{-1}w_i^2\mathbf{1}_{\{|w_i|\geq \epsilon n^{1/2}\}}) = O((p_n + s_{n1})n^{-1/2}) = o(1),$$

which completes the proof. □

**Proof of Theorem 4**. The proof is nearly identical to that of Theorem 1, except that we now set $\Delta_U = \alpha_n U$. The fact that $(\hat{\boldsymbol{\Gamma}}_{\mathbf{S}})_{ii} = 1 = \gamma_{ii}^0$ has no estimation error eliminates an order $(p_n \log p_n/n)^{1/2}$ that contributes from estimating $\text{tr}((\hat{\boldsymbol{\Gamma}}_{\mathbf{S}} - \boldsymbol{\Gamma}_0)\boldsymbol{\Psi}_0\Delta_U\boldsymbol{\Psi}_0)$ for (2.1). 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 [19]. □

**Proof of Theorem 7**. For $\boldsymbol{\Omega}$ a minimizer of (1.3), the derivative for $q_2(\boldsymbol{\Omega})$ w.r.t. $\omega_{ij}$ for $(i,j) \in S_2^c$ is

$$\frac{\partial q_2(\boldsymbol{\Omega})}{\partial \omega_{ij}} = 2(s_{ij} - \sigma_{ij} + p'_{\lambda_n}(|\omega_{ij}|)\text{sgn}(\omega_{ij})).$$

Arguments similar to those in the proof of Theorem 2 shows that

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\| = \|\mathbf{\Sigma}(\mathbf{\Omega}_0 - \mathbf{\Omega})\mathbf{\Sigma}_0\| = O(\|\mathbf{\Omega} - \mathbf{\Omega}_0\|).$$

By Lemma 2 or Lemma 3 of Bickel and Levina (2006), it follows that $\max_{i,j} |s_{ij} - \sigma_{ij}^0| = O_P(\{\log p_n/n\}^{1/2})$. Combining the last two results yield that

$$\begin{aligned}\max_{i,j} |s_{ij} - \sigma_{ij}| &= O_P(|s_{ij} - \sigma_{ij}^0| + \|\mathbf{\Omega} - \mathbf{\Omega}_0\|) \\ &= O_P(\{\log p_n/n\}^{1/2} + \|\mathbf{\Omega} - \mathbf{\Omega}_0\|) \\ &= O_P(\{(p_n + s_{n2}) \log p_n/n\}^{1/2}).\end{aligned}$$

Using arguments similar to the proof of Theorem 2, for $\omega_{ij}$ in a small neighborhood of 0, we need to have $\lambda_{n2} \succeq \{(p_n + s_{n2}) \log p_n/n\}^{1/2}$ in order to have sign of $\partial q_2(\mathbf{\Omega})/\partial \omega_{ij}$ depends on $\mathrm{sgn}(\omega_{ij})$ only with probability tending to 1. The proof of the theorem is completed. □

# 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. (2006), Regularized Estimation of Large Covariance Matrices, Technical Report, UC-Berkeley.

[3] Bickel, P.J. and Levina, E. (2006), Covariance Regularization by Thresholding, *Manuscript*

[4] d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2007), First-order Methods For Sparse Covariance Selection, *SIAM Journal on Matrix Analysis and its Applications*, to appear.

[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. (2007). Network Exploration via the Adaptive LASSO and SCAD Penalities. *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] Graybill, F.A. (2001), *Matrices with Applications in Statistics (2nd ed.),* Belmont, CA: Duxbury Press.

[12] Huang, J., Horowitz, J.L. and Ma S. (2006), Asymptotic Properties of Bridge Estimators in Sparse High-dimensional Regression Models, Technical Report No. 360, University of Iowa.

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

[14] Levina, E., Rothman, A.J. and Zhu, J. (2007), Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty, *Ann. Applied Statist.*, to appear.

[15] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*, 1436–1462.

[16] Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory,* John Wiley, New York.

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

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

[19] 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.

[20] Wagaman, A.S. and Levina, E. (2007). Discovering sparse covariance structures with the Isomap. *Manuscript.*

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

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

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

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

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

[26] Zou, H. and Li, R. (2007). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models (With Discussion). *The Annals of Statistics*, to appear.