

An empirical estimator for the sparsity of a large covariance matrix under multivariate normal assumptions

Binyan Jiang

Received: 4 April 2013 / Revised: 23 November 2013 / Published online: 6 March 2014
© The Institute of Statistical Mathematics, Tokyo 2014

Abstract Large covariance or correlation matrix is frequently assumed to be sparse in that a number of the off-diagonal elements of the matrix are zero. This paper focuses on estimating the sparsity of a large population covariance matrix using a sample correlation matrix under multivariate normal assumptions. We show that sparsity of a population covariance matrix can be well estimated by thresholding the sample correlation matrix. We then propose an empirical estimator for the sparsity and show that it is closely related to the thresholding methods. Upper bounds for the estimation error of the empirical estimator are given under mild conditions. Simulation shows that the empirical estimator can have smaller mean absolute errors than its main competitors. Furthermore, when the dimension of the covariance matrix is very large, we propose a generalized empirical estimator using simple random sampling. It is shown that the generalized empirical estimator can still estimate the sparsity well while the computation complexity can be greatly reduced.

Keywords Adaptive thresholding · Large correlation matrix · Large covariance matrix · Simple random sampling · Sparsity · Thresholding

1 Introduction

Large dimensional sparse covariance matrix estimation is frequently encountered in many fields during the past decades. The number of variables p may be much larger

This research is partially supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

B. Jiang (✉)
Living Analytics Research Centre (LARC), Heinz College, Carnegie Mellon University,
5000 Forbes Ave, Pittsburgh, PA 15213, USA
e-mail: stats.jby@gmail.com

than the sample size n and the population matrix is usually assumed to be sparse in that a number of the off-diagonal elements are zero. In this paper, sparsity of a matrix is defined as the proportion of zero in the off-diagonal elements of the matrix as p tends to infinity. Although sparsity assumptions are frequently used in many papers, the problem of estimating the sparsity of a covariance matrix is seldom studied.

Let X_1, \dots, X_n be n iid p dimensional multivariate normal random vectors with mean $\mathbf{0}$ and covariance matrix $\Sigma_{p \times p} = (\sigma_{ij})_{p \times p}$. The sample covariance matrix is given by

$$S = (s_{ij})_{p \times p} = \frac{1}{n} \sum_{i=1}^n X_i X_i^T. \quad (1)$$

Given the sample covariance matrix, one popular approach in estimating a sparse Σ is to use thresholded covariance matrices as estimators. More specifically, a thresholding estimator $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p \times p}$ is defined by:

$$\hat{\sigma}_{ij} = \hat{\sigma}_{ji} = T_{ij}(s_{ij}),$$

where $T_{ij}(s)$, $1 \leq i < j \leq p$, are general thresholding functions. For example [Bickel and Levina \(2008\)](#) and [El Karoui \(2008\)](#) considered hard thresholding functions $T_{ij}(s) = s\mathcal{I}(|s| > t)$, $1 \leq i < j \leq p$, where $\mathcal{I}(\cdot)$ is the indicator function. The thresholding parameter t controls the sparsity of the estimator $\hat{\Sigma}$. [Rothman et al. \(2009\)](#) considered more general thresholding functions possessing shrinkage properties. Overall, the threshold in [Bickel and Levina \(2008\)](#) or [Rothman et al. \(2009\)](#) is a universal-threshold in that a same thresholding parameter t is used for every off-diagonal elements in S . Lately, [Cai and Liu \(2011\)](#) proposed an adaptive thresholding method which is applicable when the p elements in X_i are not homoscedastic. They used different thresholding parameters for different s_{ij} depending on the variances of s_{ij} . For example, considering hard thresholding, for any $1 \leq i < j \leq p$, they set $T_{ij}(s_{ij}) = s_{ij}\mathcal{I}(|s_{ij}| > t\hat{\theta}_{ij}^{1/2})$, where $\hat{\theta}_{ij}$ is an estimator of the variance of s_{ij} and t is a thresholding parameter which determines the sparsity of the resulting estimator. Cross validation methods were used for finding a data-dependent thresholding parameter t in the above literatures.

Another popular approach in estimating a sparse covariance matrix is the penalization method, where estimators are obtained by minimizing penalized loss functions; see [Rothman \(2012\)](#) and the references cited therein. Similar to the thresholding approach, sparsity of these estimators relies on the choice of penalization parameters. These penalization parameters are also usually determined using cross validation methods.

Although thresholding approach and penalization approach are able to obtain sparse estimators for the covariance matrix, the resulting sparsity of the estimators may not be close to the sparsity of the true population covariance matrix in finite sample estimation. In fact, different loss functions used in determining thresholding parameters or penalization parameters in cross validation may result in different sparsity. In addition, the sparsity of the population covariance matrix is assumed to be tending to one

in some papers. Instead of estimating a sparse Σ , in this paper, we are trying to answer a relatively basic question: how sparse is the population covariance matrix?

Let ω be the proportion of zero in the off-diagonal elements of the population covariance matrix. A good estimator of ω can be used to check some sparsity assumptions. For example, we would know the assumption that ω tends to 1 is not appropriate if the consistent estimator we obtain is not close to 1. Another possible application of a good estimator of sparsity is in finding data-dependent thresholds in the thresholding approach. We can choose the thresholding parameter such that the sparsity of the thresholded sample covariance matrix equals the estimated sparsity. This way of determining the thresholding parameter is computationally more efficient than cross validation methods in [Bickel and Levina \(2008\)](#), [Rothman et al. \(2009\)](#) and [Cai and Liu \(2011\)](#) especially when p is very large. Similarly, when thresholding approach is applied to the sample correlation matrix in estimating the covariance structure as in [Jiang and Loh \(2012\)](#), the thresholding parameter can be determined based on a good estimator of ω .

