Journal of Machine Learning Research 11 (2010) 2261-2286 Submitted 12/09; Revised 4/10; Published 7/10

High Dimensional Inverse Covariance Matrix Estimation via Linear Programming

Ming Yuan MYUAN@ISYE.GATECH.EDU
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, USA

Editor: John Lafferty

Abstract

This paper considers the problem of estimating a high dimensional inverse covariance matrix that can be well approximated by "sparse" matrices. Taking advantage of the connection between multivariate linear regression and entries of the inverse covariance matrix, we propose an estimating procedure that can effectively exploit such "sparsity". The proposed method can be computed using linear programming and therefore has the potential to be used in very high dimensional problems. Oracle inequalities are established for the estimation error in terms of several operator norms, showing that the method is adaptive to different types of sparsity of the problem.

Keywords: covariance selection, Dantzig selector, Gaussian graphical model, inverse covariance matrix, Lasso, linear programming, oracle inequality, sparsity

1. Introduction

One of the classical problems in multivariate statistics is to estimate the covariance matrix or its inverse. Let $X = (X_1, \ldots, X_p)'$ be a p-dimensional random vector with an unknown covariance matrix Σ_0. The goal is to estimate Σ_0 or its inverse $\Omega_0 := \Sigma_0^{-1}$ based on n independent copies of X, $X^{(1)}, \ldots, X^{(n)}$. The usual sample covariance matrix is most often adopted for this purpose:

$$S = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \bar{X})(X^{(i)} - \bar{X})',$$

where $\bar{X} = \sum X^{(i)}/n$. The behavior of S is well understood and it is known to perform well in the classical setting when the dimensionality p is small (see, e.g., Anderson, 2003; Muirhead, 2005). On the other hand, with the recent advances in science and technology, we are more and more often faced with the problem of high dimensional covariance matrix estimation where the dimensionality p is large when compared with the sample size n. Given the large number of parameters ($p(p+1)/2$) involved, exploiting the sparse nature of the problem becomes critical. In particular, traditional estimates such as S do not take advantage of the possible sparsity and are known to perform poorly under many usual matrix norms when p is large. Motivated by the practical demands and the failure of classical methods, a number of sparse models and approaches have been introduced in recent years to deal with high dimensional covariance matrix estimation. See, for example, Ledoit and Wolf (2004), Levina, Rothman and Zhu (2007), Deng and Yuan (2008), El Karoui (2008), Fan, Fan and Lv (2008), Ravikumar, Raskutti, Wainwright and Yu (2008), Ravikumar, Wainwright, Ruskutti

and Yu (2008), Rocha, Zhao and Yu (2008), Lam and Fan (2009), and Rothman, Levina and Zhu (2009) among others.

Bickel and Levina (2008a) pioneered the theoretical study of high dimensional sparse covariance matrices. They consider the case where the magnitude of the entries of Σ_0 decays at a polynomial rate of their distance from the diagonal; and show that banding the sample covariance matrix or S leads to well-behaved estimates. More recently, Cai, Zhang and Zhou (2010) established minimax convergence rates for estimating this type of covariance matrices. A more general class of covariance matrix model is investigated in Bickel and Levina (2008b) where the rows or columns of Σ_0 is assumed to come from an ℓ_α ball with $0 < \alpha < 1$. They suggest thresholding the entries of S and study its theoretical behavior when p is large. In addition to the aforementioned methods, sparse models have also been proposed for the modified Cholesky factor of the covariance matrix in a series of papers by Pourahmadi and co-authors (Pourahmadi, 1999; Pourahmadi, 2000; Wu and Pourahmadi, 2003; Huang et al., 2006).

In this paper, we focus on another type of sparsity—sparsity in terms of the entries of the inverse covariance matrix. This type of sparsity naturally connects with the problem of covariance selection (Dempster, 1972) and Gaussian graphical models (see, e.g., Whittaker, 1990; Lauritzen, 1996; Edwards, 2000), which makes it particularly appealing in a number of applications. Methods to exploit such sparsity have been proposed recently. Inspired by the nonnegative garrote (Breiman, 1995) and Lasso (Tibshirani, 1996) for the linear regression, Yuan and Lin (2007) propose to impose ℓ_1 type of penalty on the entries of the inverse covariance matrix when maximizing the normal log-likelihood and therefore encourages some of the entries of the estimated inverse covariance matrix to be exact zero. Similar approaches are also taken by Banerjee, El Ghaoui and d'Aspremont (2008). One of the main challenges for this type of methods is computation which has been recently addressed by d'Aspremont, Banerjee and El Ghaoui (2008), Friedman, Hastie and Tibshirani (2008), Rocha, Zhao and Yu (2008), Rothman et al. (2008) and Yuan (2008). Some theoretical properties of this type of methods have also been developed by Yuan and Lin (2007), Ravikumar et al. (2008), Rothman et al. (2008) and Lam and Fan (2009) among others. In particular, the results from Ravikumar et al. (2008) and Rothman et al. (2008) suggest that, although better than the sample covariance matrix, these methods may not perform well when p is larger than the sample size n. It remains unclear to what extent the sparsity of inverse covariance matrix entails well-behaved covariance matrix estimates.

Through the study of a new estimating procedure, we show here that the estimability of a high dimensional inverse covariance matrix is related to how well it can be approximated by a graphical model with a relatively low degree. The revelation that the degree of a graph dictates the difficulty of estimating a high dimensional covariance matrix suggests that the proposed method may be more appropriate to harness sparsity in the inverse covariance matrix than those mentioned earlier in which the ℓ_1 penalty serves as a proxy to control the total number of edges in the graph as opposed to its degree. The proposed method proceeds in two steps. A preliminary estimate is first constructed using a well known relationship between inverse covariance matrix and multivariate linear regression. We show that the preliminary estimate, although often dismissed as an estimate of the inverse covariance matrix, can be easily modified to produce a satisfactory estimate for the inverse covariance matrix. We show that the resulting estimate enjoys very good theoretical properties by establishing oracle inequalities for the estimation error.

The probabilistic bounds we prove suggest that the estimation error of the proposed method adapts to the sparseness of the true inverse covariance matrix. The implications of these oracle in-

equalities are demonstrated on a couple of popular covariance matrix models. When Ω_0 corresponds to a Gaussian graphical model of degree d, we show that the proposed method can achieve convergence rate of the order $O_p[d(n^{-1}\log p)^{-1/2}]$ in terms of several matrix operator norms. We also examine the more general case where the rows or columns of Ω_0 belong to an ℓ_α ball ($0 < \alpha < 1$), the family of positive definite matrices introduced by Bickel and Levina (2008b). We show that the proposed method achieves the convergence rate of $O_p[(n^{-1}\log p)^{(1-\alpha)/2}]$, the same as that obtained by Bickel and Levina (2008b) when assuming that Σ_0 rather than Ω_0 belongs to the same family of matrices. For both examples, we also show that the obtained rates are optimal in a minimax sense when considering estimation error in terms of matrix ℓ_1 or ℓ_∞ norms.

The proposed method shares similar spirits with the neighborhood selection approach proposed by Meinshausen and Bühlmann (2006). However, the two techniques are developed for different purposes. Neighborhood selection aims at identifying the correct graphical model whereas our goal is to estimate the covariance matrix. The distinction is clear when the inverse covariance matrix is only "approximately" sparse and does not have many zero entries. Even when the inverse covariance matrix is indeed sparse, the two tasks of estimation and selection can be different. In particular, our results suggest that good estimation can be achieved under conditions weaker than those often assumed to ensure good selection.

The rest of the paper is organized as follows. In the next section, we describe in details the estimating procedure. Theoretical properties of the method are established in Section 3. All detailed proofs are relegated to Section 6. Numerical experiments are presented in Section 4 to illustrate the merits of the proposed method before concluding with some remarks in Section 5.

2. Methodology

In what follows, we shall write $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)'$. Similarly, denote by $\Sigma_{-i,-j}$ the submatrix of Σ with its ith row and jth column removed. Other notation can also be interpreted in the same fashion. For example, $\Sigma_{i,-j}$ or $\Sigma_{-i,j}$ represents the ith row of Σ with its j entry removed or the jth column with its ith entry removed respectively.

2.1 Regression and Inverse Covariance Matrix

It is well known that if X follows a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, then the conditional distribution of X_i given X_{-i} remains normally distributed (Anderson, 2003), that is,

$$X_i | X_{-i} \sim \mathcal{N}\left(\mu_i + \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}(X_{-i} - \mu_{-i}), \Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}\right).$$

This can be equivalently expressed as the following regression equation:

$$X_i = \alpha_i + X_{-i}'\theta_{(i)} + \varepsilon_i, \tag{1}$$

where $\alpha_i = \mu_i - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\mu_{-i}$ is a scalar, $\theta_{(i)} = \Sigma_{-i,-i}^{-1}\Sigma_{-i,i}$ is a $p-1$ dimensional vector and $\varepsilon_i \sim \mathcal{N}(0, \Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i})$ is independent of X_{-i}. When X follows a more general distribution, similar relationship holds in that $\alpha_i + X_{-i}'\theta_{(i)}$ is the best linear unbiased estimate of X_i given X_{-i} whereas $\mathrm{Var}(\varepsilon_i) = \Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}$.

Now by the inverse formula for block matrices, $\Omega := \Sigma^{-1}$ is given by

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix}^{-1} = \begin{pmatrix} \overbrace{\left(\Sigma_{11} - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\right)^{-1}}^{\Omega_{11}} & -\Omega_{11}\Sigma_{1,-1}\Sigma_{-1,-1}^{-1} \\ -\Sigma_{-1,-1}^{-1}\Sigma_{-1,1}\Omega_{11} & * \end{pmatrix}.$$

More generally, the ith column of Ω can be written as

$$\begin{aligned} \Omega_{ii} &= \left(\Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}\right)^{-1}; \\ \Omega_{-i,i} &= -\left(\Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}\right)^{-1}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i}. \end{aligned}$$

This immediately connects with (1):

$$\begin{aligned} \Omega_{ii} &= (\mathrm{Var}(\varepsilon_i))^{-1}; \\ \Omega_{-i,i} &= -(\mathrm{Var}(\varepsilon_i))^{-1}\theta_{(i)}. \end{aligned}$$

Therefore, an estimate of Ω can potentially be obtained by regressing X_i over X_{-i} for $i = 1, \ldots, p$. Furthermore, the sparsity in the entries of Ω can be translated into sparsity in regression coefficients $\theta_{(i)}$s.

2.2 Initial Estimate

From the aforementioned relationship, a zero entry on the ith column of the inverse covariance matrix implies a zero entry in the regression coefficient $\theta_{(i)}$ and vice versa. This property is exploited by Meinshausen and Bühlmann (2006) to identify the zero pattern of the inverse covariance matrix. Specifically, in the so-called neighborhood selection method, the zero entries of the ith column of Ω_0 are identified by doing variable selection when regressing X_i over X_{-i}. More specifically, they suggest to use Lasso (Tibshirani, 1996) for the purpose of variable selection.

