Published in final edited form as:
Ann Stat. 2009 ; 37(6B): 4254–4278. doi:10.1214/09-AOS720.

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

Clifford Lam [Lecturer] and
Department of Statistics, London School of Economics and Political Science, London, WC2A 2AE (C.Lam2@lse.ac.uk)

Jianqing Fan [Professor]
Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (jqfan@princeton.edu)

Abstract

This paper studies the sparsistency and rates of convergence for estimating sparse covariance and precision matrices based on penalized likelihood with nonconvex 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 nonzero 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 conditions on the rate with which the tuning parameter λ_n goes to 0 have been made explicit and compared under different penalties. As a result, for the L_1-penalty, to guarantee the sparsistency and optimal rate of convergence, the number of nonzero elements should be small: $s_n'=O(p_n)$ at most, among $O\left(p_n^2\right)$ 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 nonzero elements on the off-diagonal entries. On the other hand, using the SCAD or hard-thresholding penalty functions, there is no such a restriction.

Keywords

Covariance matrix; high dimensionality; consistency; nonconcave penalized likelihood; sparsistency; asymptotic normality

1 Introduction

Covariance matrix estimation is a common statistical problem 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

[*]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.

matrix or precision matrix is 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 find a banded structure for $\mathbf{\Omega}$ nonparametrically for longitudinal data. Bickel and Levina (2008b) developed consistency theories on banding methods for longitudinal data, for both $\mathbf{\Sigma}$ and $\mathbf{\Omega}$.

Various authors have used penalized likelihood methods 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 the 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 number of nonzero elements s_n in the true covariance matrix $\mathbf{\Sigma}_0$, or the precision matrix $\mathbf{\Omega}_0$. A direct application of their results to our setting can only handle a relatively small covariance matrix of size $p_n = o(n^{1/10})$.

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) used different LASSO algorithms to select zero elements in the precision matrix. A novel penalty called the nested Lasso was constructed in Levina, Rothman and Zhu (2008) to penalize off-diagonal elements. Thresholding the sample covariance matrix in high-dimensional setting was thoroughly studied by El Karoui (2008) and Bickel and Levina (2008a) and Cai, Zhang and Zhou (2009) 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 (2008) 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 (2008) 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 nonzero off-diagonal entries. However, such a rate of convergence neither addresses explicitly the issues of sparsistency such as those in Fan and Li (2001) and Zhao and Yu (2006), nor the bias issues due to the L_1-penalty and the sampling distribution of the estimated nonzero 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 weaker requirement than that of Ravikumar, Lafferty, Liu and Wasserman (2008).

In this paper, we investigate the aforementioned problems using the penalized pseudo-likelihood method. Assume a random sample $\{\mathbf{y}_i\}_{1\le i\le n}$ with mean zero and covariance matrix $\mathbf{\Sigma}_0$, satisfying some sub-Gaussian tails conditions as specified in Lemma 2 (see

Section 5). The sparsity of the true precision matrix $\mathbf{\Omega}_0$ can be explored by maximizing the Gaussian quasi-likelihood or equivalently minimizing

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

which is the penalized negative log-likelihood if the data is Gaussian. The matrix $S = n^{-1}\sum_{i=1}^{n} y_i y_i^T$ is the sample covariance matrix, $\mathbf{\Omega} = (\omega_{ij})$, 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

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

are folded-concave. 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. The computation can be done via the local linear approximation (Zhou and Li, 2008, Fan *et al.* 2009); see Section 2.1 for additional details.

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

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

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.

In studying a 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 they 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 estimating 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, at $(s_{n1} \log p_n/n)^{1/2}$ instead of $((p_n + s_{n1}) \log p_n/n)^{1/2}$, where s_{n1} is the number of nonzero correlation coefficients. We can take similar advantages over the estimation of the true inverse correlation matrix, denoted by $\mathbf{\Psi}_0$. See Section 2.5. This is an extension of the work of Rothman *et al.* (2008) using the L_1-penalty. Such an extension is important since the nonconcave penalized likelihood ameliorates the bias problem for 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, due to the bias of LASSO, an upper bounded of λ_{ni} is needed in order to achieve fast rate of convergence. On the other hand, a lower bound is required in order to achieve sparsity of estimated precision or covariance matrices. 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 the L_1-penalized estimator can achieve simultaneously the optimal rate of convergence and sparsistency for estimation of $\mathbf{\Sigma}_0$ or $\mathbf{\Omega}_0$ when the number of nonzero elements in the off-diagonal entries are no larger than $O(p_n)$, but not guaranteed so otherwise. On the other hand, using the nonconvex penalties like the SCAD or hard-thresholding penalty, such an extra restriction is not needed.

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 $\| B \| = \lambda_{\max}^{1/2}\left(B^T B\right)$ and $\|B\|_F = \mathrm{tr}^{1/2}(B^T B)$.

2 Estimation of sparse precision matrix

In this section, we present the analysis of (1.1) for estimating a sparse precision matrix. Before this, let us first present an algorithm for computing the nonconcave maximum (pseudo)-likelihood estimator and then state the conditions needed for our technical results.

2.1 Algorithm based on iterated reweighted L_1-penalty

The computation of the nonconcave maximum likelihood problems can be solved by a sequence of L_1-penalized likelihood problems via local linear approximation (Zou and Li 2008, Fan *et al.* 2009). For example, given the current estimate $\Omega_k = (\omega_{ij,k})$, by the local linear approximation to the penalty function,

$$q_1(\Omega) \approx \mathrm{tr}(S\Omega) - \log|\Omega| + \sum_{i\neq j}\left[p_{\lambda_{n1}}\left(|\omega_{ij,k}|\right) + p'_{\lambda_{n1}}\left(|\omega_{ij,k}|\right)\left(|\omega_{ij}| - |\omega_{ij,k}|\right)\right]. \tag{2.1}$$

Hence, $\mathbf{\Omega}_{k+1}$ should be taken to maximize the right-hand side of (2.1):

$$\Omega_{k+1} = \mathrm{argmax}_{\Omega}\left[\mathrm{tr}(S\Omega) - \log|\Omega| + \sum_{i\neq j} p'_{\lambda_{n1}}\left(|\omega_{ij,k}|\right)|\omega_{ij}|\right], \tag{2.2}$$

after ignoring the two constant terms. Problem (2.2) is the weighted penalized L_1-likelihood. In particular, if we take the most primitive initial value $\mathbf{\Omega}_0 = \mathbf{0}$, then

$$\Omega_1 = \mathrm{argmax}_{\Omega}\left[\mathrm{tr}(S\Omega) - \log|\Omega| + \lambda_{n1}\sum_{i\neq j}|\omega_{ij}|\right],$$

is already a good estimator. Iterations of (2.2) reduces the biases of the estimator, as larger estimated coefficients in the previous iterations receive less penalty. In fact, in a different setup, Zou and Li (2008) showed that one iteration of such a procedure is sufficient as long as the initial values are good enough.

Fan *et al.* (2009) has implemented the above algorithm for optimizing (1.1). They have also demonstrated in Section 2.2 in their paper how to utilize the graphical lasso algorithm of Friedman, Hastie and Tibshirani (2008), which is essentially a group coordinate descent procedure, to solve problem (2.2) quickly, even when $p_n > n$. Such a group coordinate decent algorithm was also used by Meier *et al.* (2008) to solve the group LASSO problem.

Thus iteratively, (2.2), and hence (1.1), can be solved quickly with the graphical lasso algorithm. See also Zhang (2007) for a general solution to the folded-concave penalized least-squares problem. The following is a brief summary of the numerical results in Fan *et al.* (2009).