[Jiang and Loh \(2012\)](#) proposed a method of moments estimator for ω . Their assumptions are similar to those in this paper but the methodology is totally different. In this paper, we propose an empirical estimator $\hat{\omega}_{em}(g)$ for ω . We show that under mild conditions, $E|\hat{\omega}_{em}(g) - \omega| = O\{(\log n/n)^{1/2} \vee p^{-1/2}\}$. Here $a \vee b = \max(a, b)$ for any constants a, b . Under multivariate normal assumptions, the rate in this upper bound is better than the one given in Theorem 1 of [Jiang and Loh \(2012\)](#). In addition, when p is very large, to further reduce the computation complexity, we propose a generalized empirical estimator $\hat{\omega}_{em}^m(\mathcal{S}_m, g)$ using simple random sampling. Let \mathcal{S}_m be a random subset generated by simple random sampling (without replacement) from the index set $\{(i, j) : 1 \leq i < j \leq p\}$ such that the cardinality of \mathcal{S}_m equals m . From classical theory in sampling (see for example Chap. 3 of [Thompson 1997](#)) we know that given Σ , the proportion of zero elements in the set $\{\sigma_{ij} : (i, j) \in \mathcal{S}_m\}$ is an unbiased estimator of the proportion of zero elements in the set $\{\sigma_{ij} : 1 \leq i < j \leq p\}$. Motivated by this, in Sect. 4 we propose a generalized empirical estimator $\hat{\omega}_{em}^m(\mathcal{S}_m, g)$, which to some degree is an empirical estimator of the sparsity of the set $\{\sigma_{ij} : (i, j) \in \mathcal{S}_m\}$. We show that under mild conditions, $E|\hat{\omega}_{em}^m(\mathcal{S}_m, g) - \omega| = O\{(\log n/n)^{1/2} \vee (m \wedge p)^{-1/2}\}$, where $a \wedge b = \min(a, b)$ for any constants a, b . From this upper bound we know that the generalized empirical estimator can still estimate ω very well while the computation complexity can be largely reduced. Particularly, if we choose $m \asymp p$, by comparing the upper bounds we obtained for the empirical estimator $\hat{\omega}_{em}(g)$ and the generalized empirical estimator $\hat{\omega}_{em}^m(\mathcal{S}_m, g)$, we immediately have that the rates of the bounds for both estimators are the same. On the other hand, when $m \asymp p$, from the definition of $\hat{\omega}_{em}(g)$ and $\hat{\omega}_{em}^m(\mathcal{S}_m, g)$ as in (4) and (6), we know that the number of terms in the summation of (4) is quadratic in p while the number of terms in the summation of (6) is linear in p , which is computationally more efficient especially when p is large.

The rest of the paper is organized as follows. In Sect. 2, we show that consistent estimators for the sparsity of the population correlation matrix can be obtained by thresholding the sample correlation matrix. In addition, upper bounds under L_1 -loss are established for some thresholding-based estimators. In Sect. 3, we propose an empirical estimator $\hat{\omega}_{em}(g)$. We show that $\hat{\omega}_{em}(g)$ is closely related to the thresholding-based estimators and possessing all the properties given in Sect. 2. In Sect. 4, we propose a

generalized empirical estimator $\hat{\omega}_{\text{em}}^m(\mathcal{S}_m, g)$ using simple random sampling. Section 5 provides some simulation studies with comparison to methods in [Bickel and Levina \(2008\)](#), [Rothman et al. \(2009\)](#), [Cai and Liu \(2011\)](#) and [Jiang and Loh \(2012\)](#).

2 Estimating the sparsity by thresholding the population correlation matrix

Suppose X_1, \dots, X_n are n independent and identically distributed p dimensional multivariate normal random vectors with mean $\mathbf{0}$ and covariance matrix $\Sigma_{p \times p} = (\sigma_{ij})_{p \times p}$. Define the sample covariance matrix S as in (1). Denote the population correlation matrix as $\Gamma = (\rho_{ij})_{p \times p}$ and the sample correlation matrix as $R = (r_{ij})_{1 \leq i, j \leq p}$ where $\rho_{ij} = \sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{1/2}$, $r_{ij} = s_{ij}/(s_{ii}s_{jj})^{1/2}$, $1 \leq i, j \leq p$.

2.1 Assumptions on the prior

Suppose Γ has a prior distribution satisfying the following assumptions:

Assumption 1 For each $1 \leq j < k \leq p$, the prior cumulative distribution function of each ρ_{jk} has the form

$$F_\rho(x) = \omega \mathcal{I}(0 \leq x \leq 1) + \sum_{i=1}^v \omega_i \mathcal{I}(\mu_i \leq x \leq 1) + \left(1 - \omega - \sum_{i=1}^v \omega_i\right) \int_{-1}^x g(x) dx,$$

where v is a nonnegative integer, $\omega, \omega_1, \dots, \omega_v$ are positive constants satisfying $\omega + \sum_{i=1}^v \omega_i \leq 1$, μ_1, \dots, μ_v are nonzero constants in $(-1, 1)$ and g is an unknown probability density function on $(-1, 1)$ such that $\sup_{\rho \in (-1, 1)} g(\rho) < \infty$. For simplicity, we assume that $0 < \omega < 1$ and $v = 0$. Results in this paper can be generalized to the case that $v \geq 1$.

Assumption 2 Let $\mathcal{F}_{ij} = \sigma(\rho_{ij})$ denote the σ -field generated by ρ_{ij} . Define for all $1 \leq i, j, s, t \leq p$,

$$\alpha(\rho_{ij}, \rho_{st}) = \sup_{A \in \mathcal{F}_{ij}, B \in \mathcal{F}_{st}} |P(A \cap B) - P(A)P(B)|.$$

Assume that, as $p \rightarrow \infty$,

$$1/p^4 \sum_{i, j, s, t: \text{all distinct}} \alpha(\rho_{ij}, \rho_{st}) \rightarrow 0.$$

Assumption 3 For all $1 \leq i, j, s, t \leq p$ such that $\{i, j\} \cap \{s, t\} = \emptyset$, we have $\alpha(\rho_{ij}, \rho_{st}) = O(p^{-1})$.