Our goal here, however, is rather different. Instead of identifying which entries of Ω_0 are zero, our focus is on estimating it. The distinction is apparent when Ω_0 is only "approximately" sparse instead of having a lot of zero entries. Even if Ω_0 indeed has lot of zeros, the two tasks can still be quite different in high dimensional problems. For example, in identifying the nonzero entries of Ω_0, it is necessary to assume that all nonzero entries are sufficiently different from zero (see, e.g., Meinshausen and Bühlmann, 2006). Such assumptions may be unrealistic and can be relaxed if the purpose is to estimate the covariance matrix. With such a distinction in mind, the question now is whether or not similar strategies of applying sparse multivariate linear regression to recover the inverse covariance matrix remains useful. The answer is affirmative.

To this end, we consider estimating Ω_0 as follows:

$$\begin{aligned} \tilde{\Omega}_{ii} &= \left(\widehat{\mathrm{Var}}(\varepsilon_i)\right)^{-1}; \\ \tilde{\Omega}_{-i,i} &= -\left(\widehat{\mathrm{Var}}(\varepsilon_i)\right)^{-1}\hat{\theta}_{(i)}, \end{aligned}$$

where $\widehat{\mathrm{Var}}(\varepsilon_i)$ and $\hat{\theta}_{(i)}$ are estimated from regressing X_i over X_{-i}. In particular, we suggest to use the so-called Dantzig selector (Candès and Tao, 2007) for estimating the regression coefficients. We

begin by centering each variable X_i to eliminate the intercept α_i in (1). Denote by $Z_i = X_i - \bar{X}_i$ where $\bar{X}_i$ is the sample average of X_i. The Dantzig selector estimate of $\theta_{(i)}$ is the solution to

$$\min_{\beta \in \mathbb{R}^{p-1}, \beta_0 \in \mathbb{R}} \|\beta\|_{\ell_1} \qquad \text{subject to } \left\| \mathbb{E}_n \left[\left(Z_i - Z'_{-i}\beta \right) Z_{-i} \right] \right\|_{\ell_\infty} \leq \delta,$$

where $\mathbb{E}_n$ represents the sample average, and $\delta > 0$ is a tuning parameter. Recall that $\mathbb{E}_n Z_i Z_j = S_{ij}$. The above problem can also be written in terms of S:

$$\min_{\beta \in \mathbb{R}^{p-1}, \beta_0 \in \mathbb{R}} \|\beta\|_{\ell_1} \qquad \text{subject to } \left\| S_{-i,i} - S_{-i,-i}\beta \right\|_{\ell_\infty} \leq \delta. \tag{2}$$

The minimization of the ℓ_1 norm of the regression coefficient reflects our preference towards sparse models which is particularly important when dealing with high dimensional problems. Once an estimate of $\theta_{(i)}$ is obtained, we can then estimate the variance of ε_i by the mean squared error of the residuals:

$$\widehat{\mathrm{Var}}(\varepsilon_i) = \mathbb{E}_n \left(X_i - X'_{-i}\hat{\theta}_{(i)} \right)^2 = S_{ii} - 2\hat{\theta}'_{(i)} S_{-i,i} + \hat{\theta}'_{(i)} S_{-i,-i} \hat{\theta}_{(i)}.$$

We obtain $\tilde{\Omega}$ by repeating this procedure for $i = 1, \ldots, p$.

We emphasize that for practical purposes, one can also use the Lasso in place of the Dantzig selector for constructing $\tilde{\Omega}$. The choice of Dantzig selector is made for the sake of our further technical developments. In the light of the results of Bickel, Ritov and Tsybakov (2009), similar performance can be expected with either the Lasso or the Dantzig selector although a more rigorous proof when using the Lasso is beyond the scope of the current paper.

2.3 Symmetrization

$\tilde{\Omega}$ is is usually dismissed as an estimate of Ω for it is not even symmetric. In fact, it is not obvious that $\tilde{\Omega}$ is in any sense a reasonable estimate of Ω_0. But a more careful examination suggests otherwise. It reveals that $\tilde{\Omega}$ could be a good estimate in a certain matrix operator norm.

The matrix operator norm is a class of matrix norms induced by vector norms. Let $\|\mathbf{x}\|_{\ell_q}$ be the ℓ_q norm of an p dimensional vector $\mathbf{x} = (x_1, \ldots, x_p)'$, that is,

$$\|\mathbf{x}\|_{\ell_q} = (|x_1|^q + \ldots + |x_p|^q)^{1/q}.$$

Then the matrix ℓ_q norm for an $p \times p$ square matrix $A = (a_{ij})_{1 \leq i,j \leq p}$ is given by

$$\|A\|_{\ell_q} = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_{\ell_q}}{\|\mathbf{x}\|_{\ell_q}}.$$

In the case of $q = 1$ and $q = \infty$, the matrix norm can be given more explicitly as

$$\begin{aligned} \|A\|_{\ell_1} &= \max_{1 \leq j \leq p} \sum_{i=1}^{p} \left| a_{ij} \right|; \\ \|A\|_{\ell_\infty} &= \max_{1 \leq i \leq p} \sum_{j=1}^{p} \left| a_{ij} \right|. \end{aligned}$$

When $q = 2$, the matrix operator norm of A amounts to its leading singular value, and is often referred to as the spectral norm.

A careful study shows that $\tilde{\Omega}$ can be a good estimate of Ω in terms of the matrix ℓ_1 norm in sparse circumstances. It is therefore of interest to consider improved estimates from $\tilde{\Omega}$ that inherits this property. To this end, we propose to adjust $\tilde{\Omega}$ by seeking a symmetric matrix $\hat{\Omega}$ that is the the closest to $\tilde{\Omega}$ in the sense of the matrix ℓ_1 norm, that is, it solves the following problem:

$$\min_{\Omega \text{ is symmetric}} \|\Omega - \tilde{\Omega}\|_{\ell_1}. \tag{3}$$

Recall that

$$\|\Omega - \tilde{\Omega}\|_{\ell_1} = \max_{1 \le j \le p} \sum_{i=1}^{p} \left|\Omega_{ij} - \tilde{\Omega}_{ij}\right|.$$

Problem (3) can therefore be re-formulated as a linear program just like the computation of $\tilde{\Omega}$.

To sum up, our estimate of the inverse covariance matrix is obtained in the following steps:

ALGORITHM FOR COMPUTING $\hat{\Omega}$

Input: Sample covariance matrix – S, tuning parameter – δ.
Output: An estimate of the inverse covariance matrix – $\hat{\Omega}$.

- **Construct $\tilde{\Omega}$**
  for $i = 1$ to p
  - Estimate $\theta_{(i)}$ by $\hat{\theta}_{(i)}$, the solution to

$$\min_{\beta \in \mathbb{R}^{p-1}} \|\beta\|_{\ell_1} \qquad \text{subject to } \left\|S_{-i,i} - S_{-i,-i}\beta\right\|_{\ell_\infty} \le \delta.$$

  - Set

$$\tilde{\Omega}_{ii} = \left(S_{ii} - 2\hat{\theta}'_{(i)}S_{-i,i} + \hat{\theta}'_{(i)}S_{-i,-i}\hat{\theta}_{(i)}\right)^{-1}.$$

  - Set

$$\tilde{\Omega}_{-i,i} = -\tilde{\Omega}_{ii}\hat{\theta}_{(i)}.$$

  end

- **Construct $\hat{\Omega}$**
  - Set $\hat{\Omega}$ as the solution to

$$\min_{\Omega \text{ is symmetric}} \|\Omega - \tilde{\Omega}\|_{\ell_1}.$$

It is worth pointing out that that proposed method depends on the data only through the sample covariance matrix. This fact is of great practical importance since it suggests that a large sample size will not affect the computational complexity in calculating $\hat{\Omega}$ more than the evaluation of S. Furthermore, only linear programs are involved in the computation of $\hat{\Omega}$, which makes the approach appealing when dealing with very high dimensional problems.

3. Theory

In what follows, we shall assume that the components of X are uniformly sub-gaussian, that is, there exist constants $c_0 \geq 0$, and $T > 0$ such that for any $|t| \leq T$

$$\mathbb{E}e^{tX_i^2} \leq c_0, \qquad i = 1, 2, \ldots, p.$$

This condition is clearly satisfied when X follows a multivariate normal distribution. It also holds true when X_is are bounded.

3.1 Oracle Inequality

Our main tool to study the theoretical properties of $\hat{\Omega}$ is an oracle type of inequality regarding the estimation error $\|\hat{\Omega} - \Omega_0\|_{\ell_1}$. To this end, we introduce the following set of "oracle" inverse covariance matrices:

$$O(\nu, \eta, \tau) = \left\{ \Omega \succ 0 : \begin{array}{ll} \nu^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq \nu & \text{(Bounded Eigenvalues)} \\ \|\Sigma_0 \Omega - I\|_{\max} \leq \eta & \text{(“Good” Approximation)} \\ \|\Omega\|_{\ell_1} \leq \tau & \text{(Sparsity)} \end{array} \right\},$$

where $A \succ 0$ indicates that a matrix A is symmetric and positive definite; $\nu > 1$, $\tau > 0$, and $\eta \geq 0$ are parameters; $\lambda_{\min}$ and $\lambda_{\max}$ represent the smallest and largest eigenvalue respectively; and $\|\cdot\|_{\max}$ represents the entry-wise ℓ_∞ norm, that is,

$$\|A\|_{\max} = \max_{1 \leq i,j \leq p} |a_{ij}|.$$

We refer to $O(\nu, \eta, \tau)$ as an "oracle" set because its definition requires the knowledge of the true covariance matrix Σ_0. Every member of $O(\nu, \eta, \tau)$ is symmetric, positive definite with eigenvalues bounded away from 0 and ∞, and belongs to an ℓ_1 ball. Moreover, $O(\nu, \eta, \tau)$ consists of matrices that approximate Ω_0 well. It is worth noting that different from the usual vector case, the choice of metric is critical when evaluating approximating error for matrices. In particular for our purpose, the approximation error is measured by $\|\Sigma_0 \Omega - I\|_{\max}$, which vanishes if and only if $\Omega = \Omega_0$. We are now in position to state our main result.

Theorem 1 *There exist constants C_1, C_2 depending only on ν, τ, $\lambda_{\min}(\Omega_0)$ and $\lambda_{\max}(\Omega_0)$, and C_3 depending only on c_0 such that, for any $A > 0$, with probability at least $1 - p^{-A}$,*

$$\left\|\hat{\Omega} - \Omega_0\right\|_{\ell_1} \leq C_1 \inf_{\Omega \in O(\nu,\eta,\tau)} \left(\left\|\Omega - \Omega_0\right\|_{\ell_1} + \deg(\Omega)\delta \right), \tag{4}$$

provided that

$$\inf_{\Omega \in O(\nu,\eta,\tau)} \left(\left\|\Omega - \Omega_0\right\|_{\ell_1} + \deg(\Omega)\delta \right) \leq C_2, \tag{5}$$

and

$$\delta \geq \nu\eta + C_3 \nu\tau \lambda_{\min}^{-1}(\Omega_0)((A+1)n^{-1}\log p)^{1/2}, \tag{6}$$

where $\deg(\Omega) = \max_i \sum_j \mathbb{I}(\Omega_{ij} \neq 0)$.