2.2 Some numerical results

We give a brief summary of a breast cancer data analysis with $p_n > n$ considered in Fan *et al.* (2009). For full details, please refer to Section 3.2 of Fan *et al.* (2009). Other simulation results are also in Section 4 in their paper.

Breast cancer data—Normalized gene expression data from 130 patients with stage I-III breast cancers are analyzed, with 33 of them belong to class 1 and 97 belong to class 2. The aim is to assess prediction accuracy in predicting which class a patient will belong to, using a set of pre-selected genes ($p_n = 110$, chosen by t-tests) as gene expression profile data. The data is randomly divided into training ($n = 109$) and testing sets. The mean vector for the genes expression levels is obtained from the training data, as well as the associated inverse covariance matrix estimated using LASSO, adaptive LASSO and SCAD penalties as three different regularization methods. A linear discriminant score is then calculated for each regularization method and applied to the testing set to predict if a patient belongs to class 1 or 2. This is repeated 100 times.

On average, the estimated precision matrix $\widehat{\Omega}$ using LASSO has many more nonzeros than that using SCAD (3923 versus 674). This is not surprising when we look at equation (2.3) in our paper, where the L_1 penalty imposes an upper bound on the tuning parameter λ_{n1} for consistency, which links to reducing the bias in the estimation. This makes the λ_{n1} in practice too small to set many of the elements in $\widehat{\Omega}$ to zero. While we do not know which elements in the true $\boldsymbol{\Omega}$ are zero, the large number of nonzero elements in the L_1 penalized estimator seems spurious, and the resulting gene network is not easy to interpret.

On the other hand, SCAD-penalized estimator has a much smaller number of nonzero elements, since the tuning parameter λ_{n1} is not bounded above under consistency of the resulting estimator. This makes the resulting gene network easier to interpret, with some clusters of genes identified.

Also, classification results on the testing set using the SCAD penalty for precision matrix estimation is better than that using the L_1 penalty, in the sense that the specificity (#True Negative/#class 2) is higher (0.794 to 0.768) while the sensitivity (#True Positive/#class 1) is similar to that using L_1-penalized precision matrix estimator.

2.3 Technical conditions

We now introduce some notations and present regularity conditions for the rate of convergence and sparsistency.

Let $S_1=\left\{(i,j):\omega^0_{ij}\neq 0\right\}$, where $\Omega_0=\left(\omega^0_{ij}\right)$ is the true precision matrix. Denote by $s_{n1} = |S_1| - p_n$, which is the number of nonzero elements in the off-diagonal entries of $\boldsymbol{\Omega}_0$. Define

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

The term a_{n1} is related to the asymptotic bias of the penalized likelihood estimate due to penalization. Note that for L_1-penalty, $a_{n1} = \lambda_{n1}$ 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 are constants τ_1 and τ_2 such that

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

B. $a_{n1} = O(\{(1 + p_n/(s_{n1} + 1)) \log p_n/n\}^{1/2})$, $b_{n1} = o(1)$, and $\min_{(i,j)\in S_1} |\omega^0_{ij}|/\lambda_{n1} \to \infty$ 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 $\theta_1, \theta_2<C\lambda_{n1}, |p''_{\lambda_{n1}}(\theta_1) - p''_{\lambda_{n1}}(\theta_2)| \le D|\theta_1 - \theta_2|$.

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 bias due to penalty can dominate the variance from the likelihood, resulting in poor estimates.

The last requirement in condition (B) states the rate at which the nonzero parameters should 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 penalty functions, 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 $\min_{(i,j)\in S_1} |\omega^0_{ij}|/\lambda_{n1}<a$.

The singularity in condition (C) gives sparsity in the 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.

2.4 Properties of sparse precision matrix estimation

Minimizing (1.1) involves nonconvex minimization, and we need to prove that there exists a local minimizer $\widehat{\Omega}$ for the minimization problem with a certain rate of convergence, which is given under the Frobenius norm. The proof is given in Section 5. It is similar to the one given in Rothman *et al.* (2008), but now the penalty function is nonconvex.

Theorem 1 (Rate of convergence). Under regularity conditions (A)-(D), if $(p_n+s_{n1}) \log p_n/n=O\left(\lambda^2_{n1}\right)$ *and* $(p_n + s_{n1})(\log p_n)^k/n = O(1)$ *for some* $k > 1$, *then there exists a local minimizer* $\widehat{\Omega}$ *such that* $\| \widehat{\Omega} - \Omega_0 \|_F^2 = O_P\{(p_n+s_{n1}) \log p_n/n\}$. *For the* L_1*-penalty, we only need* $\log p_n/n=O\left(\lambda^2_{n1}\right)$.

The proofs of this theorem and others are relegated to Section 5 so that readers can get more quickly what the results are. As in Fan and Li (2001), the asymptotic bias due to the penalty for each nonzero parameter is a_{n1}. Since we penalized only on the off-diagonal elements, the total bias induced by the penalty is asymptotically of order $s_{n1}a_{n1}$. The square of this total bias over all nonzero elements is of order $O_P\{(p_n + s_{n1}) \log p_n/n\}$ under condition (B).

Theorem 1 states explicitly how the number of nonzero elements and dimensionality affect the rate of convergence. Since there are $(p_n + s_{n1})$ nonzero elements and each of them can be estimated at best with rate $n^{-1/2}$, 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 $\log p_n$. The results holds as long as $(p_n+s_{n1})/n$ is at a rate $O((\log p_n)^{-k})$ with some $k > 1$, which decays to zero slowly. This means that in practice p_n can be comparable to n without violating the results. The condition here is not minimum possible; we expect it holds for $p \gg n$. Here, we refer the local minimizer as an interior point within a given close set such that it minimizes the target function. Following a similar argument to Huang *et al.* (2008), the local minimizer in Theorem 1 can be taken as the global minimizer with additional conditions on the tail of the penalty function.

Theorem 1 is also applicable to the L_1-penalty function, where the local minimizer becomes the global minimizer. The asymptotic bias of the L_1-penalized estimate is given in the term $s_{n1}a_{n1} = s_{n1}\lambda_{n1}$ as shown in the technical proof. In order to control the bias, we impose condition (B), which entails an upper bound on $\lambda_{n1} = O(\{(1+p_n/(s_{n1}+1)) \log p_n/n\}^{1/2})$. The bias problem due to the L_1-penalty for finite parameter has already been unveiled by Fan and Li (2001) and Zou (2006).

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

Theorem 2 (Sparsistency). Under the conditions given in Theorem 1, for any local minimizer of (1.1) satisfying $\| \widehat{\Omega} - \Omega_0 \|_F^2 = O_P\{(p_n+s_{n1}) \log p_n/n\}$ *and* $\| \widehat{\Omega} - \Omega_0 \|^2 = O_P(\eta_n)$ *for a sequence of* $\eta_n \to 0$, *if* $\log p_n/n+\eta_n=O\left(\lambda_{n1}^2\right)$, *then with probability tending to 1,* $\widehat{\omega}_{ij}=0$ *for all* $(i, j) \in S_1^c$.