These three assumptions are similar to those in [Jiang and Loh \(2012\)](#). Assumption 3 is stronger than Assumption 2. Let $\Sigma = LL^T$ be the Cholesky decomposition of Σ with L a lower triangular matrix. If the rows of L are independent of each other, Assumption 3 is then satisfied if Γ is the correlation matrix corresponding to Σ . In addition, Assumptions 1 and 3 are satisfied by adding a random permutation to the indices. Model 5.3 in Sect. 5 is an example of this. Under Assumptions 1 and 2, the

proportion of zero in the off-diagonal elements of Σ is tending to ω as p tends to infinity; see for example (7) in the Appendix. Sparsity of Σ is then quantified by ω . In the next section, we show that ω can be well estimated by thresholding the sample correlation matrix.

2.2 Sparsity estimators based on thresholding

The universal thresholding approach had been applied to the sample correlation matrix too. In the literature it is commonly assumed that for any $1 \leq i < j \leq p$, $|\rho_{ij}|$ is either equal to zero or greater than $k(n, p)$, which is a constant depending on p and n . For example El Karoui (2008) considered estimating the population correlation matrix by thresholding the sample correlation matrix and it was assumed that $k(n, p) = O(n^{-v})$ for a constant $v > 1/2$. Jiang (2013) found that to obtain covariance selection consistency by thresholding the sample correlation matrix, $k(n, p)$ should be at least of order $\sqrt{\log p/n}$. Here in this section, we construct sparsity estimators by thresholding the sample correlation matrix and provide some results on their asymptotic performance. Particularly, we do not assume that the nonzero $|\rho_{ij}|$ are greater than some constant $k(n, p)$. For a threshold $t > 0$, define

$$\hat{\omega}(t) = \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}(|r_{ij}| < t). \tag{2}$$

Recall that for any constants a and b , we denote $a \vee b = \max(a, b)$, and $a \wedge b = \min(a, b)$. The following theorem indicates that when t converges to zero slowly enough, $\hat{\omega}(t)$ is consistent in estimating ω .

Theorem 1 *Let $\hat{\omega}(t)$ be defined as in (2). Under Assumptions 1 and 2, for any t satisfying $t \rightarrow 0$ and $tn^{1/2} \rightarrow \infty$, we have $\hat{\omega}(t) \rightarrow \omega$ in probability when $n \wedge p \rightarrow \infty$.*

Theorem 1 is true when $t = C_0(\log n/n)^{1/2}$ for some constant $C_0 > 0$. The following theorem provides an upper bound for the estimation error of $\hat{\omega}(t)$ when $\{(\log n - \log \log n)/n\}^{1/2} \leq t < C_1(\log n/n)^{1/2}$ for some constant $C_1 > 0$.

Theorem 2 *Under Assumptions 1 and 3, for any threshold t such that $\{(\log n - \log \log n)/n\}^{1/2} \leq t < C_1(\log n/n)^{1/2}$ for some constant $C_1 > 0$, when $n \wedge p \rightarrow \infty$, we have*

$$E|\hat{\omega}(t) - \omega| = O \left\{ (\log n/n)^{1/2} \vee p^{-1/2} \right\}.$$

From Theorems 1 and 2 we know that sparsity of a population matrix can be well estimated by thresholding the sample correlation matrix. In the next section, we propose an empirical estimator. We shall see from Theorem 3 that the empirical estimator is closely related to the thresholding estimator $\hat{\omega}(t)$.

3 An empirical estimator of sparsity under multivariate normal assumption

Under multivariate normal assumption, the density of r_{ij} given ρ_{ij} is (see for example Anderson 2003)

$$f_{r_{ij}|\rho_{ij}}(r|\rho) = \frac{2^{n-2} (1 - \rho^2)^{\frac{n}{2}} (1 - r^2)^{\frac{n-3}{2}}}{(n - 2)! \pi} \sum_{i=0}^{\infty} \frac{(2\rho r)^i}{i!} \Gamma^2\left(\frac{n+i}{2}\right), \quad \forall r \in (-1, 1).$$

Under Assumption 1, the marginal density of r_{ij} is:

$$\begin{aligned} f_{r_{ij}}(r; \omega) &= \omega f_{r_{ij}|\rho_{ij}}(r|\rho = 0) + (1 - \omega) \int_{-1}^1 f_{r_{ij}|\rho_{ij}}(r|\rho) g(\rho) d\rho \\ &= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right) \pi^{1/2}} (1 - r^2)^{\frac{n-3}{2}} \{1 + (1 - \omega) a_{ij}(r_{ij}, g)\}, \end{aligned}$$

where

$$\begin{aligned} a_{ij}(r_{ij}, g) &= -1 + \frac{\int_{-1}^1 f_{r_{ij}|\rho_{ij}}(r_{ij}|\rho) g(\rho) d\rho}{f_{r_{ij}|\rho_{ij}=0}(r_{ij}|\rho = 0)} \\ &= -1 + \sum_{k=0}^{\infty} \frac{\Gamma^2\left(\frac{n+2k}{2}\right) 2^{2k} r_{ij}^{2k}}{\Gamma^2\left(\frac{n}{2}\right) (2k)!} \int_{-1}^1 (1 - \rho^2)^{\frac{n}{2}} \rho^{2k} g(\rho) d\rho \\ &\quad + \sum_{k=0}^{\infty} \frac{\Gamma^2\left(\frac{n+2k+1}{2}\right) 2^{2k+1} r_{ij}^{2k+1}}{\Gamma^2\left(\frac{n}{2}\right) (2k+1)!} \int_{-1}^1 (1 - \rho^2)^{\frac{n}{2}} \rho^{2k+1} g(\rho) d\rho. \end{aligned} \tag{3}$$

Notice that when $a_{ij}(r_{ij}, g) > (<)0$, $f_{r_{ij}}(r; \omega)$ is maximized when $\omega = 0(1)$. In other words, when $a_{ij}(r_{ij}, g) > (<)0$, it tends to estimate ρ_{ij} as nonzero (zero). Therefore, we propose the following empirical estimator for ω :

$$\hat{\omega}_{em}(g) = \frac{\sum_{1 \leq i < j \leq p} \mathcal{I}_{\{a_{ij}(r_{ij}, g) < 0\}}}{p(p - 1)/2}. \tag{4}$$

Definition For any constant $d > 0$, let \mathcal{H}_d denote the set of probability density functions in $(-1, 1)$ such that for any $h \in \mathcal{H}_d$,

$$n^{-d} \leq \inf_{\rho \in (-1, 1)} h(\rho) \leq \sup_{\rho \in (-1, 1)} h(\rho) < \infty.$$

We shall see in Sect. 3.1 that for any $g \in \mathcal{H}_d$, $\hat{\omega}_{em}(g)$ is a consistent estimator of ω . In other words, asymptotically speaking, the functional $\hat{\omega}_{em}(\cdot)$ maps \mathcal{H}_d to a small neighborhood of ω .