We remark that the oracle inequality given in Theorem 1 is of probabilistic nature and non-asymptotic. However, (4) holds with overwhelming probability as we are interested in the case when p is very large. The requirement (5) is in place to ensure that the true inverse covariance matrix is indeed "approximately" sparse. Another note is on the choice of the tuning parameter δ. To ensure an tight upper bound in (4), smaller δs are preferred. On the other hand, Condition (6) specifies how small they can be. For simplicity, we have used the same tuning parameter δ for estimating all $\theta_{(i)}$s. In practice, it may be beneficial to use different δs for different $\theta_{(i)}$s. Following the same argument, it can be shown that the statement of Theorem 1 continue to hold if all tuning parameters used satisfy Condition (6).

Recall that for a symmetric matrix A, $\|A\|_{\ell_\infty} = \|A\|_{\ell_1}$ and

$$\|A\|_{\ell_2} \le (\|A\|_{\ell_1}\|A\|_{\ell_\infty})^{1/2} = \|A\|_{\ell_1}.$$

A direct consequence of Theorem 1 is that the same upper bound holds true under matrix ℓ_∞ and ℓ_2 norms.

Corollary 2 *There exist constants C_1, C_2 depending only on ν, τ, $\lambda_{\min}(\Omega_0)$ and $\lambda_{\max}(\Omega_0)$, and C_3 depending only on c_0 such that, for any $A > 0$, with probability at least $1 - p^{-A}$,*

$$\left\|\hat{\Omega} - \Omega_0\right\|_{\ell_\infty}, \left\|\hat{\Omega} - \Omega_0\right\|_{\ell_2} \le C_1 \inf_{\Omega \in O(\nu,\eta,\tau)} \left(\left\|\Omega - \Omega_0\right\|_{\ell_1} + \deg(\Omega)\delta\right),$$

provided that

$$\inf_{\Omega \in O(\nu,\eta,\tau)} \left(\left\|\Omega - \Omega_0\right\|_{\ell_1} + \deg(\Omega)\delta\right) \le C_2,$$

and

$$\delta \ge \nu\eta + C_3\nu\tau\lambda_{\min}^{-1}(\Omega_0)((A+1)n^{-1}\log p)^{1/2}.$$

The bound on the matrix ℓ_2 has great practical implications when we are interested in estimating the covariance matrix or need to a positive definite estimate of Ω. The proposed estimate $\hat{\Omega}$ is symmetric but not guaranteed to be positive definite. However, Corollary 2 suggests that with overwhelming probability, it is indeed positive definite provided that the upper bound is sufficiently small because

$$\lambda_{\min}(\hat{\Omega}) \ge \lambda_{\min}(\Omega_0) - \left\|\hat{\Omega} - \Omega_0\right\|_{\ell_2}.$$

Moreover, a positive definite estimate of Ω can always be constructed by replacing its negative eigenvalues with δ. Denote the resulting estimate by $\hat{\hat{\Omega}}$. By Corollary 2, it can be shown that

Corollary 3 *There exist constants C_1, C_2 depending only on ν, τ, $\lambda_{\min}(\Omega_0)$ and $\lambda_{\max}(\Omega_0)$, and C_3 depending only on c_0 such that, for any $A > 0$, with probability at least $1 - p^{-A}$,*

$$\left\|\hat{\hat{\Omega}}^{-1} - \Sigma_0\right\|_{\ell_2}, \left\|\hat{\hat{\Omega}} - \Omega_0\right\|_{\ell_2} \le C_1 \inf_{\Omega \in O(\nu,\eta,\tau)} \left(\left\|\Omega - \Omega_0\right\|_{\ell_1} + \deg(\Omega)\delta\right),$$

provided that

$$\inf_{\Omega \in O(\nu,\eta,\tau)} \left(\left\|\Omega - \Omega_0\right\|_{\ell_1} + \deg(\Omega)\delta\right) \le C_2,$$

and

$$\delta \ge \nu\eta + C_3\nu\tau\lambda_{\min}^{-1}(\Omega_0)((A+1)n^{-1}\log p)^{1/2}.$$

When considering a particular class of inverse covariance matrices, we can use the oracle inequalities established here with a proper choice of the oracle set O. Typically in choosing a good oracle set O, we take ν and τ to be of finite magnitude whereas the approximation error η sufficiently small. To further illustrate their practical implications, we now turn to a couple of more concrete examples.

3.2 Sparse Models

We begin with a class of matrix models that are closely connected with graphical models. When X follows a multivariate normal distribution, the sparsity of the entries of the inverse covariance matrix relates to the notion of conditional independence: the (i, j) entry of Ω_0 being zero implies that X_i is independent of X_j conditional on the remaining variables and vice versa. The conditional independence relationships among the coordinates of the Gaussian random vector X can be represented by an undirected graph $G = (V, E)$, often referred to as a Gaussian graphical model, where V contains p vertices corresponding to the p coordinates and the edge between X_i and X_j is present if and only if X_i and X_j are not independent conditional on the others. The complexity of a graphical model is commonly measured by its degree:

$$\deg(G) = \max_{1 \le i \le p} \sum_j e_{ij},$$

where $e_{ij} = 1$ if there is an edge between X_i and X_j and 0 otherwise. Gaussian graphical models are an indispensable statistical tool in studying communication networks and gene pathways among many other subjects. The readers are referred to Whittaker (1990), Lauritzen (1996) and Edwards (2000) for further details.

Motivated by this connection, we consider the following class of inverse covariance matrices:

$$\mathcal{M}_1(\tau_0, \nu_0, d) = \left\{ A \succ 0 : \|A\|_{\ell_1} < \tau_0, \nu_0^{-1} < \lambda_{\min}(A) < \lambda_{\max}(A) < \nu_0, \deg(A) < d \right\},$$

where $\tau_0, \nu_0 > 1$, and $\deg(A) = \max_i \sum_j \mathbb{I}(A_{ij} \neq 0)$. In this case, taking an oracle set such that $\Omega_0 \in O$ yields the following result:

Theorem 4 *Assume that $d(n^{-1}\log p)^{1/2} = o(1)$. Then*

$$\sup_{\Omega_0 \in \mathcal{M}_1(\tau_0, \nu_0, d)} \left\| \hat{\Omega} - \Omega_0 \right\|_{\ell_q} = O_p\left(d\sqrt{\frac{\log p}{n}} \right), \tag{7}$$

provided that $\delta = C(n^{-1}\log p)^{1/2}$ and C is large enough.

Theorem 4 follows immediately from Theorem 1 and Corollary 2 by taking $\eta = 0$, $\tau = \|\Omega_0\|_{\ell_1}$, and $\nu = \max\{\lambda_{\min}^{-1}(\Omega_0), \lambda_{\max}(\Omega_0)\}$, which ensures that $\Omega_0 \in O(\nu, \eta, \tau)$. We note that the rate of convergence given by (7) is also optimal in the minimax sense when considering matrix ℓ_1 norm.

Theorem 5 *Assume that $d(n^{-1}\log p)^{1/2} = o(1)$. Then there exists a constant $C > 0$ depending only on τ_0, and ν_0 such that*

$$\inf_{\bar{\Omega}} \sup_{\Omega_0 \in \mathcal{M}_1(\tau_0, \nu_0, d)} \mathbb{P}\left\{ \left\| \bar{\Omega} - \Omega_0 \right\|_{\ell_1} \ge Cd\sqrt{\frac{\log p}{n}} \right\} > 0,$$

where the infimum is taken over all estimate $\bar{\Omega}$ based on observations $X^{(1)}, \ldots, X^{(n)}$.

Theorem 5 indicates that the estimability of a sparse inverse covariance matrix is dictated by its degree as opposed to the total number of nonzero entries. This observation gives a plausible explanation on why the usual ℓ_1 penalized likelihood estimate (see, e.g., Yuan and Lin, 2007; Banarjee, El Ghaoui and d'Aspremont, 2008) may not be the best to exploit this type of sparsity because the penalty employed by these methods is convex relaxations of the constraint on total number of edges in a graphical model instead of its degree.

It is also of interest to compare our results with those from Meinshausen and Bühlmann (2006). As mentioned before, the goal of the neighborhood selection from Meinshausen and Bühlmann (2006) is to select the correct graphical model whereas our focus here is on estimating the covariance matrix. However, the neighborhood selection method can be followed by the maximum likelihood estimation based on the selected graphical model to yield a covariance matrix estimate. Clearly the success of this method hinges upon the ability of the neighborhood selection to choose a correct graphical model. It turns out that selecting the graphical model can be more difficult than estimating the covariance matrix as reflected by the more restrictive assumptions made in Meinshausen and Bühlmann (2006). In particular, to be able to identify the nonzero entries of the inverse covariance matrix, it is necessary that they are sufficiently large in magnitude whereas such requirement is generally not needed for the purpose of estimation. Moreover, Meinshausen and Bühlmann (2006) only deals with the case when the dimensionality is of a polynomial order of the sample size, that is, $p = O(n^\gamma)$ for some $\gamma > 0$.

3.3 Approximately Sparse Models

In many applications, the inverse covariance matrix is only approximately sparse. A popular way to model this class of covariance matrix is to assume that its rows or columns belong to an ℓ_α ball ($0 < \alpha < 1$):

$$\mathcal{M}_2(\tau_0, \nu_0, \alpha, M) = \left\{ A \succ 0 : \|A^{-1}\|_{\ell_1} < \tau_0, \nu_0^{-1} \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le \nu_0, \sum_{j=1}^{p} \left|A_{ij}\right|^\alpha \le M \right\},$$

where $\tau_0, \nu_0 > 1$ and $0 < \alpha < 1$. $\mathcal{M}_2$ can be viewed as a natural extension of the sparse model $\mathcal{M}_1$. In particular, $\mathcal{M}_1$ can be viewed as the limiting case of $\mathcal{M}_2$ when α approaches 0. By relaxing α, $\mathcal{M}_2$ includes matrices that are less sparse than those included in $\mathcal{M}_1$. The particular class of matrices were first introduced by Bickel and Levina (2008b) who investigate the case when $\Sigma_0 \in \mathcal{M}_2(\tau_0, \nu_0, \alpha, M)$. We note that their setting is different from ours as $\mathcal{M}_2$ is not closed with respect to inversion. An application of Theorem 1 and Corollary 2 yields:

Theorem 6 *Assume that* $M\left(n^{-1}\log p\right)^{\frac{1-\alpha}{2}} = o(1)$. *Then*

$$\sup_{\Omega_0 \in \mathcal{M}_2(\tau_0, \nu_0, \alpha, M)} \|\hat{\Omega} - \Omega_0\|_{\ell_q} = O_p\left(M \left(\frac{\log p}{n} \right)^{\frac{1-\alpha}{2}} \right), \tag{8}$$

provided that $\delta = C(n^{-1}\log p)^{1/2}$ *and* C *is sufficiently large.*

Assuming that $\Sigma_0 \in \mathcal{M}_2$, Bickel and Levina (2008b) study thresholding estimator of the covariance matrix. Their setting is different from ours because $\mathcal{M}_2$ is not closed under inversion. It is

however interesting to note that Bickel and Levina (2008b) show that thresholding the sample covariance matrix S at an appropriate level can achieve the same rate given by right hand side of (8). The coincidence should not come as a surprise despite the difference in problem setting because the size of the parameter space in both problems are the same. Moreover, the following theorem shows that in both settings, the rate is optimal in the minimax sense.