First, 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 zero 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,

$$\log p_n/n+\eta_n=O\left(\lambda_{n1}^2\right) = (1+p_n/ (s_{n1}+1)) \log p_n/n \tag{2.3}$$

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 $\| \widehat{\Omega} - \Omega_0 \| \le \| \widehat{\Omega} - \Omega_0 \|_F$. In the worst case scenario where they have the same order, $\| \widehat{\Omega} - \Omega_0 \|^2 = O_P((p_n+s_{n1}) \log p_n/n)$, so that $\eta_n = (p_n+s_{n1}) \log p_n/n$. It is then easy to see from (2.3) that the two bounds are compatible only when $s_{n1} = O(1)$.
2. We also have $\| \widehat{\Omega} - \Omega_0 \|_F^2/p_n \le \| \widehat{\Omega} - \Omega_0 \|^2$. In the optimistic scenario where they have the same order,

$$\| \widehat{\Omega} - \Omega_0 \|^2 = O_P\left((1+s_{n1}/p_n)\log p_n/n\right).$$

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 penalty, 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 penalty functions, when n is sufficiently large; see condition (B) and the explanation thereof). Thus, λ_{n1} is allowed to decay to zero slowly, allowing even the largest order $s_{n1} = O\left(p_n^2\right)$.

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.

2.5 Properties of sparse inverse correlation matrix estimation

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

$$\mathrm{tr}\left(\Psi\widehat{\Gamma}_S\right) - \log|\Psi| + \sum_{i\neq j} p_{\nu_{n1}}\left(|\psi_{ij}|\right), \tag{2.4}$$

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

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

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

where $\Psi_0 = \left(\psi^0_{ij}\right)$ and modify condition (D) to (D') with λ_{n1} there replaced by υ_{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 $(s_{n1}+1)(\log p_n)^k/n = O(1)$ *for some* $k > 1$ *and* $(s_{n1}+1)\log p_n/n = o\left(\nu_{n1}^2\right)$*, then there exists a local minimizer* $\widehat{\Psi}$ *for* (2.4) *such that* $\| \widehat{\Psi} - \Psi_0 \|_F^2 = O_P\left(s_{n1}\log p_n/n\right)$ *and* $\|\tilde{\Omega} - \Omega_0\|^2 = O_P\left((s_{n1}+1)\log p_n/n\right)$ *under the operator norm. For the* L_1*-penalty, we only need* $\log p_n/n = O\left(\nu_{n1}^2\right)$

Note that we can allow $p_n \gg n$ without violating the result as long as $\log p_n/n = o(1)$. Note also 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 the inverse correlation matrix, unlike the 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 the off-diagonal elements are zero. Indeed, there are at most $O(s_{n1})$ columns of the inverse correlation matrix which contain at least one nonzero element. 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 (2.4) 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.* (2008). We offer the sparsistency result as follows.

Theorem 4 (Sparsistency) Under the conditions given in Theorem 3, for any local minimizer of (2.4) satisfying $\|\widehat{\Psi}-\Psi_0\|_F^2=O_P(s_{n1}\log p_n/n)$ *and* $\|\widehat{\Psi}-\Psi_0\|^2=O_P(n_n)$ *for some* $\eta_n \to 0$, *if* $\log p_n/n+n_n=O(\nu_{n1}^2)$, *then with probability tending to 1,* $\widehat{\psi}_{ij}=0$ *for all* $(i,j) \in S_1^c$.

The proof follows exactly the same as that for Theorem 2 in Section 2.4, and is thus omitted.

For the L_1-penalty, control of bias and sparsistency require υ_{n1} to satisfy bounds like (2.3):

$$\log p_n/n+\eta_n=O\left(\nu_{n1}^2\right)=\log p_n/n. \tag{2.5}$$

This leads to two scenarios:

1. The worst case scenario has

$$\|\widehat{\Psi}-\Psi_0\|^2=\|\widehat{\Psi}-\Psi_0\|_F^2=O_P\left(s_{n1}\log p_n/n\right),$$

meaning $\eta_n = s_{n1} \log p_n/n$. Then compatibility of the bounds in (2.5) requires $s_{n1} = O(1)$.

2. The optimistic scenario has

$$\|\widehat{\Psi}-\Psi_0\|^2=\|\widehat{\Psi}-\Psi_0\|_F^2/p_n=O_P\left(s_{n1}/p_n\cdot\log p_n/n\right),$$

meaning $\eta_n = s_{n1}/p_n \cdot \log p_n/n$. Then compatibility of the bounds in (2.5) 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, we only need $(s_n + 1)(\log p_n)^k/n = O(1)$ as $n \to \infty$ for some $k > 1$. 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 matrix estimation using the penalized likelihood (1.3). Then it is modified to estimating the correlation matrix, which improves the rate of convergence. We assume that the $\mathbf{y}_i$'s are i.i.d. $N(\mathbf{0}, \mathbf{\Sigma}_0)$ throughout this section.

3.1 Properties of sparse covariance matrix estimation

Let $S_2=\left\{(i,j):\sigma^0_{ij}\neq 0\right\}$, where $\Sigma_0=\left(\sigma^0_{ij}\right)$. Denote $s_{n2} = |S_2| - p_n$, so that s_{n2} is the number of nonzero elements in $\mathbf{\Sigma}_0$ on the off-diagonal entries. Put

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

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) $a_{n2} = \mathrm{O}(\{(1 + p_n/(s_n2 + 1)) \log p_n/n\}^{1/2})$, $b_{n2} = o(1)$, and $\min_{(i,j)\in S_2}|\sigma^0_{ij}|/\lambda_{n2}\to\infty$ as $n \to \infty$.

Theorem 5 (Rate of convergence). Under regularity conditions (A), (B2), (C) and (D2), if $(p_n + s_{n2})(\log p_n)^k/n = O(1)$ *for some* $k > 1$ *and* $(p_n + s_{n2}) \log p_n/n=O\left(\lambda^2_{n2}\right)$, *then there exists a local minimizer* $\widehat{\Sigma}$ *such that* $\|\widehat{\Sigma}-\Sigma_0\|_F^2=O_P\left((p_n+s_{n2})\log p_n/n\right)$. *For the* L_1*-penalty, we only need* $\log p_n/n=O\left(\lambda^2_{n2}\right)$.

Like the case for precision matrix estimation, the asymptotic bias due to the L_1-penalty is of order $s_{n2}a_{n2} = s_{n2}\lambda_{n2}$. To control this term, for the L_1-penalty, we require $\lambda_{n2} = O(\{(1 + p_n/(s_{n2} + 1)) \log p_n/n\}^{1/2})$.

Theorem 6 (Sparsistency). Under the conditions given in Theorem 5, for any local minimizer $\widehat{\Sigma}$ *of* (1.3) *satisfying* $\|\widehat{\Sigma}-\Sigma_0\|_F^2=O_P\left((p_n+s_{n2})\log p_n/n\right)$ *and* $\|\widehat{\Sigma}-\Sigma_0\|^2=O_P\left(n_n\right)$ *for some* $\eta_n \to 0$, *if* $\log p_n/n+n_n=O\left(\lambda^2_{n2}\right)$, *then with probability tending to 1,* $\widehat{\sigma}_{ij}=0$ *for all* $(i, j) \in S_2^c$.