3.1 Asymptotic properties of $\hat{\omega}_{em}(g)$ for any $g \in \mathcal{H}_d$

For any $g \in \mathcal{H}_d$, the empirical estimator $\hat{\omega}_{em}(g)$ is closely related to the thresholding approach. For example, suppose g is an odd function. It can be easily seen from (3) that $a_{ij}(r_{ij}, g)$ is a monotone function of r_{ij}^2 . Given g , let $t_0 > 0$ be the solution of

the equation $a_{ij}(t, g) = 0$, it is easy to see that $\hat{\omega}(t_0) = \hat{\omega}_{em}(g)$. For a general density function $g \in \mathcal{H}_d$, we have the following proposition:

Proposition 1 *Suppose Assumption 1 holds. When n is large enough we have for any $g \in \mathcal{H}_d$,*

$$a_{ij}(r_{ij}, g) > 0 \Rightarrow r_{ij}^2 > \frac{\log n - \log \log n}{n};$$

$$a_{ij}(r_{ij}, g) < 0 \Rightarrow r_{ij}^2 < \frac{2(d + 1) \log n}{n}.$$

From Proposition 1 and the definition of $a_{ij}(r_{ij}, g)$, we immediately have:

Theorem 3 *Suppose Assumption 1 holds. When n is large enough we have for any $g \in \mathcal{H}_d$,*

$$\hat{\omega}(\{(\log n - \log \log n)/n\}^{1/2}) \leq \hat{\omega}_{em}(g) \leq \hat{\omega}(\{2(d + 1) \log n/n\}^{1/2}).$$

From Theorems 1 and 3 we immediately have

Theorem 4 *Suppose Assumptions 1 and 2 hold. For any $g \in \mathcal{H}_d$, we have $\hat{\omega}_{em}(g) \rightarrow \omega$ in probability when $n \wedge p \rightarrow \infty$.*

Similarly, from Theorems 2 and 3 and the triangular inequality, we have

Theorem 5 *Suppose Assumptions 1 and 3 hold. For any $g \in \mathcal{H}_d$, when $n \wedge p \rightarrow \infty$, we have*

$$E|\hat{\omega}_{em}(g) - \omega| = O\left\{(\log n/n)^{1/2} \vee p^{-1/2}\right\}.$$

Under multivariate normal assumptions the upper bound obtained in Theorem 5 is better than the upper bound given in Theorem 1 of Jiang and Loh (2012), which is $O(n^{-1/4} \vee p^{-1/2})$.

3.2 An empirical version of $\hat{\omega}_{eb}(g)$

To compute $\hat{\omega}_{eb}(g)$, we need to determine the prior density $g(\rho)$. However, the theoretical results in Sect. 3.1 to some degree imply that the choice of the density function g is not crucial in the sense that for any prior density $g \in \mathcal{H}_d$, all the asymptotic properties given in Sect. 3.1 are true for the corresponding empirical estimator $\hat{\omega}_{em}(g)$. This is similar to the Bayes approach, where as long as the sample size is large enough, the choice of the prior (from a proper set of distribution families) is theoretically not very crucial. Here in this section we propose to compute $\hat{\omega}_{em}(g)$ based on a sequence of “crude” samples from g .

Notice that $a_{ij}(r_{ij}, g)$ can be written as

$$a_{ij}(r_{ij}, g) = -1 + \frac{E_g f_{r_{ij}|\rho_{ij}}(r_{ij}|\rho)}{f_{r_{ij}|\rho_{ij}=0}(r_{ij}|\rho = 0)}, \tag{5}$$

where E_g denotes expectation under g . Suppose g is bounded both from below and above, that is, there exists constants $D_1 > 0$ and $D_2 > 0$ such that $D_1 < \inf_{\rho \in (-1, 1)} g(\rho) \leq \sup_{\rho \in (-1, 1)} g(\rho) < D_2$. From Theorem 3 we have that when n is large enough, $\hat{\omega}(\{(\log n - \log \log n)/n\}^{1/2}) \leq \hat{\omega}_{\text{em}}(g) \leq \hat{\omega}(\{2 \log n/n\}^{1/2})$. Motivated by this we first of all threshold $|r_{ij}|, 1 \leq i < j \leq p$, by $(2 \log n/n)^{1/2}$ and obtain a sequence: $r_{ij} \mathcal{I}_{\{|r_{ij}| > (2 \log n/n)^{1/2}\}}, 1 \leq i < j \leq p$. We then use those nonzero $r_{ij} \mathcal{I}_{\{|r_{ij}| > (2 \log n/n)^{1/2}\}}$ as samples from g and use Monte Carlo approximation in computing the expectation in (5). In addition, we suggest using the following form of $f_{r_{ij}|\rho_{ij}}(r|\rho)$ since it converges more rapidly; see for example Anderson (2003):

$$f_{r_{ij}|\rho_{ij}}(r|\rho) = \frac{n-1}{\sqrt{2\pi}} (1-\rho^2)^{n/2} (1-r^2)^{(n-3)/2} (1-\rho r)^{-n+1/2} \sum_{j=0}^{\infty} \frac{\Gamma(1/2+j)\Gamma(1/2+j)\Gamma(n)(1+\rho r)^j}{\Gamma(1/2)\Gamma(1/2)\Gamma(n+1/2+j)\Gamma(j+1)2^j}.$$

From the second statement of Proposition 1, we know that when n is large enough, $|r_{ij}| > \sqrt{\frac{2(d+1)\log n}{n}}$ implies $a_{ij}(r_{ij}, g) > 0$. Therefore, to further reduce the computation complexity, we only need to compute $a_{ij}(r_{ij}, g)$ when $|r_{ij}|$ is small. This can further reduce the computation complexity. For example, when n is large, we can compute the $a_{ij}(r_{ij}, g)$ values for those $|r_{ij}| \leq (2 \log n/n)^{1/2}$ and simply set $\mathcal{I}_{\{a_{ij}(r_{ij}, g) < 0\}} = 0$ if $|r_{ij}| > (2 \log n/n)^{1/2}$.