Theorem 7 *Assume that* $M\left(n^{-1}\log p\right)^{\frac{1-\alpha}{2}} = o(1)$. *Then there exists a constant* $C > 0$ *depending only on* τ_0, *and* ν_0 *such that*

$$\inf_{\bar{\Omega}} \sup_{\Omega_0 \in \mathcal{M}_2(\tau_0,\nu_0,\alpha,M)} \mathbb{P}\left\{ \left\|\bar{\Omega} - \Omega_0\right\|_{\ell_1} \geq CM\left(\frac{\log p}{n}\right)^{\frac{1-\alpha}{2}} \right\} > 0, \tag{9}$$

and

$$\inf_{\bar{\Sigma}} \sup_{\Sigma_0 \in \mathcal{M}_2(\tau_0,\nu_0,\alpha,M)} \mathbb{P}\left\{ \left\|\bar{\Sigma} - \Sigma_0\right\|_{\ell_1} \geq CM\left(\frac{\log p}{n}\right)^{\frac{1-\alpha}{2}} \right\} > 0, \tag{10}$$

where the infimum is taken over all estimate, $\bar{\Omega}$ *or* $\bar{\Sigma}$, *based on observations* $X^{(1)},\ldots,X^{(n)}$.

4. Numerical Experiments

To illustrate the merits of the proposed method and compare it with other popular alternatives, we now conduct a set of numerical studies. Specifically, we generated $n = 50$ observations from a multivariate normal distribution with mean 0 and variance covariance matrix given by $\Sigma^0_{ij} = \rho^{|i-j|}$ for some $\rho \neq 0$. Such covariance structure corresponds to an AR(1) model. Its inverse covariance matrix is banded with the magnitude of ρ determining the strength of the dependence among the coordinates. We consider combinations of seven different values of ρ, 0.1, 0.2,..., 0.7 and four values of the dimensionality, $p = 25$, 50, 100 or 200. Two hundred data sets were simulated for each combination. For each simulated data set, we ran the proposed method to construct estimate of the inverse covariance matrix. As suggested by the theoretical developments, we set $\delta = (2n^{-1}\log p)^{-1}$ throughout all simulation studies. For comparison purposes, we included a couple of popular alternative covariance matrix estimates in the study. The first is the ℓ_1 penalized likelihood estimate of Yuan and Lin (2007). As suggested by Yuan and Lin (2007), the BIC criterion was used to choose the tuning parameter among a total of 20 pre-specified values. The second is a variant of the neighborhood selection approach of Meinshausen and Bühlmann (2006). As pointed out earlier, the goal of the neighborhood selection is to identify the underlying graphical model rather than estimating the covariance matrix. We consider here a simple two-step procedure where the maximum likelihood estimate based on the selected graphical model is employed. As advocated by Meinshausen and Bühlmann (2006), the level of significance is set at $\alpha = 0.05$ in identifying the graphical model. Figure 1 summarizes the estimation error measured by the spectral norm, that is, $\|\hat{C} - C\|_{\ell_2}$, for the three methods, averaged over two hundred runs.

A few observations can be made from Figure 1. We first note that the proposed method tends to outperform the other two methods when ρ is small and the advantage becomes more evident as the dimensionality increases. On the other hand, the advantage over the neighborhood selection based method gradually vanishes as ρ increases yet the proposed method remains competitive. A plausible explanation is the distinction between estimation and selection in high dimensional problems as pointed out earlier. The success of the neighbor selection based method hinges upon a good selection

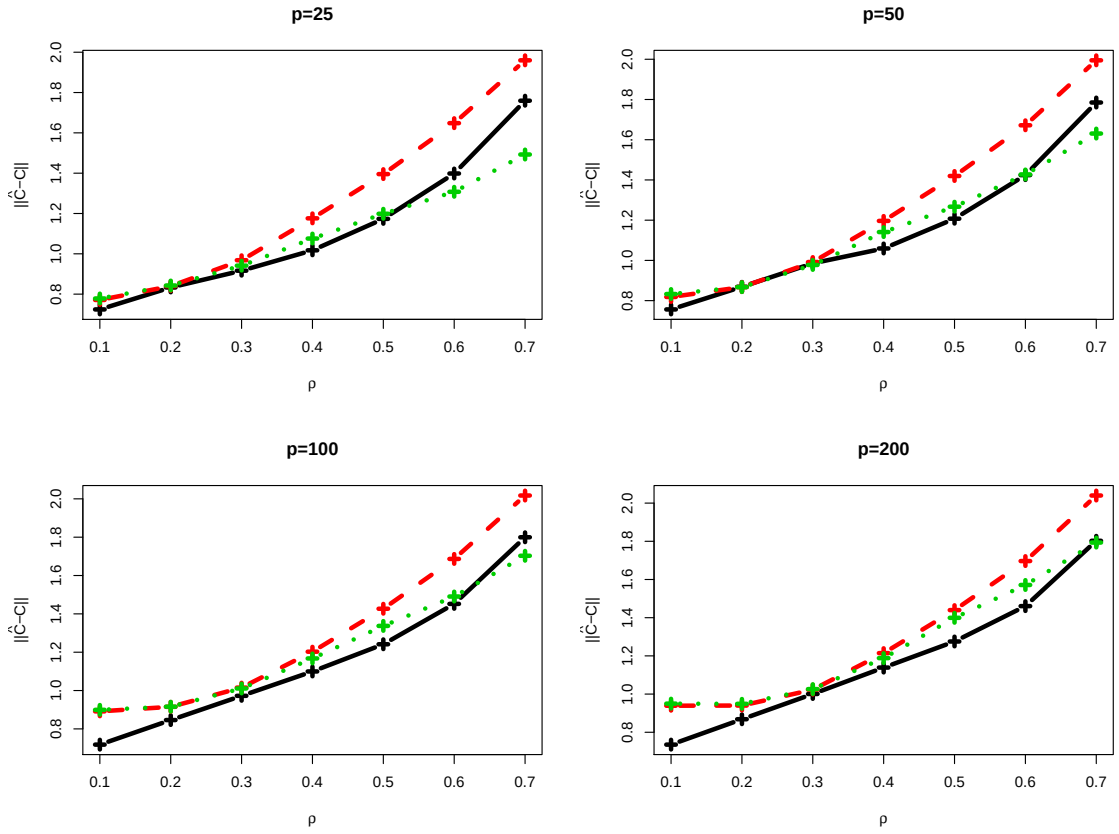

Figure 1: Estimation error of the proposed method (black solid lines), the ℓ_1 penalized likelihood estimate (red dashed lines) and the maximum likelihood estimate based on the graphical model selected through neighborhood selection (green dotted lines). Each panel corresponds to a different value of the dimensionality. X-axes represent the value of ρ. The estimation errors are averaged over two hundred runs.

of the graphical model. Recall that the inverse covariance matrix is banded with nonzero entries increasing in magnitude with ρ. For large values of ρ, the task of identifying the correct graphical model is relatively easier. With a good graphical model chosen, refitting it with the maximum likelihood estimator could reduce biases often associated with regularization approaches. Such benefit diminishes for small values of ρ as identifying nonzero entries in the inverse covariance matrix becomes more difficult.

At last, we note that all three methods are relatively efficient to compute. For example, when $p = 200$ and $\rho = 0.5$, the averaged CPU time for the ℓ_1 penalized likelihood estimate is 0.53 seconds, for the neighborhood selection based method is 1.42 seconds, and for the proposed method is 3.21 seconds. Both the ℓ_1 penalized likelihood estimate and the neighborhood selection based method are computed using the graphical Lasso algorithm of Friedman, Hastie and Tibshirani (2008) which iteratively solves a sequence of p Lasso problems using a modified Lars algorithm (Efron et al., 2004). The algorithm is specially developed to take advantage of the sparse nature of the problem

and available in the `glasso` package in R. The proposed method is implemented in MATLAB using its general purpose interior-point algorithm based linear programming solver and could be further improved using more specialized algorithms (see, e.g., Asif, 2008).

5. Discussions

High dimensional (inverse) covariance matrix estimation is becoming more and more common in various scientific and technological areas. Most of the existing methods are designed to benefit from sparsity of the covariance matrix, and based on banding or thresholding the sample covariance matrix. Sparse models for the inverse covariance matrix, despite its practical appeal and close connection to graphical modeling, are more difficult to be taken advantage of due to heavy computational cost as well as the lack of a coherent theory on how such sparsity can be effectively exploited. In this paper, we propose an estimating procedure that addresses both challenges. The proposed method can be formulated using linear programming and therefore computed very efficiently. We also show that the resulting estimate enjoys nice probabilistic properties, which translates to sharp convergence rates in terms of matrix operator norms under a couple of common settings.

The choice of the tuning parameter δ is of great practical importance. Our theoretical developments have suggested reasonable choices of the tuning parameter and it seems to work well in the well-controlled simulation settings. In practice, however, a data-driven choice such as those determined by multi-fold cross-validation may yield improved performance.

We also note that the method can be easily extended to handle prior information regarding the sparsity patterns of the inverse covariance matrices. Such situations often arise in the context of, for example, genomics. In a typical gene expression experiment, tens of thousands of genes are often studied simultaneously. Among these genes, there are often known pathways which corresponding to conditional (in)dependence among a subset of the genes, or in our notation, variables. This can be naturally interpreted as some of the entries of Ω_0 being known to be nonzero or zero. Such prior information can be easily incorporated in our procedure. In particular, it suffices to set some of the entries of β to be exact zero apriori in (2). Likewise, if a particular entry of β is known to be nonzero, we can also opt to minimize the ℓ_1 norm of only the remaining entries.

6. Proofs

We now present the proofs to Theorems 1, 5, 6 and 7.

6.1 Proof of Theorem 1

We begin by comparing $\hat{\theta}_{(i)}$ with $\theta_{(i)}$. For brevity, we shall abbreviate the subscript (i) in what follows when no confusion occurs. Recall that $\theta = -\Omega^0_{-i,i}/\Omega^0_{ii}$ and

$$\hat{\theta} = \underset{\beta \in \mathcal{F}}{\operatorname{argmin}} \|\beta\|_{\ell_1},$$

where $\mathcal{F} = \{\beta : \|S_{-i,i} - S_{-i,-i}\beta\|_{\ell_\infty} \le \delta\}$. For a given $\Omega \in O(\nu, \eta, \tau)$, let $\Omega \in O$ and $\gamma = -\Omega_{-i,i}/\Omega_{ii}$. We first show that $\gamma \in \mathcal{F}$.