For the L_1-penalized likelihood, controlling of bias for consistency together with sparsistency requires

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

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

$$\|\widehat{\Sigma}-\Sigma_0\|^2=\|\widehat{\Sigma}-\Sigma_0\|_F^2=O_P\left((p_n+s_{n2})\log p_n/n\right),$$

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

$$\|\widehat{\Sigma}-\Sigma_0\|^2=\|\widehat{\Sigma}-\Sigma_0\|_F^2/p_n,$$

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 penalty are used, we do not need an upper bound on λ_{n2} since the bias $a_{n2} = 0$ for sufficiently large n. This gives 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 sparsity structure of $\mathbf{\Sigma}_0$ with known diagonal elements. This special structure allows us to estimate $\mathbf{\Gamma}_0$ more accurately. To take advantage of the known diagonal elements, the sparse correlation matrix $\mathbf{\Gamma}_0$ is estimated by minimizing w.r.t. $\mathbf{\Gamma} = (\gamma_{ij})$,

$$\mathrm{tr}\left(\Gamma^{-1}\widehat{\Gamma}_S\right) + \log|\Gamma| + \sum_{i \neq j} p_{\nu_{n2}}\left(|\gamma_{ij}|\right), \tag{3.2}$$

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

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

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

where $\Gamma_0 = \left(\gamma^0_{ij}\right)$. We modify condition (D) to (D2′) with λ_{n2} there replaced by υ_{n2}, and (B) to (B2′) as follows: (B2′) $c_{n2} = O(\{\log p_n/n\}^{1/2})$, $d_{n2} = o(1)$, and $\min_{(i,j)\in S_2} |\gamma^0_{ij}|/\nu n_2 \to \infty$ as $n \to \infty$.

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

$$\| \widehat{\Gamma} - \Gamma_0 \|^2_F = O_P\left(s_{n2}\log p_n/n\right).$$

In addition, for the operator norm, we have

$$\|\tilde{\Sigma} - \Sigma_0\|^2 = O_P\left\{(s_{n2}+1)\log p_n/n\right\}.$$

For the L_1*-penalty, we only need* $\log\ p_n/n = O\left(\nu^2_{n2}\right)$.

The proof is sketched in Section 5. 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 the conditions given in Theorem 7, for any local minimizer $\widehat{\Gamma}$ *of* (3.2) *satisfying* $\| \widehat{\Gamma} - \Gamma_0 \|^2_F = O_P\left(s_{n2}\log p_n/n\right)$ *and* $\| \widehat{\Gamma} - \Gamma_0 \|^2 = O_P\left(\eta_n\right)$ *for some* $\eta_n \to 0$, *if* $\log\ p_n/n + \eta_n = O\left(\nu^2_{n2}\right)$*, then with probability tending to 1,* $\widehat{\gamma}_{ij} = 0$ *for all* $(i, j) \in S^c_2$.

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

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

This is the same condition as (2.5), hence in the worst scenario where

$$\| \widehat{\Gamma}-\Gamma_0 \|^2=\| \widehat{\Gamma}-\Gamma_0 \|_F^2=O_P\left(s_{n2}\log p_n/n\right),$$

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

$$\| \widehat{\Gamma}-\Gamma_0 \|^2=\| \widehat{\Gamma}-\Gamma_0 \|_F^2/p_n=O_P\left(s_{n2}/p_n\cdot\log p_n/n\right),$$

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

The use of unbiased penalty functions like the SCAD or the hard-thresholding penalty, similar to results in the 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 $\mathbf{y} = (y_1, \cdots, y_{pn})^T$ using the autoregressive model:

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

where $\mathbf{T}$ is the unique unit lower triangular matrix with ones on its diagonal and $(i, j)^{\text{th}}$ element being $-\phi_{ij}$ for $j < i$, and $\mathbf{D}$ is diagonal with i^{th} element being $\sigma_i^2=\mathrm{var}(\epsilon_i)$. 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 estimating $\mathbf{\Omega}_0$. 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 $\widehat{\mathbf{T}}$. The latter corresponds to the original Cholesky decomposition. The sparse Cholesky factor can be estimated through minimizing

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

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 $\mathbf{T}\mathbf{y}=\varepsilon$, the least square version is to minimize $\mathrm{tr}\left(\varepsilon\varepsilon^T\right)=\mathrm{tr}\left(\mathbf{T}^T\mathbf{T}\mathbf{y}\mathbf{y}^T\right)$ in the matrix notation. Aggregating the n observations and adding penalty functions, the least-square criterion is to minimize

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

In view of the results in Sections 2.5 and 3.2, we can also write the sample covariance matrix in (4.2) as $\mathbf{S}=\widehat{\mathbf{W}}\widehat{\Gamma}_S\widehat{\mathbf{W}}$ and then replace $\mathbf{D}^{-1/2}\mathbf{T}\widehat{\mathbf{W}}$ by $\mathbf{T}$, resulting in the normalized (NL) version as follows:

$$(NL): \quad q_5(\mathbf{T})=\mathrm{tr}\left(\mathbf{T}^T\mathbf{T}\widehat{\Gamma}_S\right)-2\log|\mathbf{T}|+2\sum_{i<j}p_{\lambda_{n5}}\left(|t_{ij}|\right). \tag{4.4}$$

We will also assume the $\mathbf{y}_i$'s are i.i.d. $N(\mathbf{0}, \mathbf{\Sigma}_0)$ as in the last section.

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 nonzero set and number of nonzeros associated with each $\mathbf{T}$ above. Define

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

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

Theorem 9 *Under regularity conditions (A),(B3),(C),(D3), if* $(p_n + s_{n3})(\log p_n)^k/n = O(1)$ *for some* $k > 1$ *and* $(p_n+s_{n3})\log p_n/n=O\left(\lambda^2_{n3}\right)$, *then there exists a local minimizer* $\widehat{\mathbf{T}}$ *and* $\widehat{\mathbf{D}}$ *for (ML) such that* $\|\widehat{\mathbf{T}}-\mathbf{T}_0\|^2_F=O_P(s_{n3}\log p_n/n)$, $\|\widehat{\mathbf{D}}-\mathbf{D}_0\|^2_F=O_P(p_n\log p_n/n)$ *and* $\|\widehat{\Omega}-\Omega_0\|^2_F=O_P\{(p_n+s_{n3})\log p_n/n\}$. *For the* L_1*-penalty, we only need* $\log p_n/n=O\left(\lambda^2_{n3}\right)$.

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 $\widehat{\Omega}$.

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

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

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 $s_{n3} = O(1)$. 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 the SCAD or hard-thresholding penalty, giving 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

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

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 $s_{n3}(\log p_n)^k/n = O(1)$ *for some* $k > 1$ *and* $(s_{n3}+1)\log p_n/n=o\left(\lambda_{n5}^2\right)$, *then there exists a local minimizer* $\widehat{\mathbf{T}}$ *for (NL) such that* $\|\widehat{\mathbf{T}}-\mathbf{T}_0\|_F^2=O_P(s_{n3}\log p_n/n)$ *and rate of convergence in the Frobenius norm*

$$\|\widehat{\Omega}-\Omega_0\|_F^2=O_P\{(p_n+s_{n3})\log p_n/n\},$$

and in the operator norm, it is improved to

$$\|\widehat{\Omega}-\Omega_0\|^2=O_P\{(s_{n3}+1)\log p_n/n)\}.$$

For the L_1*-penalty, we only need* $\log p_n/n=O\left(\lambda_{n5}^2\right)$.

The proof is similar to that of Theorems 5 and 7 and is omitted. In this theorem, like Lemma 3, we can have p_n so that p_n/n goes to a constant less than 1. It is evident that normalizing with $\widehat{\mathbf{W}}$ results in an improvement in the rate of convergence in operator norm.