4 A generalized empirical estimator

Suppose we randomly choose a subset \mathcal{S} from $\{\sigma_{ij} : 1 \leq i < j \leq p\}$ by simple random sampling without replacement. When the size of \mathcal{S} is large enough, we would expect that the proportion of zero in \mathcal{S} is close to the proportion of zero in the off-diagonal elements of Σ ; see for example Chap. 3 of Thompson (1997). Motivated by this, we propose a generalized version of the empirical estimator:

$$\hat{\omega}_{\text{em}}^m(\mathcal{S}_m, g) = 1/m \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{a_{ij}(r_{ij}, g) < 0\}}, \tag{6}$$

where \mathcal{S}_m is a random subset generated by simple random sampling (without replacement) from the index set $\{(i, j) : 1 \leq i < j \leq p\}$ such that the cardinality of \mathcal{S}_m equals m . When $m = p(p-1)/2$, $\hat{\omega}_{\text{em}}^m(\cdot)$ reduces to $\hat{\omega}_{\text{em}}(\cdot)$.

Similar to Theorems 4 and 5, we have:

Theorem 6 For any $g \in \mathcal{H}_d$, let $\hat{\omega}_{\text{em}}^m(\mathcal{S}_m, g)$ be defined as in (6). Under the assumptions of Theorem 4, we have $\hat{\omega}_{\text{em}}^m(\mathcal{S}_m, g) \rightarrow \omega$ in probability when $n \wedge m \rightarrow \infty$.

Theorem 7 For any $g \in \mathcal{H}_d$, let $\hat{\omega}_{\text{em}}^m(\mathcal{S}_m, g)$ be defined as in (6). Under the assumptions of Theorem 5, when $n \wedge m \rightarrow \infty$, we have

$$E|\hat{\omega}_{\text{em}}^m(\mathcal{S}_m, g) - \omega| = O\left\{(\log n/n)^{1/2} \vee (m \wedge p)^{-1/2}\right\}.$$

To compute $\hat{\omega}_{em}^m(\mathcal{S}_m, g)$, we first of all generate the index set \mathcal{S}_m from $\{(i, j) : 1 \leq i < j \leq p\}$ using simple random sampling, and then for any $(i, j) \in \mathcal{S}_m$ compute $a_{ij}(r_{ij}, g)$ as in Sect. 3.2.

5 Simulation study

This section provides some simulation results for estimating the sparsity of the population covariance matrix. More simulations can be found in the author’s PhD thesis. In this simulation study, we consider the following three types of sparse covariance matrices.

Model 5.1 Following [Jiang and Loh \(2012\)](#), let $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq p}$, where $\sigma_{ii} = 1, 1 \leq i \leq p; \sigma_{ij} = 0.3$ if $1 \leq i, j \leq p/2, i \neq j$ and $\sigma_{ij} = 0$ otherwise.

Model 5.2 Following [Cai and Liu \(2011\)](#), let $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq p}$, where $\sigma_{ij} = (1 - |i - j|/20)_+$ if $1 \leq i, j \leq p/2, \sigma_{ii} = 4$ if $p/2 + 1 \leq i \leq p$, and $\sigma_{ij} = 0$ otherwise.

Model 5.3 Let $L = (l_{ij})_{1 \leq i, j \leq p}$ be a lower triangular matrix and the elements in L are generated as follows: $l_{ii} = 1, i = 1, \dots, p, l_{i1} = 2\sqrt{U(0, 1)} \times B(0, 0.5), i = 2, \dots, p$ and $l_{ij} = U(0, 1) \times B(1, 0.01), 2 \leq j < i \leq p$. Here $U(0, 1)$ denotes a random variable uniformly distributed on $(0, 1)$ and $B(1, \alpha)$ is a Bernoulli random variable which equals 1 with probability α and 0 with probability $1 - \alpha$. Let $C = (c_{ij})_{1 \leq i, j \leq p}$ be a lower triangular matrix such that $c_{ij} = l_{ij} / (\sum_{k=1}^i l_{ik}^2)^{1/2}, 1 \leq j \leq i \leq p$. We then set $\Sigma = MCC^T M^T$, where M is a random permutation matrix uniformly distributed in set of all $p \times p$ permutation matrices.

In the first simulation, we set $n = 100$ and $p = 50, 100, 200$. X_1, \dots, X_n are generated independently from $N_p(0, \Sigma)$. We compare the empirical estimator $\hat{\omega}_{em}(g)$ to the following three estimators:

- (i) $\hat{\omega}_{mm}$, estimator based on moment matrices; see [Jiang and Loh \(2012\)](#).
- (ii) $\hat{\omega}_{cv}$, sparsity of the hard thresholding estimator derived using cross validation. Following [Bickel and Levina \(2008\)](#), the threshold is chosen in the following way: randomly split the n sample into two sets of size $n_1 = n - \lfloor n / \log n \rfloor$ and $n_2 = \lfloor n / \log n \rfloor$ and repeat this N times. Here $\lfloor \cdot \rfloor$ is the greatest integer function. For the k th split, let $S_{1,k}, S_{2,k}$ be the sample covariance matrix based on the n_1 and n_2 observations, respectively. For a given threshold t define the thresholding operator by $T_t(S) = [s_{ij} \mathcal{I}(|s_{ij}| > t)]_{p \times p}$. We then choose t such that

$$CV(t) = \frac{1}{N} \sum_{k=1}^N \|T_t(S_{1,k}) - S_{2,k}\|_F^2,$$

is minimized. Here $\|S\|_F^2 = \sum_{1 \leq i, j \leq p} s_{ij}^2$ is the Frobenius norm of a matrix $S = (s_{ij})_{p \times p}$. In this simulation, we set $N = 100$ and the set of thresholds over which an optimal threshold was searched is $\{0.02, 0.04, \dots, 0.38, 0.40\}$.