Lemma 8 *Under the event that* $\|S - \Sigma_0\|_{\max} < C_0 \lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}$,

$$\|S_{-i,i} - S_{-i,-i}\gamma\|_{\ell_\infty} \le \delta,$$

provided that

$$\delta \geq \eta\nu + C_0\tau\nu\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}.$$

Proof By the definition of $O(\nu,\eta,\tau)$, for any $j \neq i$,

$$\left|\Sigma^0_{j\cdot}\Omega_{\cdot i}\right| = \Omega_{ii}\left|\Sigma^0_{ji} - \Sigma^0_{j,-i}\gamma\right| \leq \|\Sigma_0\Omega - I\|_{\max} \leq \eta,$$

which implies that

$$\left\|\Sigma^0_{-i,i} - \Sigma^0_{-i,-i}\gamma\right\|_{\ell_\infty} = \max_{j\neq i}\left|\Sigma^0_{ji} - \Sigma^0_{j,-i}\gamma\right| \leq \Omega^{-1}_{ii}\eta \leq \lambda^{-1}_{\min}(\Omega)\eta \leq \eta\nu.$$

An application of the triangular inequality now yields

$$\begin{aligned}
\|S_{-i,i} - S_{-i,-i}\gamma\|_{\ell_\infty} &\leq \left\|S_{-i,i} - \Sigma^0_{-i,i}\right\|_{\ell_\infty} + \left\|\left(S_{-i,-i} - \Sigma^0_{-i,-i}\right)\gamma\right\|_{\ell_\infty} + \left\|\Sigma^0_{-i,i} - \Sigma^0_{-i,-i}\gamma\right\|_{\ell_\infty} \\
&\leq \|S - \Sigma_0\|_{\max} + \|S - \Sigma_0\|_{\max}\|\gamma\|_{\ell_1} + \eta\nu \\
&= \|S - \Sigma_0\|_{\max}\|\Omega_{\cdot i}\|_{\ell_1}/\Omega_{ii} + \eta\nu \\
&\leq \tau\nu\|S - \Sigma_0\|_{\max} + \eta\nu.
\end{aligned}$$

The claim now follows. ■

Now that $\gamma \in \mathcal{F}$, by the definition of $\hat{\theta}$,

$$\|\hat{\theta}\|_{\ell_1} \leq \|\gamma\|_{\ell_1} \leq \Omega^{-1}_{ii}\|\Omega_{\cdot i}\|_{\ell_1} - 1 \leq \lambda^{-1}_{\min}(\Omega)\|\Omega\|_{\ell_1} - 1 \leq \nu\tau - 1. \tag{11}$$

Write $\mathcal{J} = \{j : \gamma_j \neq 0\}$. Denote by $d_{\mathcal{J}} = \mathrm{card}(\mathcal{J})$. It is clear that $d_{\mathcal{J}} \leq \deg(\Omega)$. From (11),

$$0 \leq \|\gamma\|_{\ell_1} - \|\hat{\theta}\|_{\ell_1} \leq \|\hat{\theta}_{\mathcal{J}} - \gamma_{\mathcal{J}}\|_{\ell_1} - \|\hat{\theta}_{\mathcal{J}^c}\|_{\ell_1}.$$

Thus,

$$\begin{aligned}
\|\hat{\theta} - \gamma\|_{\ell_1} &= \|\hat{\theta}_{\mathcal{J}} - \gamma_{\mathcal{J}}\|_{\ell_1} + \|\hat{\theta}_{\mathcal{J}^c}\|_{\ell_1} \\
&\leq 2\|\hat{\theta}_{\mathcal{J}} - \gamma_{\mathcal{J}}\|_{\ell_1} \\
&\leq 2d^{1/2}_{\mathcal{J}}\|\hat{\theta}_{\mathcal{J}} - \gamma_{\mathcal{J}}\|_{\ell_2} \\
&\leq 2d^{1/2}_{\mathcal{J}}\|\hat{\theta} - \gamma\|_{\ell_2} \\
&\leq 2d^{1/2}_{\mathcal{J}}\lambda^{-1}_{\min}(\Sigma^0_{-i,-i})\left[\left(\hat{\theta} - \gamma\right)'\Sigma^0_{-i,-i}\left(\hat{\theta} - \gamma\right)\right]^{1/2} \\
&\leq 2\lambda^{-1}_{\min}(\Sigma_0)d^{1/2}_{\mathcal{J}}\left[\left(\hat{\theta} - \gamma\right)'\Sigma^0_{-i,-i}\left(\hat{\theta} - \gamma\right)\right]^{1/2} \\
&= 2\lambda_{\max}(\Omega_0)d^{1/2}_{\mathcal{J}}\left[\left(\hat{\theta} - \gamma\right)'\Sigma^0_{-i,-i}\left(\hat{\theta} - \gamma\right)\right]^{1/2}.
\end{aligned}$$

Observe that

$$\left(\hat{\theta} - \gamma\right)'\Sigma^0_{-i,-i}\left(\hat{\theta} - \gamma\right) \leq \|\hat{\theta} - \gamma\|_{\ell_1}\left\|\Sigma^0_{-i,-i}\left(\hat{\theta} - \gamma\right)\right\|_{\ell_\infty}.$$

Therefore,

$$\|\hat{\theta} - \gamma\|_{\ell_1} \leq 2\lambda_{\max}(\Omega_0)d^{1/2}_{\mathcal{J}}\|\hat{\theta} - \gamma\|^{1/2}_{\ell_1}\left\|\Sigma^0_{-i,-i}\left(\hat{\theta} - \gamma\right)\right\|^{1/2}_{\ell_\infty},$$

which implies that

$$\|\hat{\theta}-\gamma\|_{\ell_1} \leq 4\lambda_{\max}^2(\Omega_0) d_{\mathcal{J}} \left\|\Sigma^0_{-i,-i}\left(\hat{\theta}-\gamma\right)\right\|_{\ell_\infty}. \tag{12}$$

We now set up to further bound the last term on the right hand side. We appeal to the following result.

Lemma 9 *Under the event that* $\|S-\Sigma_0\|_{\max} < C_0\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}$, *we have*

$$\left\|\Sigma^0_{-i,-i}\left(\hat{\theta}-\gamma\right)\right\|_{\ell_\infty} \leq 2\delta,$$

provided that

$$\delta \geq \eta\nu + C_0\tau\nu\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}.$$

Proof By triangular inequality,

$$\left\|\Sigma^0_{-i,-i}\left(\hat{\theta}-\gamma\right)\right\|_{\ell_\infty} \leq \left\|\Sigma^0_{-i,-i}(\theta-\gamma)\right\|_{\ell_\infty} + \left\|\Sigma^0_{-i,-i}\left(\hat{\theta}-\theta\right)\right\|_{\ell_\infty}. \tag{13}$$

We begin with the first term on the right hand side. Recall that

$$\Sigma^0_{-i,i}\Omega^0_{ii} + \Sigma^0_{-i,-i}\Omega^0_{-i,-i} = \mathbf{0},$$

which implies that

$$\Sigma^0_{-i,-i}\theta = \Sigma^0_{-i,i}. \tag{14}$$

Hence,

$$\left\|\Sigma^0_{-i,-i}(\theta-\gamma)\right\|_{\ell_\infty} = \left\|\Sigma^0_{-i,i}-\Sigma^0_{-i,-i}\gamma\right\|_{\ell_\infty} = \Omega_{ii}^{-1}\left\|\Sigma^0_{-i,i}\Omega_{ii}+\Sigma^0_{-i,-i}\Omega_{-i,i}\right\|_{\ell_\infty} \leq \nu\eta.$$

We now turn to the second term on the right hand side of (13). Again by triangular inequality

$$\left\|\Sigma^0_{-i,-i}\left(\hat{\theta}-\theta\right)\right\|_{\ell_\infty} \leq \left\|\left(S_{-i,-i}-\Sigma^0_{-i,-i}\right)\hat{\theta}\right\|_{\ell_\infty} + \left\|S_{-i,-i}\hat{\theta}-\Sigma^0_{-i,-i}\theta\right\|_{\ell_\infty}.$$

To bound the first term on the right hand side, note that

$$\left\|\left(S_{-i,-i}-\Sigma^0_{-i,-i}\right)\hat{\theta}\right\|_{\ell_\infty} \leq \left\|S_{-i,-i}-\Sigma^0_{-i,-i}\right\|_{\max}\|\hat{\theta}\|_{\ell_1} \leq \|S-\Sigma_0\|_{\max}\|\hat{\theta}\|_{\ell_1}.$$

Also recall that

$$\left\|S_{-i,i}-S_{-i,-i}\hat{\theta}\right\|_{\ell_\infty} \leq \delta,$$

and $\Sigma^0_{-i,-i}\theta = \Sigma^0_{-i,i}$. Therefore, by triangular inequality and (14),

$$\left\|S_{-i,-i}\hat{\theta}-\Sigma^0_{-i,-i}\theta\right\|_{\ell_\infty} \leq \delta + \|\Sigma^0_{-i,i}-S_{-i,i}\|_{\ell_\infty} \leq \delta + \|S-\Sigma_0\|_{\max}.$$

To sum up,

$$\begin{aligned}
\left\|\Sigma^0_{-i,-i}\left(\hat{\theta}-\gamma\right)\right\|_{\ell_\infty} &\leq \delta+\nu\eta+\|S-\Sigma_0\|_{\max}+\|S-\Sigma_0\|_{\max}\|\hat{\theta}\|_{\ell_1} \\
&\leq \delta+\nu\eta+\|S-\Sigma_0\|_{\max}(1+\|\gamma\|_{\ell_1}) \\
&\leq \delta+\nu\eta+\|S-\Sigma_0\|_{\max}\|\Omega\|_{\ell_1}/\Omega_{ii} \\
&\leq \delta+\nu\eta+\|S-\Sigma_0\|_{\max}\|\Omega\|_{\ell_1}\lambda_{\min}^{-1}(\Omega) \\
&\leq \delta+\nu\eta+\tau\nu\|S-\Sigma_0\|_{\max},
\end{aligned}$$

which, under the event that $\|S-\Sigma_0\|_{\max} < C_0\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}$, can be further bounded by 2δ by Lemma 8. ■

Together with (12), Lemma 9 implies that for all $i = 1,\ldots,p$

$$\|\hat{\theta}-\gamma\|_{\ell_1} \le 8\lambda^2_{\max}(\Omega_0)d_{\mathcal{J}}\delta,$$

if $\|S-\Sigma_0\|_{\max} < C_0\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}$. We are now in position to bound $\|\tilde{\Omega}-\Omega_0\|_{\ell_1}$. We begin with the diagonal elements $|\tilde{\Omega}_{ii}-\Omega^0_{ii}|$.