Theorem 12 (Sparsistency). Under the conditions given in Theorem 11, for any local minimizer $\widehat{\mathbf{T}}$ *of* (4.4) *satisfying* $\|\widehat{\mathbf{T}}-\mathbf{T}_0\|_F^2=O_P(s_{n3}\log p_n/n)$ *if* $\log p_n/n+\eta_n=O\left(\lambda_{n5}^2\right)$, *then sparsistency holds for* $\widehat{\mathbf{T}}$, *provided that* $\|\widehat{\mathbf{T}}-\mathbf{T}_0\|^2=O(\eta_n)$ *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-penalized estimator. For simultaneous persistency and optimal rate of convergence using the L_1-penalty, the biases inherent in it 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 three lemmas. The first one concerns with inequalities involving the operator and the Frobenius norms. The other two concern with order estimation for elements in a matrix of the form $\mathbf{A}(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}$, which are 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* $\| \mathbf{A} \|_{\min}^2 = \lambda_{\min}\left(\mathbf{A}^T\mathbf{A}\right)$, *we have*

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

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

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

Similarly,

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

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

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

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* $E(\mathbf{y}_i) = 0$ *and* $var(\mathbf{y}_i) = \mathbf{\Sigma}_0$. *Let* $\mathbf{y}_i = (y_{i1}, \cdots, \mathbf{y}ip_n)$ *with* $y_{ij} \sim F_j$, where $\mathrm{F_j}$ is the c.d.f. of $\mathrm{y_{ij}}$, and let $\mathrm{G_j}$ be the c.d.f. of y_{ij}^2, *with*

$$\max_{1 \leq i \leq p_n} \int_0^\infty \exp(\lambda t)\, dG_j(t) < \infty, \qquad 0 < |\lambda| < \lambda_0 \tag{5.2}$$

for some $\lambda_0 > 0$. *Assume* $\log p_n/n = o(1)$, *and that* $\mathbf{\Sigma}_0$ *has eigenvalues uniformly bounded above as* $n \to \infty$. *Then for constant matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *with* $\|\mathbf{A}\|, \|\mathbf{B}\| = O(1)$, *we have* $\max_{i,j} |(\mathbf{A}(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B})_{ij}| = O_P(\{\log p_n/n\}^{1/2})$.

Remark: The conditions on the y_{ij}'s above are the same as those used in Bickel and Levina (2008b) for relaxing the normality assumption.

Proof of Lemma 2. Let $\mathbf{x}_i = \mathbf{A}\mathbf{y}_i$ and $\mathrm{w}_i = \mathbf{B}^T \mathrm{y}_i$. Define $\mathrm{u}_i = \left(\mathrm{x}_i^T, \mathrm{w}_i^T\right)^T$, with covariance matrix

$$\Sigma_{\mathrm{u}} = \mathrm{var}\,(\mathrm{u}_i) = \begin{pmatrix} \mathrm{A}\Sigma_0\mathrm{A}^T & \mathrm{A}\Sigma_0\mathrm{B} \\ \mathrm{B}^T\Sigma_0\mathrm{A}^T & \mathrm{B}^T\Sigma_0\mathrm{B} \end{pmatrix}.$$

Since $\|(\mathbf{A}^T\,\mathbf{B})^T\| \le (\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2)^{1/2} = O(1)$ and $\|\mathbf{\Sigma}_0\| = O(1)$ uniformly, we have $\|\mathbf{\Sigma}_{\mathbf{u}}\| = O(1)$ uniformly, Then, with $\mathrm{S}_{\mathrm{u}} = n^{-1}\sum_{i=1}^{n} \mathrm{u}_i\mathrm{u}_i^T$, which is the sample covariance matrix for the random sample $\{\mathbf{u}_i\}_{1\le i\le n}$, by Lemma A.3 of Bickel and Levina (2008b) which holds under the assumption for the y_{ij}'s and $\log p_n/n = o(1)$, we have

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

In particular, it means that

$$\max_{i,j} |(\mathrm{A}\,(\mathrm{S} - \Sigma_0)\,\mathrm{B})_{ij}| = \left(n^{-1}\sum_{r=1}^{n} \mathrm{x}_r\mathrm{w}_r^T - \mathrm{A}\Sigma_0\mathrm{B}\right)_{ij} = O_P\left(\{\log p_n/n\}^{1/2}\right),$$

which completes the proof of the lemma. □

Lemma 3 *Let* **S** *be a sample covariance matrix of a random sample* $\mathbf{y}_{i1\le i\le n}$ *with* $\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$. *Assume* $p_n/n \to y \in [0, 1)$, $\mathbf{\Sigma}_0$ *has eigenvalues uniformly bounded as* $n \to \infty$, *and* $\mathbf{A} = \mathbf{A}_0 + \Delta_1$, $\mathbf{B} = \mathbf{B}_0 + \Delta_2$ *are such that the constant matrices* $\|A_0\|, \|B_0\| = O(1)$, *with* $\|\Delta_1\|, \|\Delta_2\| = o_P(1)$. *Then we still have* $\max_{i,j}|(\mathbf{A}(\mathbf{S}-\mathbf{\Sigma}_0)\mathbf{B})_{ij}| = O_P(\{\log p_n/n\}^{1/2})$.

Proof of Lemma 3. Consider

$$\mathrm{A}\,(\mathrm{S} - \Sigma_0)\,\mathrm{B} = K_1 + K_2 + K_3 + K_4, \tag{5.3}$$

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 $\max_{i,j} |(K_1)_{ij}| = O_P(\{\log p_n/n\}^{1/2})$ by Lemma 2. Consider K_2. Suppose the maximum element of the matrix is at the (i, j)-th position. Consider $((\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0)_{ij}$, the (i, j)-th element of $(\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0$. Since each element in $\mathbf{S} - \mathbf{\Sigma}_0$ has a rate $O_P(n^{-1/2})$, the i-th row of $\mathbf{S} - \mathbf{\Sigma}_0$ has a norm of $O_P(\{p_n/n\}^{1/2})$. Also, the j-th column of $\mathbf{B}_0$ has $\|\mathbf{B}_0\mathbf{e}_j\| \le \|\mathbf{B}_0\| = O(1)$. Hence, $((\mathbf{S} - \mathbf{\Sigma}_0)\mathbf{B}_0)_{ij} = O_P(\{p_n/n\}^{1/2})$.

Hence, we can find $c_n = o(\{n/p_n\}^{1/2})$ such that each element in $c_n\mathbf{B}_0^T\,(\mathrm{S} - \Sigma_0)$ has an order larger than that in Δ_1, since $\|\Delta_1\| = o_P(1)$ implies that each element in Δ_1 is also $o_P(1)$ by Lemma 1.

Then suitable choice of c_n leads to

$$\max_{i,j} |(\Delta_1\,(\mathrm{S} - \Sigma_0)\,\mathrm{B}_0)_{ij}| \le c_n \max_k \left|\left(\mathrm{B}_0^T(\mathrm{S} - \Sigma_0)^2\mathrm{B}_0\right)_{kk}\right|. \tag{5.4}$$

At the same time, Theorem 5.10 in Bai and Silverstein (2006) implies that, for $\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$ and $p_n/n \to y \in (0, 1)$, with probability one,

$$\begin{aligned} -2\sqrt{y}-y \le & \liminf_{n\to\infty}\lambda_{\min}\left(\Sigma_0^{-1/2}\mathrm{S}\Sigma_0^{-1/2}-\mathbf{I}\right) \\ \le & \limsup_{n\to\infty}\lambda_{\max}\left(\Sigma_0^{-1/2}\mathrm{S}\Sigma_0^{-1/2}-\mathrm{I}\right) \le 2\sqrt{y}+y. \end{aligned}$$

Hence, if we have $p_n/n = o(1)$, we must have $\| \Sigma_0^{-1/2}\mathrm{S}\Sigma_0^{-1/2}-\mathbf{I} \| = o_P(1)$, or it will contradict the above. It means that $\|\mathbf{S}-\mathbf{\Sigma}_0\| = o_P(1)$ since $\mathbf{\Sigma}_0$ has eigenvalues uniformly bounded. Or, if $p_n/n \to y \in (0, 1)$, then we have $\|\mathbf{S} - \mathbf{\Sigma}_0\| = O_P(1)$ by the above.