- (iii) $\hat{\omega}_{cv}^s$, sparsity of the thresholding estimator derived using soft thresholding function and cross validation as in Rothman et al. (2009). $\hat{\omega}_{cv}^s$ is computed similar to $\hat{\omega}_{cv}$ except that the thresholding operator is now given as $T_t(S) = [\text{sign}(s_{ij})(|s_{ij}| - t)\mathcal{I}(|s_{ij}| > t)]_{p \times p}$.
- (iv) $\hat{\omega}_{acv}$, sparsity of the adaptive thresholding estimator derived using hard thresholding function and cross validation. Denote the i th observation as $X_i = (X_{i1}, \dots, X_{ip})^T$. Following Cai and Liu (2011), for a thresholding parameter t , the adaptive thresholding estimator is defined as $T_t^a(S) = [T_{ij}^a(s_{ij}, t)]_{1 \leq i, j \leq p}$, where $T_{ij}^a(s_{ij}, t) = s_{ij}\mathcal{I}\left(|s_{ij}| > t\sqrt{\hat{\theta}_{ij} \log p/n}\right)$, and $\hat{\theta}_{ij}$ is defined as:

$$\hat{\theta}_{ij} = \frac{1}{n} \sum_{k=1}^n \left[(X_{ki} - \bar{X}^i) (X_{kj} - \bar{X}^j) - s_{ij} \right]^2, \quad \bar{X}^i = \frac{1}{n} \sum_{k=1}^n X_{ki}.$$

Similar to the cross validation procedure for computing $\hat{\omega}_{cv}$, we randomly split the n observations for N times and choose t such that

$$ACV(t) = \frac{1}{N} \sum_{k=1}^N \|T_t^a(S_{1,k}) - S_{2,k}\|_F^2,$$

is minimized. In this simulation, we set $N = 100$ and the set of thresholds over which an optimal threshold was searched is $\{0.1, 0.2, \dots, 3.9, 4\}$.

For each case, the simulation is repeated 100 times. The mean and its standard deviation (SD) of the following quantities are computed over 100 replications: (i) L_1 -loss: $|\text{estimator} - \omega|$; (ii) Error1 = $\#\{(i, j) : 1 \leq i < j \leq p, \rho_{ij} = 0, \hat{r}_{ij} \neq 0\}$; (iii) Error2 = $\#\{(i, j) : 1 \leq i < j \leq p, \rho_{ij} \neq 0, \hat{r}_{ij} = 0\}$. Here $\hat{R} = (\hat{r}_{ij})_{p \times p}$ is obtained in the following ways: for $\hat{\omega}_{mm}$, let t be the $[\hat{\omega}_{mm}p(p-1)/2]$ th smallest number among $|r_{ij}|, 1 \leq i < j \leq p$. Here $[\cdot]$ is the greatest integer function. We construct $\hat{R} = (\hat{r}_{ij})_{p \times p}$ with $\hat{r}_{ji} = \hat{r}_{ij} = r_{ij}\mathcal{I}(|r_{ij}| > t), 1 \leq i < j \leq p$ and $\hat{r}_{ii} = 1, i = 1, \dots, p$, i.e., \hat{R} is obtained by applying the universal thresholding approach to the sample correlation matrix R such that the resulting sparsity of \hat{R} equals $\hat{\omega}_{mm}$; for $\hat{\omega}_{cv}$ and $\hat{\omega}_{acv}$ we use the correlation matrices corresponding to the thresholding estimator and the adaptive thresholding estimator as \hat{R} , respectively; For $\hat{\omega}_{eb}(g)$, we set $\hat{r}_{ji} = \hat{r}_{ij} = r_{ij}\mathcal{I}(|a_{ij}| > 0), 1 \leq i < j \leq p$.

Tables 1, 2 and 3 indicate that under Models 5.1, 5.2 and 5.3, $\hat{\omega}_{em}(g)$ has smaller mean L_1 -loss and Error1+Error2 values than the other four estimators. From Tables 1, 2 and 3 and all other simulations we have done, we found that generally both $\hat{\omega}_{mm}$ and $\hat{\omega}_{em}(g)$ can estimate ω well while $\hat{\omega}_{cv}$ and $\hat{\omega}_{acv}$ can estimate ω well only when most of the nonzero off-diagonal elements of Γ are away from zero. $\hat{\omega}_{cv}^s$ is not doing well in terms of covariance selection and sparsity estimation under the models used in the simulation study. Similar results can be observed from Table 3 of Rothman et al. (2009), which shows that soft thresholding method can have much larger false-positive rates than hard thresholding method. When not too many of the nonzero off-diagonal elements of Γ are close to zero, $\hat{\omega}_{em}(g)$ generally would outperform $\hat{\omega}_{mm}$ while