Lemma 10 *Assume that* $\|S-\Sigma_0\|_{\max} < C_0\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}$ *and*

$$\delta\lambda_{\max}(\Omega_0)\left(\nu\tau + 8\lambda^2_{\max}(\Omega_0)\lambda^{-1}_{\min}(\Omega_0)d_{\mathcal{J}}\right) + \nu\tau\lambda_{\max}(\Omega_0)\lambda^{-2}_{\min}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1} \le c_0$$

for some numerical constant $0 < c_0 < 1$. *Then*

$$\left|\Omega^0_{ii}-\tilde{\Omega}_{ii}\right| \le \frac{1}{1-c_0}\left(\delta\lambda^2_{\max}(\Omega_0)\left(\nu\tau + 8\lambda^2_{\max}(\Omega_0)d_{\mathcal{J}}\lambda^{-1}_{\min}(\Omega_0)\right) + \nu\tau\lambda^{-2}_{\min}(\Omega_0)\lambda^2_{\max}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1}\right),$$

provided that

$$\delta \ge \eta\nu + C_0\tau\nu\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}.$$

Proof Recall that $\Sigma^0_{-i,i} = \Sigma^0_{-i,-i}\theta$. Therefore

$$\Omega^0_{ii} = \left(\Sigma^0_{ii} - 2\Sigma^0_{i,-i}\theta + \theta'\Sigma^0_{-i,-i}\theta\right)^{-1} = \left(\Sigma^0_{ii} - \Sigma^0_{i,-i}\theta\right)^{-1}.$$

Because

$$\tilde{\Omega}_{ii} = \left(S_{ii} - 2S_{i,-i}\hat{\theta} + \hat{\theta}'S_{-i,-i}\hat{\theta}\right)^{-1},$$

we have

$$\left|\tilde{\Omega}^{-1}_{ii} - \left(\Omega^0_{ii}\right)^{-1}\right| \le |S_{ii}-\Sigma^0_{ii}| + \left|\hat{\theta}'S_{-i,-i}\hat{\theta} - S_{i,-i}\hat{\theta}\right| + \left|S_{i,-i}\hat{\theta} - \Sigma^0_{i,-i}\theta\right|. \tag{15}$$

We now bound the three terms on the right hand side separately. It is clear that the first term can be bounded by $\|S-\Sigma_0\|_{\max}$. Recall that $\hat{\theta} \in \mathcal{F}$. Hence the second term can be bounded as follows:

$$\left|\hat{\theta}'S_{-i,-i}\hat{\theta} - S_{i,-i}\hat{\theta}\right| \le \left\|S_{-i,-i}\hat{\theta} - S_{-i,i}\right\|_{\ell_\infty}\|\hat{\theta}\|_{\ell_1} \le \delta\|\hat{\theta}\|_{\ell_1}.$$

The last term on the right hand side of (15) can also be bounded similarly.

$$\begin{aligned}
\left|S_{i,-i}\hat{\theta} - \Sigma^0_{i,-i}\theta\right| &\le \left|\left(S_{i,-i}-\Sigma^0_{i,-i}\right)\hat{\theta}\right| + \left|\Sigma^0_{i,-i}\left(\hat{\theta}-\theta\right)\right| \\
&\le \|S-\Sigma_0\|_{\max}\|\hat{\theta}\|_{\ell_1} + \|\Sigma^0_{i,-i}\|_{\ell_\infty}\|\hat{\theta}-\theta\|_{\ell_1} \\
&\le \|S-\Sigma_0\|_{\max}\|\hat{\theta}\|_{\ell_1} + \lambda_{\max}(\Sigma_0)\left(\|\hat{\theta}-\gamma\|_{\ell_1} + \|\gamma-\theta\|_{\ell_1}\right) \\
&= \|S-\Sigma_0\|_{\max}\|\hat{\theta}\|_{\ell_1} + \lambda^{-1}_{\min}(\Omega_0)\left(\|\hat{\theta}-\gamma\|_{\ell_1} + \|\gamma-\theta\|_{\ell_1}\right).
\end{aligned}$$

In summary, we have

$$\begin{aligned}
\left|\tilde{\Omega}^{-1}_{ii} - \left(\Omega^0_{ii}\right)^{-1}\right| &\le \|S-\Sigma_0\|_{\max} + \delta\|\hat{\theta}\|_{\ell_1} + \|S-\Sigma_0\|_{\max}\|\hat{\theta}\|_{\ell_1} \\
&\qquad + \lambda^{-1}_{\min}(\Omega_0)\left(\|\hat{\theta}-\gamma\|_{\ell_1} + \|\gamma-\theta\|_{\ell_1}\right) \\
&\le \nu\tau\|S-\Sigma_0\|_{\max} + \delta\|\hat{\theta}\|_{\ell_1} + \lambda^{-1}_{\min}(\Omega_0)\left(8\lambda^2_{\max}(\Omega_0)d_{\mathcal{J}}\delta + \|\gamma-\theta\|_{\ell_1}\right) \\
&\le \delta\left(\nu\tau + 8\lambda^2_{\max}(\Omega_0)\lambda^{-1}_{\min}(\Omega_0)d_{\mathcal{J}}\right) + \lambda^{-1}_{\min}(\Omega_0)\|\gamma-\theta\|_{\ell_1},
\end{aligned}$$

provided that $\|S-\Sigma_0\|_{\max} < C_0\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}$. Together with the fact that $\Omega^0_{ii} \leq \lambda_{\max}(\Omega_0)$, this yields

$$\left|\frac{\Omega^0_{ii}}{\tilde{\Omega}_{ii}} - 1\right| \leq \delta\lambda_{\max}(\Omega_0)\left(\nu\tau + 8\lambda^2_{\max}(\Omega_0)\lambda^{-1}_{\min}(\Omega_0)d_{\mathcal{J}}\right) + \lambda_{\max}(\Omega_0)\lambda^{-1}_{\min}(\Omega_0)\|\gamma-\theta\|_{\ell_1}.$$

Moreover, observe that

$$\begin{aligned}\|\gamma-\theta\|_{\ell_1} &\leq \left(\Omega^0_{ii}\right)^{-1}\|\Omega_{-i,i}-\Omega^0_{-i,i}\|_{\ell_1} + \Omega^{-1}_{ii}\left(\Omega^0_{ii}\right)^{-1}|\Omega_{ii}-\Omega^0_{ii}|\|\Omega_{-i,i}\|_{\ell_1} \\ &\leq \lambda^{-1}_{\min}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1} + \lambda^{-1}_{\min}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1}(\nu\tau-1) \\ &\leq \nu\tau\lambda^{-1}_{\min}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1}.\end{aligned}$$

Therefore,

$$\left|\frac{\Omega^0_{ii}}{\tilde{\Omega}_{ii}} - 1\right| \leq \delta\lambda_{\max}(\Omega_0)\left(\nu\tau + 8\lambda^2_{\max}(\Omega_0)\lambda^{-1}_{\min}(\Omega_0)d_{\mathcal{J}}\right) + \nu\tau\lambda_{\max}(\Omega_0)\lambda^{-2}_{\min}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1}, \quad (16)$$

which implies that

$$\frac{\Omega^0_{ii}}{\tilde{\Omega}_{ii}} \geq 1 - c_0.$$

Subsequently,

$$\tilde{\Omega}_{ii} \leq \frac{1}{1-c_0}\Omega^0_{ii} \leq \frac{1}{1-c_0}\lambda_{\max}(\Omega_0).$$

Together with (16), this implies

$$\begin{aligned}\left|\Omega^0_{ii}-\tilde{\Omega}_{ii}\right| &\leq \tilde{\Omega}_{ii}\left|\frac{\Omega^0_{ii}}{\tilde{\Omega}_{ii}} - 1\right| \\ &\leq \frac{1}{1-c_0}\delta\lambda^2_{\max}(\Omega_0)\left(\nu\tau + 8\lambda^2_{\max}(\Omega_0)\lambda^{-1}_{\min}(\Omega_0)d_{\mathcal{J}}\right) \\ &\qquad + \frac{1}{1-c_0}\nu\tau\lambda^{-2}_{\min}(\Omega_0)\lambda^2_{\max}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1}.\end{aligned}$$

■

We now turn to the off-diagonal entries of $\tilde{\Omega}-\Omega_0$.

Lemma 11 *Under the assumptions of Lemma 10, there exist positive constants C_1, C_2 and C_3 depending only on ν, τ, $\lambda_{\min}(\Omega_0)$ and $\lambda_{\max}(\Omega_0)$ such that*

$$\|\tilde{\Omega}_{-i,\cdot} - \Omega^0_{-i,\cdot}\|_{\ell_1} \leq (C_1 + C_2 d_{\mathcal{J}})\delta + C_3\|\Omega-\Omega_0\|_{\ell_1}.$$

Proof Note that

$$\begin{aligned}
\|\tilde{\Omega}_{-i,i}-\Omega^0_{-i,i}\|_{\ell_1} &= \left\|\tilde{\Omega}_{ii}\hat{\theta}-\Omega^0_{ii}\theta\right\|_{\ell_1} \\
&\le \Omega^0_{ii}\|\hat{\theta}-\theta\|_{\ell_1}+\left|\tilde{\Omega}_{ii}-\Omega^0_{ii}\right|\|\hat{\theta}\|_{\ell_1} \\
&\le \lambda^{-1}_{\min}(\Omega_0)\left(\|\hat{\theta}-\gamma\|_{\ell_1}+\|\gamma-\theta\|_{\ell_1}\right) \\
&\qquad +\frac{\nu\tau-1}{1-c_0}\delta\lambda^2_{\max}(\Omega_0)\left(\nu\tau+8\lambda^2_{\max}(\Omega_0)d_{\mathcal{J}}\lambda^{-1}_{\min}(\Omega_0)\right) \\
&\qquad +\frac{\nu\tau-1}{1-c_0}\nu\tau\lambda^{-2}_{\min}(\Omega_0)\lambda^2_{\max}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1} \\
&\le 8\nu^2 d_{\mathcal{J}}\lambda^{-1}_{\min}(\Omega_0)\delta+\nu\tau\lambda^{-2}_{\min}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1} \\
&\qquad +\frac{\nu\tau-1}{1-c_0}\delta\lambda^2_{\max}(\Omega_0)\left(\nu\tau+8\lambda^2_{\max}(\Omega_0)d_{\mathcal{J}}\lambda^{-1}_{\min}(\Omega_0)\right) \\
&\qquad +\frac{\nu\tau-1}{1-c_0}\nu\tau\lambda^{-2}_{\min}(\Omega_0)\lambda^2_{\max}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\tilde{\Omega}_{-i,\cdot}-\Omega^0_{-i,\cdot}\|_{\ell_1} &= \left|\Omega^0_{ii}-\tilde{\Omega}_{ii}\right|+\|\tilde{\Omega}_{-i,i}-\Omega^0_{-i,i}\|_{\ell_1} \\
&\le \delta\left(\frac{1}{1-c_0}\nu^2\tau^2\lambda^2_{\max}(\Omega_0)+8\left(1+\frac{\nu\tau}{1-c_0}\right)\lambda^2_{\max}(\Omega_0)d_{\mathcal{J}}\lambda^{-1}_{\min}(\Omega_0)\right) \\
&\qquad +\left(1+\frac{\nu\tau}{1-c_0}\lambda^2_{\max}(\Omega_0)\right)\nu\tau\lambda^{-2}_{\min}(\Omega_0)\|\Omega-\Omega_0\|_{\ell_1}.
\end{aligned}$$

■

From Lemma 11, it is clear that under the assumptions of Lemma 10,

$$\|\tilde{\Omega}-\Omega^0\|_{\ell_1} \le C\inf_{\Omega\in O(\nu,\eta,\tau)}\left(\|\Omega-\Omega_0\|_{\ell_1}+\deg(\Omega)\delta\right), \tag{17}$$

where $C=\max\{C_1,C_2,C_3\}$ is a positive constant depending only on ν,τ, $\lambda_{\min}(\Omega_0)$ and $\lambda_{\max}(\Omega_0)$. By the definition of $\hat{\Omega}$,

$$\|\hat{\Omega}-\tilde{\Omega}\|_{\ell_1} \le \|\tilde{\Omega}-\Omega_0\|_{\ell_1} \le C\inf_{\Omega\in O(\nu,\eta,\tau)}\left(\|\Omega-\Omega_0\|_{\ell_1}+\deg(\Omega)\delta\right).$$

An application of triangular inequality immediately gives

$$\begin{aligned}
\|\hat{\Omega}-\Omega_0\|_{\ell_1} &\le \|\hat{\Omega}-\tilde{\Omega}\|_{\ell_1}+\|\tilde{\Omega}-\Omega_0\|_{\ell_1} \\
&\le 2C\inf_{\Omega\in O(\nu,\eta,\tau)}\left(\|\Omega-\Omega_0\|_{\ell_1}+\deg(\Omega)\delta\right).
\end{aligned}$$

To complete the proof, we appeal to the following lemma showing that

$$\|S-\Sigma_0\|_{\max} \le C_0\lambda_{\max}(\Sigma_0)\sqrt{\frac{t+\log p}{n}},$$

for a numerical constant $C_0>0$, with probability at least $1-e^{-t}$. Taking $t=A\log p$ yields

$$\mathbb{P}\left\{\|S-\Sigma_0\|_{\max} < C_0\lambda_{\max}(\Sigma_0)((A+1)n^{-1}\log p)^{1/2}\right\} \ge 1-p^{-A}.$$

Lemma 12 *Assume that there exist constants $c_0 \geq 0$, and $T > 0$ such that for any $|t| \leq T$*

$$\mathbb{E}e^{tX_i^2} \leq c_0, \qquad i = 1, 2, \ldots, p.$$

Then there exists a constant $C > 0$ depending on c_0 and T such that

$$\|S - \Sigma_0\|_{\max} \leq C\sqrt{\frac{t + \log p}{n}}$$

with probability at least $1 - e^{-t}$ for all $t > 0$.