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

$$\mathrm{S}-\Sigma_0=\mathrm{Q}\Lambda\mathrm{Q}^T,$$

where Λ is a diagonal matrix with real entries. Then we are free to control c_n again so as to satisfy further that $c_n\|\Lambda\|^2 = o_P(\|\Lambda\|)$, since $\|\Lambda\| = \|\mathbf{S} - \mathbf{\Sigma}_0\| = O_P(1)$ at most. Hence,

$$\begin{aligned} c_n\max_k|\left(\mathrm{B}_0^T(\mathrm{S}-\Sigma_0)^2\mathrm{B}_0\right)_{kk}|= & \max_k|\left(\mathrm{B}_0^T\mathrm{Q}c_n\Lambda^2\mathrm{Q}^T\mathrm{B}_0\right)_{kk}| \\ \le & \max_k|\left(\mathrm{B}_0^T\mathrm{Q}\Lambda\mathrm{Q}^T\mathrm{B}_0\right)_{kk}| \\ = & \max_k|\left(\mathrm{B}_0^T(\mathrm{S}-\Sigma_0)\mathrm{B}_0\right)_{kk}|=O_P\left(\{\log p_n/n\}^{1/2}\right), \end{aligned}$$

where the last line used the previous proof for constant matrix $\mathbf{B}_0$. Hence, combining this with (5.4), we have $\max_{i,j} |(K_2)_{ij}| = O_P(\{\log p_n/n\}^{1/2})$. Similar arguments go for K_3 and K_4. □

Proof of Theorem 1. The main idea of the proof is inspired by Fan and Li (2001) and Rothman *et al.* (2008). 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 $\mathcal{A}=\left\{U:\| \Delta_U \|_F^2=C_1^2\alpha_n^2+C_2^2\beta_n^2\right\}$,

$$P\left(\inf_{U\in\mathcal{A}}q_1(\Omega_0+\Delta_U)>q_1(\Omega_0)\right)\to 1,$$

for sufficiently large constants C_1 and C_2. This implies that there is a local minimizer in $\left\{\Omega_0+\Delta_U:\| \Delta_U \|_F^2 \le C_1^2\alpha_n^2+C_2^2\beta_n^2\right\}$ such that $\| \widehat{\Omega}-\Omega_0 \|_F=O_P(\alpha_n+\beta_n)$ for sufficiently large n, since $\Omega_0 + \Delta_U$ is positive definite. This is shown by noting that

$$\lambda_{\min}(\Omega_0+\Delta_U) \ge \lambda_{\min}(\Omega_0)+\lambda_{\min}(\Delta_U) \ge \lambda_{\min}(\Omega_0)-\| \Delta_U \|_F>0,$$

since Ω_0 has eigenvalues uniformly bounded away from 0 and ∞ by condition (A), and $\|\Delta_U\|_F = O(\alpha_n + \beta_n) = o(1)$.

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

$$q_1(\Omega)-q_1(\Omega_0)=I_1+I_2+I_3,$$

where

$$I_1=\mathrm{tr}\,(\mathrm{S}\Omega)-\log|\Omega|-\left(\mathrm{tr}\,(\mathrm{S}\Omega_0)-\log|\Omega_0|\right),$$

$$I_2=\sum_{(i,j)\in S_1^c}\left(p_{\lambda_{n1}}\left(|\omega_{ij}|\right)-p_{\lambda_{n1}}\left(|\omega_{ij}^0|\right)\right),$$

$$I_3=\sum_{(i,j)\in S_1,i\neq j}\left(p_{\lambda_{n1}}\left(|\omega_{ij}|\right)-p_{\lambda_{n1}}\left(|\omega_{ij}^0|\right)\right).$$

It is sufficient 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

$$\begin{aligned}&K_1=\mathrm{tr}\,((\mathrm{S}-\Sigma_0)\,\Delta_U)\,,\\&K_2=\mathrm{vec}(\Delta_U)^T\left\{\int_0^1 g\,(v,\Omega_v)\,(1-v)\,dv\right\}\mathrm{vec}\,(\Delta_U)\,,\end{aligned}\tag{5.5}$$

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

$$\begin{aligned}K_2\geq\;&\int_0^1(1-v)\min_{0\leq v\leq 1}\lambda_{\min}\left(\Omega_v^{-1}\otimes\Omega_v^{-1}\right)dv\cdot\|\,\mathrm{vec}\,(\Delta_U)\,\|^2\\=\;&\|\,\mathrm{vec}\,(\Delta_U)\,\|^2/2\cdot\min_{0\leq v\leq 1}\lambda_{\max}^{-2}\,(\Omega_v)\\\geq\;&\|\,\mathrm{vec}\,(\Delta_U)\,\|^2/2\cdot(\|\,\Omega_0\,\|+\|\,\Delta_U\,\|)^{-2}\\\geq\;&\left(C_1^2\alpha_n^2+C_2^2\beta_n^2\right)/2\cdot\left(\tau_1^{-1}+o\,(1)\right)-2,\end{aligned}$$

where we used $\|\Delta_U\| \leq C_1\alpha_n + C_2\beta_n = O((\log p_n)^{(1-k)/2}) = o(1)$ by our assumption.

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

$$L_1=\left|\sum_{(i,j)\in S_1}(\mathrm{S}-\Sigma_0)_{ij}(\Delta_U)_{ij}\right|,$$

$$L_2=\left|\sum_{(i,j)\in S_1^c}(\mathrm{S}-\Sigma_0)_{ij}(\Delta_U)_{ij}\right|.$$

Using Lemmas 1 and 2, we have

$$\begin{aligned}L_1\leq\;&(s_{n1}+p_n)^{1/2}\max_{i,j}|(\mathrm{S}-\Sigma_0)_{ij}|\cdot\|\,\Delta_U\,\|_F\\\leq\;&O_P\,(\alpha_n+\beta_n)\cdot\|\,\Delta_U\,\|_F\\=\;&O_P\left(C_1\alpha_n^2+C_2\beta_n^2\right),\end{aligned}$$

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

Now, consider $I_2 - L_2$ for penalties other than L_1. Since $\| \Delta_U \|_F^2 = C_1^2\alpha_n^2 + C_2^2\beta_n^2$ on $\mathcal{A}$, we have that $|\omega_{ij}| = O(C_1\alpha_n + C_2\beta_n) = o(1)$ for all $(i, j) \in S_1^c$. Also, note that the condition on λ_{n1} ensures that, for $(i, j) \in S_1^c$, $|\omega_{ij}| = O(\alpha_n + \beta_n) = o(\lambda_{n1})$. Hence, by condition (C), for all $(i, j) \in S_1^c$, we can find a constant $k_1 > 0$ such that

$$p_{\lambda_{n1}}\left(|\omega_{ij}|\right) \geq \lambda_{n1} k_1 |\omega_{ij}|.$$

This implies that

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

Hence,

$$\begin{aligned} I_2 - L_2 \geq & \sum_{(i,j)\in S_1^c} \left\{\lambda_{n1} k_1 |\omega_{ij}| - |(\mathrm{S} - \Sigma_0)_{ij}| \cdot |\omega_{ij}|\right\} \\ \geq & \sum_{(i,j)\in S_1^c} \left[\lambda_{n1} k_1 - O_P\left(\{\log p_n/n\}^{1/2}\right)\right] \cdot |\omega_{ij}| \\ = & \lambda_{n1} \sum_{(i,j)\in S_1^c} \left[k_1 - O_P\left(\lambda_{n1}^{-1}\{\log p_n/n\}^{1/2}\right)\right] \cdot |\omega_{ij}|. \end{aligned}$$

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

For the L_1-penalty, since we have $\max_{i\neq j} |\mathbf{S} - \mathbf{\Sigma}_0| = O_P((\log p_n/n)^{1/2})$ by Lemma 2, we can find a positive $W = O_P(1)$ such that

$$\max_{i\neq j} |\mathrm{S} - \Sigma_0| = W(\log p_n/n)^{1/2}.$$

Then we can set $\lambda_{n1} = 2W(\log p_n/n)^{1/2}$ or one with order greater than $(\log p_n/n)^{1/2}$, and the above arguments are still valid, so that $I_2 - L_2 > 0$.