Table 1 Simulation results under Model 5.1 over 100 replications

	$\hat{\omega}_{em}(g)$	$\hat{\omega}_{mm}$	$\hat{\omega}_{cv}$	$\hat{\omega}_{cv}^s$	$\hat{\omega}_{acv}$
$p = 50; \omega = 0.755$					
L_1 -loss (SD)	0.011 (0.001)	0.025 (0.002)	0.045 (0.004)	0.237 (0.010)	0.052 (0.005)
Error1 (SD)	41.3 (0.9)	49.0 (2.9)	62.1 (4.7)	299.7 (11.9)	67.1 (4.7)
Error2 (SD)	40.7 (2.1)	39.8 (2.1)	45.5 (4.3)	9.4 (1.0)	43.6 (5.3)
$p = 100; \omega = 0.753$					
L_1 -loss (SD)	0.011 (0.001)	0.013 (0.001)	0.037 (0.004)	0.216 (0.008)	0.058 (0.008)
Error1 (SD)	122.3 (2.3)	176.1 (8.0)	218.4 (13.5)	1,107.9 (38.1)	333.7 (34.4)
Error2 (SD)	122.9 (8.3)	155.8 (7.1)	190.5 (19.9)	36.8 (4.3)	147.7 (19.0)
$p = 200; \omega = 0.751$					
L_1 -loss (SD)	0.011 (0.001)	0.011 (0.008)	0.037 (0.004)	0.233 (0.008)	0.059 (0.006)
Error1 (SD)	562.8 (7.0)	692.1 (25.0)	834.4 (48.0)	4,777.2 (138.9)	1,134.8 (71.5)
Error2 (SD)	496.0 (29.7)	634.3 (27.9)	846.1 (76.4)	138.9 (21.0)	817.7 (114.5)

Table 2 Simulation results under Model 5.2 over 100 replications

	$\hat{\omega}_{em}(g)$	$\hat{\omega}_{mm}$	$\hat{\omega}_{cv}$	$\hat{\omega}_{cv}^s$	$\hat{\omega}_{acv}$
$p = 50; \omega = 0.767$					
L_1 -loss (SD)	0.018 (0.001)	0.034 (0.003)	0.027 (0.002)	0.179 (0.008)	0.051 (0.002)
Error1 (SD)	12.9 (0.8)	34.9 (3.7)	92.1 (2.7)	261.1 (8.9)	1.3 (0.4)
Error2 (SD)	32.3 (1.3)	35.4 (2.3)	87.7 (2.4)	42.2 (2.2)	63.7 (2.1)
$p = 100; \omega = 0.846$					
L_1 -loss (SD)	0.014 (0.001)	0.032 (0.002)	0.015 (0.001)	0.112 (0.004)	0.051 (0.001)
Error1 (SD)	70.0 (2.6)	141.6 (14.8)	333.9 (5.4)	770.6 (29.0)	2.2 (0.5)
Error2 (SD)	135.2 (3.3)	143.1 (6.5)	326.9 (5.6)	217.2 (5.4)	253.0 (4.8)
$p = 200; \omega = 0.914$					
L_1 -loss (SD)	0.004 (<0.001)	0.021 (0.002)	0.023 (0.001)	0.092 (0.002)	0.032 (0.001)
Error1 (SD)	299.6 (7.8)	361.6 (404.5)	1,257.0 (11.1)	2,404.7 (39.3)	4.5 (0.7)
Error2 (SD)	339.4 (5.8)	414.5 (184.6)	795.6 (7.2)	577.7 (9.1)	649.4 (11.1)

$\hat{\omega}_{mm}$ would perform better when most of the nonzero off-diagonal elements of Γ are close to zero. On the other hand, $\hat{\omega}_{em}(g)$ generally has smaller Error1+Error2 values than $\hat{\omega}_{mm}$, indicating the covariance selection procedure based on $\hat{R} = (\hat{r}_{ij})_{p \times p}$ with $\hat{r}_{ji} = r_{ij}\mathcal{I}(|a_{ij}| > 0)$ can do better than simply thresholding R based on $\hat{\omega}_{mm}$.

In the second simulation, we study the performance of the generalized empirical estimator $\hat{\omega}_{em}^m(\mathcal{S}_m, g)$. Similar to the previous simulation, we let $n = 100$ and generate n independent samples from $N_p(0, \Sigma)$ with Σ given by Models 5.1, 5.2 and 5.3. We let $m = 5,000$ and $p = 200, 600, 1,000$. The mean and its standard deviation over 100 replications of the L_1 -loss are reported in Table 4.

Table 3 Simulation results under Model 5.3 over 100 replications

	$\hat{\omega}_{em}(g)$	$\hat{\omega}_{mm}$	$\hat{\omega}_{cv}$	$\hat{\omega}_{cv}^s$	$\hat{\omega}_{acv}$
$p = 50; \omega = 0.639$					
L_1 -loss (SD)	0.037 (0.001)	0.037 (0.028)	0.249 (0.014)	0.261 (0.012)	0.067 (0.001)
Error1(SD)	11.1 (0.7)	44.5 (4.1)	318.9 (16.4)	331.6 (13.7)	47.2 (11.0)
Error2 (SD)	57.8 (1.7)	47.0 (2.1)	13.8 (0.1)	11.9 (1.0)	62.2 (3.1)
$p = 100; \omega = 0.743$					
L_1 -loss (SD)	0.024 (0.001)	0.024 (0.002)	0.244 (0.012)	0.231 (0.012)	0.036 (0.002)
Error1 (SD)	51.0 (1.7)	142.7 (11.9)	1,273.2 (60.7)	1,210.9 (58.8)	50.5 (7.9)
Error2 (SD)	174.2 (2.9)	160.2 (4.7)	65.6 (3.5)	68.5 (3.8)	209.9 (4.8)
$p = 200; \omega = 0.745$					
L_1 -loss (SD)	0.017 (0.001)	0.018 (0.001)	0.247 (0.012)	0.229 (0.011)	0.027 (0.001)
Error1 (SD)	225.8 (5.6)	425.2 (31.6)	5,121.7 (230.8)	4,757.5 (219.3)	206.0 (19.0)
Error2 (SD)	575.4 (9.9)	559.4 (13.1)	196.9 (9.1)	208.4 (10.8)	713.6 (18.2)

Table 4 Simulation results for the generalized empirical estimator under Models 5.1 and 5.2

Model 5.1	p	200	600	1,000
	L_1 -loss (SD)	0.011 (0.001)	0.010 (0.001)	0.010 (0.001)
Model 5.2	p	200	600	1,000
	L_1 -loss (SD)	0.006 (<0.001)	0.009 (<0.001)	0.015 (<0.001)
Model 5.3	p	200	600	1,000
	L_1 -loss (SD)	0.049 (0.001)	0.071 (0.001)	0.096 (0.001)

Simulation results in Table 4 indicate that $\hat{\omega}_{em}^m(S_m, g)$ can still estimate ω very well while the computation is largely reduced. For example, when $p = 200$ and $m = 5,000$, the time used for calculating $\hat{\omega}_{em}^m(S_m, g)$ is about 1/4 of the time used for calculating $\hat{\omega}_{em}(g)$, while the L_1 -loss and mean error values of $\hat{\omega}_{em}^m(S_m, g)$ are still very small.