Proof Observe that S is invariant to $\mathbb{E}X$. We shall assume that $\mathbb{E}X = \mathbf{0}$ without loss of generality. Note that $S_{ij} = \mathbb{E}_n X_i X_j - \mathbb{E}_n X_i \mathbb{E}_n X_j$. We have

$$|S_{ij} - \Sigma^0_{ij}| \leq |\mathbb{E}_n X_i X_j - \mathbb{E}X_i X_j| + |\mathbb{E}_n X_i||\mathbb{E}_n X_j| =: \Delta_1 + \Delta_2.$$

We begin by bounding Δ_1.

$$\begin{aligned} \left|(\mathbb{E}_n - \mathbb{E})X_i X_j\right| &= \frac{1}{4}\left|(\mathbb{E}_n - \mathbb{E})\left((X_i + X_j)^2 - (X_i - X_j)^2\right)\right| \\ &\leq \frac{1}{4}\left|(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2\right| + \frac{1}{4}\left|(\mathbb{E}_n - \mathbb{E})(X_i - X_j)^2\right|. \end{aligned}$$

The two terms in the upper bound can be bounded similarly and we focus only on the first term. By the sub-Gaussianity of X_i and X_j, for any $|t| \leq T/4$,

$$\mathbb{E}e^{t(X_i + X_j)^2} \leq \mathbb{E}e^{2tX_i^2} e^{2tX_j^2} \leq \mathbb{E}^{1/2} e^{4tX_i^2} \mathbb{E}^{1/2} e^{4tX_j^2} \leq c_0.$$

In other words, $\{X_i + X_j : 1 \leq i, j \leq p\}$ are also sub-Gaussian. Observe that

$$\begin{aligned} \ln\left(\mathbb{E}e^{t[(X_i + X_j)^2 - \mathbb{E}(X_i + X_j)^2]}\right) &= \ln \mathbb{E}e^{t(X_i + X_j)^2} - t\mathbb{E}(X_i + X_j)^2 \\ &\leq \mathbb{E}\left[e^{t(X_i + X_j)^2} - t(X_i + X_j)^2 - 1\right], \end{aligned}$$

where we used the fact that $\ln u \leq u - 1$ for all $u > 0$. An application of the Taylor expansion now yields that there exist constants $c_1, T_1 > 0$ such that

$$\ln\left(\mathbb{E}e^{t[(X_i + X_j)^2 - \mathbb{E}(X_i + X_j)^2]}\right) \leq c_1 t^2$$

for all $|t| < T_1$. In other words,

$$\mathbb{E}e^{t[(X_i + X_j)^2 - \mathbb{E}(X_i + X_j)^2]} \leq e^{c_1 t^2}$$

for any $|t| < T_1$. Therefore,

$$\mathbb{E}e^{t[(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2]} \leq e^{c_1 t^2/n}.$$

By Markov inequality

$$\mathbb{P}\left\{(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2 \geq x\right\} \leq e^{-tx}\mathbb{E}e^{t[(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2]} \leq \exp\left(c_1 t^2/n - tx\right).$$

Taking $t = nx/2c_1$ yields

$$\mathbb{P}\left\{(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2 \geq x\right\} \leq \exp\left\{-\frac{nx^2}{4c_1}\right\}.$$

Similarly,

$$\mathbb{P}\left\{(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2 \leq -x\right\} \leq \exp\left\{-\frac{nx^2}{4c_1}\right\}.$$

Therefore,

$$\mathbb{P}\left\{\left|(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2\right| \geq x\right\} \leq 2\exp\left\{-\frac{nx^2}{4c_1}\right\}.$$

Following the same argument, one can show that

$$\mathbb{P}\left\{\left|(\mathbb{E}_n - \mathbb{E})(X_i - X_j)^2\right| \geq x\right\} \leq 2\exp\left\{-\frac{nx^2}{4c_1}\right\}.$$

Note also that this inequality holds trivially when $i = j$. In summary, we have

$$\begin{aligned}\mathbb{P}\{\Delta_1 \geq x\} &\leq \mathbb{P}\left\{\left|(\mathbb{E}_n - \mathbb{E})(X_i + X_j)^2\right| \geq 2x\right\} + \mathbb{P}\left\{\left|(\mathbb{E}_n - \mathbb{E})(X_i - X_j)^2\right| \geq 2x\right\} \\ &\leq 4\exp\left\{-c_1^{-1}nx^2\right\}.\end{aligned}$$

Now consider Δ_2.

$$\mathbb{E}e^{t\mathbb{E}_n X_i \mathbb{E}_n X_j} \leq \mathbb{E}^{1/2} e^{2t\mathbb{E}_n X_i} \mathbb{E}^{1/2} e^{2t\mathbb{E}_n X_j} \leq \max_{1\leq i\leq p} \mathbb{E}e^{2t\mathbb{E}_n X_i}.$$

Following a similar argument as before, we can show that there exist constants $c_2, T_2 > 0$ such that

$$\mathbb{E}e^{2t\mathbb{E}_n X_i} \leq e^{c_2 t^2/n}$$

for all $|t| < T_2$. This further leads to, similar to before,

$$\mathbb{P}\{\Delta_1 \geq x\} \leq 2\exp\left\{-c_2^{-1}nx^2\right\}.$$

To sum up,

$$\begin{aligned}\mathbb{P}\left\{\|S - \Sigma_0\|_{\max} \geq x\right\} &\leq p^2 \max_{1\leq i,j\leq p} \mathbb{P}\left\{|S_{ij} - \Sigma^0_{ij}\| \geq x\right\} \\ &\leq p^2\left[\mathbb{P}\left(\Delta_1 \geq x/2\right) + \mathbb{P}\left(\Delta_2 \geq x/2\right)\right] \\ &\leq 4p^2\exp\left\{-c_3 nx^2\right\}.\end{aligned}$$

for some constant $c_3 > 0$. The claimed result now follows. ■

6.2 Proof of Theorem 5

First note that the claim follows from

$$\inf_{\bar{\Omega}} \sup_{\Omega_0 \in \mathcal{M}_1(\tau_0,\nu_0,d)} \mathbb{E}\big\|\bar{\Omega}-\Omega_0\big\|_{\ell_1} \geq C' d\sqrt{\frac{\log p}{n}} \tag{18}$$

for some constant $C' > 0$. We establish the minimax lower bound (18) using the tools from Korostelev and Tsybakov (1993), which is based upon testing many hypotheses as well as statistical applications of Fano's lemma and the Varshamov-Gilbert bound. More specifically, it suffices to find a collection of inverse covariance matrices $\mathcal{M}' = \{\Omega_1, \ldots, \Omega_{K+1}\} \subset \mathcal{M}_1(\tau_0, \nu_0, d)$ such that

(a) for any two distinct members $\Omega_j, \Omega_k \in \mathcal{M}'$, $\|\Omega_j - \Omega_k\|_{\ell_1} > Ad(n^{-1}\log p)^{1/2}$ for some constant $A > 0$;

(b) there exists a numerical constant $0 < c_0 < 1/8$ such that

$$\frac{1}{K}\sum_{k=1}^{K} \mathcal{K}(\mathcal{P}(\Omega_k), \mathcal{P}(\Omega_{K+1})) \leq c_0 \log K,$$

where $\mathcal{K}$ stands for the Kullback-Leibler divergence and $\mathcal{P}(\Omega)$ is the probability measure $\mathcal{N}(\mathbf{0}, \Omega)$.

To construct $\mathcal{M}'$, we assume that $d(n^{-1}\log p)^{1/2} < 1/2$, $\tau_0, \nu_0 > 2$ without loss of generality. As shown by Birgé and Massart (1998), from the Varshamov-Gilbert bound, there is a set of binary vectors of length $p-1$, $\mathcal{B} = \{b_1, \ldots, b_K\} \subset \{0,1\}^{p-1}$ such that (i) there are d ones in a vector b_j for any $j = 1, \ldots, p-1$; (ii) the Hamming distance between b_j and b_k is at least $d/2$ for any $j \neq k$; (iii) $\log K > 0.233 d \log(p/d)$. We now take Ω_k for $k = 1, \ldots, K$ as follows. It differs from the identity matrix only by its first row and column. More specifically, $\Omega^k_{11} = 1$, $\Omega^k_{-1,1} = (\Omega^k_{1,-1})' = a_n b_k$, $\Omega^k_{-1,-1} = I_{p-1}$, that is,

$$\Omega_k = \begin{pmatrix} 1 & a_n b_k' \\ a_n b_k & I \end{pmatrix},$$

where $a_n = a_0(n^{-1}\log p)^{1/2}$ with a constant $0 < a_0 < 1$ to be determined later. Finally, we take $\Omega_{K+1} = I$. It is clear that Condition (a) is satisfied with this choice of $\mathcal{M}'$ and $A = a_0/2$. It remains to verify Condition (b). Simple algebraic manipulations yield that for any $1 \leq k \leq K$,

$$\mathcal{K}(\mathcal{P}(\Omega_k), \mathcal{P}(\Omega_{K+1})) = -\frac{n}{2}\log\det(\Omega_k) = -\frac{n}{2}\log(1 - a_n^2 b_k' b_k).$$

Recall that $a_n^2 b_k' b_k = d a_n^2 < d^2 a_n^2 < 1/4$. Together with the fact that $-\log(1-x) \leq \log(1+2x) \leq 2x$ for $0 < x < 1/2$, we have

$$\mathcal{K}(\mathcal{P}(\Omega_k), \mathcal{P}(\Omega_{K+1})) \leq n d a_n^2.$$

By setting a_0 small enough, this can be further bounded by $0.233 c_0 d \log(p/d)$ and subsequently $c_0 \log K$. The proof is now completed. ∎

6.3 Proof of Theorem 6

We prove the theorem by applying the oracle inequality from Theorem 1. To this end, we need to find an "oracle" inverse covariance matrix Ω. Let

$$\Omega_{ij} = \Omega^0_{ij} \mathbf{1}\left(\left|\Omega^0_{ij}\right| \geq \zeta\right),$$

where $\zeta > 0$ is to be specified later. We now verify that $\Omega \in O(\nu, \eta, \tau)$ with appropriate choices of the three parameters.