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

$$|I_3| \leq \min(C_1, C_2)^{-1} \cdot O(1) \cdot \left(C_1^2\alpha_n^2 + C_2^2\beta_n^2\right) + o(1) \cdot \left(C_1^2\alpha_n^2 + C_2^2\beta_n^2\right),$$

where $o(1)$ and $O(1)$ are the terms independent of C_1 and C_2. By condition (B), we have

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

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

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

where sgn(a) denotes the sign of a. If we can show that the sign of $\partial q_1(\mathbf{\Omega})/\partial \omega_{ij}$ depends on sgn(ω_{ij}) only with probability tending to 1, the optimum will be at 0, so that $\widehat{\omega}_{ij}=0$ for all $(i, j) \in S_2^c$ with probability tending to 1. 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

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

By Lemma 2 or Lemma A.3 of Bickel and Levina (2008b), 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, $|\sigma_{ij} - \sigma_{ij}^0| \le \| \Sigma - \Sigma_0 \|$, which has order

$$\begin{aligned} \| \Sigma - \Sigma_0 \| = & \quad \| \Sigma (\Omega - \Omega_0) \Sigma_0 \| \\ \le & \quad \| \Sigma \| \cdot \| \Omega - \Omega_0 \| \cdot \| \Sigma_0 \| \\ = & \quad O(\| \Omega - \Omega_0 \|), \end{aligned}$$

where we used condition (A) to get $\|\mathbf{\Sigma}_0\| = O(1)$, and using $\eta_n \to 0$ so that $\lambda_{\min}(\Omega - \Omega_0) = o(1)$ for $\| \Omega - \Omega_0 \| = O\left(\eta_n^{1/2}\right)$,

$$\begin{aligned} \| \Sigma \| = \lambda_{\min}^{-1}(\Omega) \le & \quad (\lambda_{\min}(\Omega_0) + \lambda_{\min}(\Omega - \Omega_0))^{-1} \\ = & \quad (O(1) + o(1))^{-1} = O(1). \end{aligned}$$

Hence, $\| \Omega - \Omega_0 \| = O\left(\eta_n^{1/2}\right)$ implies $|I_2| = O\left(\eta_n^{1/2}\right)$.

Combining the last two results yields that

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

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

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

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 $\log \quad p_n/n+n_n=O\left(\lambda_{n1}^2\right)$ in order to have the sign of $\partial q_1(\mathbf{\Omega})/\partial\omega_{ij}$ depends on sgn(ω_{ij}) only with probability tending to 1. The proof of the theorem is completed. □

Proof of Theorem 3. Because of the similarity between equations (2.4) 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.* (2008). □

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 $\mathcal{A}=\left\{U:\| \Delta_U \|_F^2=C_1^2\alpha_n^2+C_2^2\beta_n^2\right\}$. Want to show

$$P\left(\inf_{U\in\mathcal{A}} q_2\left(\Sigma_0+\Delta_U\right)>q_2\left(\Sigma_0\right)\right)\to 1,$$

for sufficiently large constants $C_1$and C_2.

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

$$q_2\left(\Sigma\right)-q_2\left(\Sigma_0\right)=I_1+I_2+I_3,$$

where

$$I_1=\text{tr}\left(S\,\Omega\right)+\log|\Sigma|-\left(\text{tr}\left(S\,\Omega_0\right)+\log|\Sigma_0|\right),$$

$$I_2=\sum_{(i,j)\in S_2^c}\left(p_{\lambda_{n2}}\left(|\sigma_{ij}|\right)-p_{\lambda_{n2}}\left(|\sigma_{ij}^0|\right)\right),$$

$$I_3=\sum_{(i,j)\in S_2,i\neq j}\left(p_{\lambda_{n2}}\left(|\sigma_{ij}|\right)-p_{\lambda_{n2}}\left(|\sigma_{ij}^0|\right)\right),$$

with $I_1 = K_1 + K_2$, where

$$\begin{aligned}K_1&=-\text{tr}\left(\left(\mathrm{S}-\Sigma_0\right)\Omega_0\Delta_U\Omega_0\right)=-\text{tr}\left(\left(\mathrm{S}_{\Omega_0}-\Omega_0\right)\Delta_U\right),\\ K_2&=\text{vec}(\Delta_U)^T\left\{\int_0^1 g\left(v,\Sigma_v\right)(1-v)\,dv,\right\}\text{vec}\left(\Delta_U\right),\end{aligned}\qquad(5.6)$$

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,

$$g\left(v,\Sigma_v\right)=\Sigma_v^{-1}\otimes\Sigma_v^{-1}\mathrm{S}\Sigma_v^{-1}+\Sigma_v^{-1}\mathrm{S}\Sigma_v^{-1}\otimes\Sigma_v^{-1}-\Sigma_v^{-1}\otimes\Sigma_v^{-1}.\qquad(5.7)$$

The treatment of $K_2$is different from that in Theorem 1. By condition (A), and $(p_n + s_{n2})(\log p_n)^k/n = O(1)$ for some $k > 1$, we have

$$\| v\Delta_U \Omega_0 \| \le \| \Delta_U \| \| \Omega_0 \| \le \tau_1^{-1}(C_1\alpha_n + C_2\beta_n) = O\left((\log p_n)^{1-k}\right) = o(1).$$

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

$$\Sigma_v^{-1} = \Omega_0(I + v\Delta_U\Omega_0)^{-1} = \Omega_0(I - v\Delta_U\Omega_0 + o(1)),$$

where the little o (or o_P, O or O_P in any matrix expansions in the remainder of this proof) represents a function of the L_2 norm of the residual matrix in the expansion. That is, $\sum_v^{-1} = \Omega_0 + O_P(\alpha_n + \beta_n)$, and $\| \sum_v^{-1} \| = \tau_1^{-1} + O_P(\alpha_n + \beta_n)$. With $\mathbf{S}_I$ difined as the sample covariance matrix formed from a random sample $\{\mathbf{x}_i\}_{1 \le i \le n}$ having $\mathbf{x}_i \sim N(\mathbf{0}, I)$,

$$\| \mathrm{S} - \Sigma_0 \| = O_P(\| \mathrm{S}_I - I \|) = o_P(1)$$

(see arguments in Lemma 3). These entail

$$\begin{aligned} \mathrm{S}\Sigma_v^{-1} = &\ (\mathrm{S} - \Sigma_0)\Sigma_v^{-1} + \Sigma_0\Sigma_v^{-1} \\ = &\ o_P(1) + I + O_P(\alpha_n + \beta_n) \\ = &\ I + o_P(1). \end{aligned}$$

Combining these results, we have

$$g(v, \Sigma_v) = \Omega_0 \otimes \Omega_0 + O_P(\alpha_n + \beta_n).$$

Consequently,

$$\begin{aligned} K_2 = &\ \mathrm{vec}(\Delta_U)^T \left\{ \int_0^1 \Omega_0 \otimes \Omega_0 (1 + o_P(1))(1 - v)\, dv \right\} \mathrm{vec}(\Delta_U) \\ \ge &\ \lambda_{\min}(\Omega_0 \otimes \Omega_0) \| \mathrm{vec}(\Delta_U) \|^2 / 2 \cdot (1 + o_P(1)) \\ = &\ \tau_1^{-2}\left(C_1^2\alpha_n^2 + C_2^2\beta_n^2\right)/2 \cdot (1 + o_P(1)). \end{aligned}$$

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

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

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

$$-\Omega S\Omega+\Omega=I_1+I_2,$$

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

$$\begin{aligned}\|\Omega\| &= \lambda_{\min}^{-1}(\Sigma) \le (\lambda_{\min}(\Sigma_0)+\lambda_{\min}(\Sigma-\Sigma_0))^{-1}\\ &= \tau_1^{-1}+o(1),\end{aligned}$$

we have

$$\Omega=\Omega_0+(\Omega-\Omega_0)=\Omega_0-\Omega(\Sigma-\Sigma_0)\Omega_0=\Omega_0+\Delta,$$

where $\|\Delta\|\le\|\Omega\|\cdot\|\Sigma-\Sigma_0\|\cdot\|\Omega_0\|=O\left(\eta_n^{1/2}\right)=o(1)$ by Lemma 1, with $\|\mathbf{\Sigma}-\mathbf{\Sigma}_0\|^2 = O(\eta_n)$. Hence, we can apply Lemma 3 and conclude that $\max_{i,j} |(I_1)_{ij}| = O_P(\{\log p_n/n\}^{1/2})$.

For I_2, we have

$$\max_{i,j}|(I_2)_{ij}|\le\|\Omega\|\cdot\|\Sigma-\Sigma_0\|\cdot\|\Omega\|=O(\|\Sigma-\Sigma_0\|)=O\left(\eta_n^{1/2}\right).$$

Hence, we have

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

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 $\left(\widehat{\Gamma}_S\right)_{ii}=1=\gamma_{ii}^0$ has no estimation error eliminates an order $(p_n \log p_n/n)^{1/2}$ that contributes from estimating $\mathrm{tr}\left(\left(\widehat{\Gamma}_S-\Gamma_0\right)\Psi_0\Delta_U\Psi_0\right)$ for (3.2). This is why we can estimate a sparse correlation matrix more accurately.

For the operator norm result, we refer readers to the proof of Theorem 2 of Rothman *et al.* (2008). □

Proof of Theorem 10. For (**T**,**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

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

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

$$I_1=(S-\Sigma_0)T^TD^{-1}\quad I_2=\Sigma_0(T-T_0)^TD^{-1},$$

$$I_3=\Sigma_0\mathrm{T}_0^T\left(\mathrm{D}^{-1}-\mathrm{D}_0^{-1}\right), \quad I_4=\Sigma_0\mathrm{T}_0^T\mathrm{D}_0^{-1}.$$

By the MCD (4.1), $I_4=T_0^{-1}$. Since $i > j$ for $(i, j) \in S_3^c$, we must have $\left(T_0^{-1}\right)_{ji}=0$. 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 $\| \mathrm{D}^{-1}-\mathrm{D}_0^{-1} \| = O\left(\| \mathrm{D}-\mathrm{D}_0 \|\right) = O\left(\zeta_n^{1/2}\right)$. Then we can apply Lemma 3 to show that $\max_{ij} |(I_1)_{ij}| = (\log p_n/n)^{1/2}$.

For I_2, we have $\max_{ij}|(I_2)_{ij}| \le \| \Sigma_0 \| \cdot \| \mathrm{T}-\mathrm{T}_0 \| \cdot \| \mathrm{D}^{-1} \| = O\left(\eta_n^{1/2}\right)$. And finally. $\max_{ij}|(I_3)_{ij}| \le \| \Sigma_0 \| \cdot \| \mathrm{T}_0 \| \cdot \| \mathrm{D}^{-1}-\mathrm{D}_0^{-1} \| = O\left(\zeta_n^{1/2}\right)$.

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

References

[1]. Bai, Z.; Silverstein, JW. Spectral Analysis of Large Dimensional Random Matrices. Science Press; Beijing: 2006.

[2]. Bickel PJ, Levina E. Covariance Regularization by Thresholding. Ann. Statist 2008a;36(6):2577–2604.

[3]. Bickel PJ, Levina E. Regularized Estimation of Large Covariance Matrices. Ann. Statist 2008b;36(1):199–227.

[4]. Cai, T.; Zhang, C-H.; Zhou, H. Optimal rates of convergence for co-variance matrix estimaiton. the Wharton School, University of Pennsylvania; 2009. Technical report

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

[6]. Dempster AP. Covariance Selection. Biometrics 1972;28:157–175.

[7]. Diggle P, Verbyla A. Nonparametric Estimation of Covariance Structure in Longitudinal Data. Biometrics 1998;54(2):401–415. [PubMed: 9629635]

[8]. El Karoui N. Operator Norm Consistent Estimation of a Large Dimensional Sparse Covariance Matrices. Ann. Statist 2008;36(6):2717–2756.

[9]. Fan J, Feng Y, Wu Y. Network Exploration via the Adaptive LASSO and SCAD Penalties. Annals of Applied Statistics 2009;3(2):521–541.

[10]. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. J. Amer. Statist. Assoc 2001;96:1348–1360.

[11]. Fan J, Peng H. Nonconcave Penalized Likelihood With a Diverging Number of Parameters. Ann. Statist 2004;32:928–961.

[12]. Friedman J, Hastie T, Tibshirani R. Sparse Inverse Covariance Estimation with the Graphical LASSO. Biostatistics 2008;9(3):432–441. [PubMed: 18079126]

[13]. Huang J, Horowitz J, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist 2008;36:587–613.

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

[15]. Levina E, Rothman AJ, Zhu J. Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty. Ann. Applied Statist 2008;2(1):245–263.

[16]. Meier L, van de Geer S, Bühlmann P. The group Lasso for logistic regression. Journal of the Royal Statistical Society, B 2008;70:53–71.

[17]. Meinshausen N, Bühlmann P. High dimensional graphs and variable selection with the Lasso. Ann. Statist 2006;34:1436–1462.

[18]. Pourahmadi M. Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation. Biometrika 1999;86:677–690.

[19]. Ravikumar, P.; Lafferty, J.; Liu, H.; Wasserman, L. Advances in Neural Information Processing Systems. MIT Press; 2008. Sparse additive models; p. 20

[20]. Rothman AJ, Bickel PJ, Levina E, Zhu J. Sparse Permutation Invariant Covariance Estimation. Electron. J. Statist 2008;2:494–515.

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

[22]. Wagaman AS, Levina E. Discovering sparse covariance structures with the Isomap. Journal of Computational and Graphical Statistics 2008;18 to appear.

[23]. Wong F, Carter C, Kohn R. Efficient Estimation of Covariance Selection Models. Biometrika 2003;90:809–830.

[24]. Wu WB, Pourahmadi M. Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data. Biometrika 2003;94:1–17.

[25]. Yuan M, Lin Y. Model Selection and Estimation in the Gaussian Graphical Model. Biometrika 2007;90:831–844.

[26]. Zhang, CH. Penalized Linear Unbiased Selection. the statistics dept., Rutgers University; 2007. Technical report 2007-003

[27]. Zhao P, Yu B. On Model Selection Consistency of Lasso. Journal of Machine Learning Research 2006;7:2541–2563.

[28]. Zou H. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc 2006;101:1418–1429.

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