First observe that

$$\begin{aligned}
\|\Omega - \Omega_0\|_{\ell_1} &\leq \max_{1\leq i\leq p} \sum_{j=1}^{p} \left|\Omega^0_{ij}\right| \mathbf{1}\left(\left|\Omega^0_{ij}\right| \leq \zeta\right) \\
&\leq \zeta^{1-\alpha} \max_{1\leq i\leq p} \sum_{j=1}^{p} \left|\Omega^0_{ij}\right|^{\alpha} \mathbf{1}\left(\left|\Omega^0_{ij}\right| \leq \zeta\right) \\
&\leq M\zeta^{1-\alpha}.
\end{aligned}$$

Thus,

$$\nu_0^{-1} - M\zeta^{1-\alpha} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq \nu_0 + M\zeta^{1-\alpha}.$$

In particular, setting ζ small enough such that $M\zeta^{1-\alpha} < (2\nu_0)^{-1}$ yields

$$(2\nu_0)^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq 2\nu_0.$$

We can therefore take $\nu = 2\nu_0$.

Now consider the approximation error $\|\Sigma_0\Omega - I\|_{\max}$. Note that the (i, j) entry of $\Sigma_0\Omega - I = \Sigma_0(\Omega - \Omega_0)$ can be bounded as follows

$$\begin{aligned}
\left|\sum_{k=1}^{p} \Sigma^0_{ik}\Omega^0_{kj} \mathbf{1}\left(\left|\Omega^0_{kj}\right| \leq \zeta\right)\right| &\leq \sum_{k=1}^{p} \left|\Sigma^0_{ik}\right| \left|\Omega^0_{kj}\right| \mathbf{1}\left(\left|\Omega^0_{kj}\right| \leq \zeta\right) \\
&\leq \zeta \max_{1\leq k\leq p} \sum_{k=1}^{p} \left|\Sigma^0_{ik}\right| \\
&= \zeta\|\Sigma_0\|_{\ell_1}.
\end{aligned}$$

This implies that

$$\|\Sigma_0\Omega - I\|_{\max} \leq \zeta\|\Sigma_0\|_{\ell_1}.$$

In other words, we can take $\eta = \zeta\|\Sigma_0\|_{\ell_1}$.

Furthermore, it is clear that we can take $\|\Omega\|_{\ell_1} \leq \|\Omega_0\|_{\ell_1}$. Therefore, by Theorem 1, there exist constants $C_1, C_2 > 0$ depending only on $\|\Omega_0\|_{\ell_1}$, $\|\Sigma_0\|_{\ell_1}$, and ν_0 such that for any

$$\delta \geq C_1 \left(\zeta + \sqrt{\frac{(A+1)\log p}{n}}\right) \qquad (i = 1, 2, \ldots, p),$$

we have

$$\|\hat{\Omega} - \Omega_0\|_{\ell_1} \leq C_2 \left(M\zeta^{1-\alpha} + \deg(\Omega)\delta\right)$$

with probability at least $1-p^{-A}$. Now note that

$$\deg(\Omega) = \max_{1\le i\le p} \sum_{j=1}^{p} \mathbf{1}\left(\left|\Omega_{ij}^0\right| \ge \zeta\right) \le M\zeta^{-\alpha}.$$

The claimed the results then follows by taking $\zeta = C_3((A+1)n^{-1}\log p)^{1/2}$ for a small enough constant $C_3 > 0$ such that $M\zeta^{1-\alpha} < (2\nu_0)^{-1}$.

6.4 Proof of Theorem 7

Assume that $M(n^{-1}\log p)^{(1-\alpha)/2} < 1/2$, $\tau_0, \nu_0 > 2$ without loss of generality. Similar to Theorem 5, there eixs a collection of inverse covariance matrices $\mathcal{M}' = \{\Omega_1, \ldots, \Omega_{K+1}\} \subset \mathcal{M}_2$ such that

(a) for any two distinct members $\Omega_j, \Omega_k \in \mathcal{M}'$, $\|\Omega_j - \Omega_k\|_{\ell_1} > AM(n^{-1}\log p)^{(1-\alpha)/2}$ for some constant $A > 0$;

(b) there exists a numerical constant $0 < c_0 < 1/8$ such that

$$\frac{1}{K}\sum_{k=1}^{K} \mathcal{K}(\mathcal{P}(\Omega_k), \mathcal{P}(\Omega_{K+1})) \le c_0 \log K.$$

To this end, we follow the same construction as in the proof of Theorem 5 by taking

$$d = \left\lfloor M\left(\frac{\log p}{n}\right)^{-\frac{\alpha}{2}} \right\rfloor,$$

and $\lfloor x \rfloor$ stands for the integer part of x. First, we need to show that $\mathcal{M}' \subset \mathcal{M}_2$. Because $d(n^{-1}\log p)^{1/2} < 1/2$, it is clear that the bounded eigenvalue condition and $\|\Omega_k\|_{\ell_1} < \tau_0$ can ensured by setting a_0 small enough. It is also obvious that

$$\max_{1\le j\le p} \sum_{i=1}^{p} |\Omega_{ij}^k|^{\alpha} \le M.$$

It remains to check that $\|\Sigma_k\|_{\ell_1}$ is bounded. By the block matrix inversion formula

$$\Sigma_k = \begin{pmatrix} \frac{1}{1-a_n^2 b_k' b_k} & -\frac{a_n}{1-a_n^2 b_k' b_k} b_k' \\ -\frac{a_n}{1-a_n^2 b_k' b_k} b_k & I + \frac{a_n^2}{1-a_n^2 b_k' b_k} b_k b_k' \end{pmatrix}.$$

It can then be readily checked that $\|\Sigma_k\|_{\ell_1}$ can also be bounded from above by setting a_0 small enough.

Next we verify Conditions (a) and (b). It is clear that Condition (a) is satisfied with this choice of $\mathcal{M}'$ and $A = a_0/2$. It remains to verify Condition (b). Simple algebraic manipulations yield that for any $1 \le k \le K$,

$$\mathcal{K}(\mathcal{P}(\Omega_k), \mathcal{P}(\Omega_{K+1})) = -\frac{n}{2}\log\det(\Omega_k) = -\frac{n}{2}\log(1 - a_n^2 b_k' b_k).$$

Recall that $a_n^2 b_k' b_k = d a_n^2 < d^2 a_n^2 < 1/4$. Together with the fact that $-\log(1-x) \le \log(1+2x) \le 2x$ for $0 < x < 1/2$, we have

$$\mathcal{K}(\mathcal{P}(\Omega_k), \mathcal{P}(\Omega_{K+1})) \le n d a_n^2.$$

By setting a_0 small enough, this can be further bounded by $0.233c_0d\log(p/d)$ and subsequently $c_0\log K$. The proof of (9) is now completed.

The proof of (10) follows from a similar argument. We essentially construct the same subset $\mathcal{M}'$ but with

$$\Sigma_k = \begin{pmatrix} 1 & a_n b_k' \\ a_n b_k & I \end{pmatrix}.$$

The only difference from before is the calculation of Kullback-Leibler divergence, which in this case is

$$\mathcal{K}(\mathcal{P}(\Sigma_k), \mathcal{P}(\Sigma_{K+1})) = \frac{n}{2}\left(\text{trace}(\Omega_k) + \log\det(\Sigma_k) - p\right)$$

where

$$\Omega_k = \begin{pmatrix} \frac{1}{1-a_n^2 b_k' b_k} & -\frac{a_n}{1-a_n^2 b_k' b_k} b_k' \\ -\frac{a_n}{1-a_n^2 b_k' b_k} b_k & I + \frac{a_n^2}{1-a_n^2 b_k' b_k} b_k b_k' \end{pmatrix}.$$

Therefore, $\text{trace}(\Omega_k) = p + 2da_n^2/(1-da_n^2)$. Together with the fact that $\det(\Sigma_k) = 1 - da_n^2$, we conclude that

$$\mathcal{K}(\mathcal{P}(\Sigma_k), \mathcal{P}(\Sigma_{K+1})) = \frac{n}{2}\left(\frac{2da_n^2}{(1-da_n^2)} + \log(1-da_n^2)\right) \leq \frac{1}{2} n d a_n^2.$$

where we used the fact that $\log(1-x) \leq -x$. The rest of the argument proceeds in the same fashion as before.

Acknowledgments

This was supported in part by NSF grant DMS-0846234 (CAREER) and a grant from Georgia Cancer Coalition. The author wish to thank the editor and three anonymous referees for their comments that help greatly improve the manuscript.

References

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, London, 2003.

M. Asif. *Primal dual pursuit: a homotopy based algorithm for the Dantzig selector*. Master Thesis, School of Electrical and Computer Enginnering, Georgia Institute of Technology, 2008.

O. Banerjee, L. El Ghaoui and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485-516, 2008.

P. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:199-227, 2008a.

P. Bickel and E. Levina. Covariance regularization by thresholding. *Annals of Statistics*, 36:2577-2604, 2008b.

P. Bickel, Y. Ritov and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705-1732, 2009.

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329-375, 1998.

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373-384, 1995.

T.T. Cai, C. Zhang, and H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38:2118-2144, 2010.

E.J. Candés and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313-2351, 2007.

A. d'Aspremont, O. Banerjee and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and it Applications*, 30:56-66, 2008.

A. Dempster. Covariance selection. *Biometrika*, 32:95-108, 1972.

X. Deng and M. Yuan. Large Gaussian covariance matrix estimation with Markov structures. *Journal of Computational and Graphical Statistics*, 18:640-657, 2008.

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

B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407-499, 2004.

N. El Karoui. Operator norm consistent estimation of large dimensional sparse covariance matrices. *Annals of Statistics*, 36:2717-2756, 2008.

J. Fan, Y. Fan and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186-197, 2008.

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

J. Huang, N. Liu, M. Pourahmadi and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93:85-98, 2006.

A. Korostelev and A. Tsybakov. *Minimax Theory of Image Reconstruction*. Springer, New York, 1993.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrices estimation. *Annals of Statistics*, 37:4254-4278, 2009.

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

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365-411, 2004.

E. Levina, A.J. Rothman and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics*, 2:245-263, 2007.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436-1462, 2006.

R. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, London, 2005.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86:677-690, 1999.

M. Pourahmadi. Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*, 87:425-435, 2000.

P. Ravikumar, G. Raskutti, M. Wainwright and B. Yu. Model selection in Gaussian graphical models: high-dimensional consistency of ℓ_1-regularized MLE. In *Advances in Neural Information Processing Systems (NIPS)* 21, 2008.

P. Ravikumar, M. Wainwright, G. Raskutti and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Technical Report*, 2008.

G. Rocha, P. Zhao and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation. *Technical Report*, 2008.

A. Rothman, P. Bickel, E. Levina and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494-515, 2008.

A. Rothman, E. Levina and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104:177-186, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267-288, 1996.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, Chichester, 1990.

W. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90:831-844, 2003.

M. Yuan. Efficient computation of the ℓ_1 regularized solution path in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17:809-826, 2008.

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