6 Appendix: Technical details

The following lemma is a special case of Proposition 1 in Jiang (2013). It gives a Bernstein-type inequality for elements of the sample correlation matrix.

Lemma 1 For any $0 < v \leq 2$ and $1 \leq j, k \leq p$, there exist constants $d_1 > 0$ and $d_2 > 0$ such that

$$P(|r_{jk} - \rho_{jk}| \geq v|\rho_{jk}) \leq d_1 e^{-d_2 n v^2}.$$

Next we provide the proof for Theorem 2. Theorem 1 can be proved similarly.

Proof of Theorem 2 Notice that

$$E|\hat{\omega}(t) - \omega| \leq E\left|\frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}} - \omega\right| + E\left|\hat{\omega}(t) - \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}}\right|,$$

we shall bound $E|\hat{\omega}(t) - \omega|$ by bounding the two terms on the right hand side of the above inequality.

Under Assumption 3, By Jensen’s inequality we have

$$\begin{aligned} & E\left|\frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}} - \omega\right| \\ & \leq \left\{E\left|\frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} (\mathcal{I}_{\{\rho_{ij}=0\}} - E\mathcal{I}_{\{\rho_{ij}=0\}})\right|^2\right\}^{1/2} \\ & \leq \left[\frac{4\omega(1-\omega)}{p^2(p-1)^2} \left\{\frac{p(p-1)}{2} + p(p-1)(p-2)\right\} + O(p^{-1})\right]^{1/2} \\ & = O(p^{-1/2}), \end{aligned} \tag{7}$$

where in the last step we have used Assumption 3 and the fact that

$$\text{Var}(\mathcal{I}_{\{\rho_{12}=0\}}) = \omega(1-\omega),$$

and

$$\begin{aligned} & E(\mathcal{I}_{\{\rho_{12}=0\}} - E\mathcal{I}_{\{\rho_{12}=0\}})(\mathcal{I}_{\{\rho_{23}=0\}} - E\mathcal{I}_{\{\rho_{23}=0\}}) \\ & \leq \{\text{Var}(\mathcal{I}_{\{\rho_{12}=0\}}) \text{Var}(\mathcal{I}_{\{\rho_{23}=0\}})\}^{1/2} \\ & = \omega(1-\omega). \end{aligned}$$

On the other hand,

$$\begin{aligned} & E\left|\hat{\omega}(t) - \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}}\right| \\ & \leq \frac{2}{p(p-1)} E \sum_{1 \leq i < j \leq p} \left|\mathcal{I}_{\{|r_{ij}|<t\}} - \mathcal{I}_{\{\rho_{ij}=0\}}\right| \\ & = E|Y_{12}|, \end{aligned} \tag{8}$$

where

$$Y_{ij} = \mathcal{I}_{\{|r_{ij}|<t\}} - \mathcal{I}_{\{\rho_{ij}=0\}}, \quad \forall 1 \leq i < j \leq p.$$

Since $\{(\log n - \log \log n)/n\}^{1/2} \leq t < C_1(\log n/n)^{1/2}$, for any constant $C_2 > 0$ we have,

$$\begin{aligned}
 E|Y_{ij}| &= P(|r_{ij}| \geq t, \rho_{ij} = 0) + P(|r_{ij}| < t, \rho_{ij} \neq 0) \\
 &\leq \omega P \left\{ |r_{ij}| \geq \{(\log n - \log \log n)/n\}^{1/2} | \rho_{ij} = 0 \right\} \\
 &\quad + P \left\{ 0 < |\rho_{ij}| \leq C_2(\log n/n)^{1/2} \right\} \\
 &\quad + P \left\{ |r_{ij}| < C_1(\log n/n)^{1/2}, |\rho_{ij}| > C_2(\log n/n)^{1/2} \right\}. \tag{9}
 \end{aligned}$$

By choosing $C_2 > (2d_2)^{-1/2} + C_1$, from the assumption $\sup_{\rho \in (-1,1)} g(\rho) < \infty$ and Lemma 1 we have:

$$P \left\{ 0 < |\rho_{ij}| \leq C_2(\log n/n)^{1/2} \right\} = O \left\{ (\log n/n)^{1/2} \right\}, \tag{10}$$

and

$$P \left\{ |r_{ij}| < (C_1 \log n/n)^{1/2}, |\rho_{ij}| > C_2(\log n/n)^{1/2} \right\} = o \left(n^{1/2} \right). \tag{11}$$

Let $C_3 > (2d_2)^{-1}$ be a constant. Using the density function of r_{ij} given $\rho_{ij} = 0$ and Lemma 1, we have,

$$\begin{aligned}
 &P \left\{ |r_{ij}| \geq \{(\log n - \log \log n)/n\}^{1/2} | \rho_{ij} = 0 \right\} \\
 &= P \left[\{(\log n - \log \log n)/n\}^{1/2} \leq |r_{ij}| \leq (C_3 \log n/n)^{1/2} | \rho_{ij} = 0 \right] \\
 &\quad + P \left\{ |r_{ij}| \geq (C_3 \log n/n)^{1/2} | \rho_{ij} = 0 \right\} \\
 &\leq \int_{\{(\log n - \log \log n)/n\}^{1/2}}^{(C_3 \log n/n)^{1/2}} \frac{\Gamma(n/2)r(1-r^2)^{(n-3)/2}}{\{(\log n - \log \log n)/n\}^{1/2} \Gamma(n/2 - 1/2)(\pi/2)^{1/2}} dr \\
 &\quad + d_1 e^{-d_2 C_3 \log n} \\
 &\leq \frac{\Gamma(n/2)\{1 - (\log n - \log \log n)/n\}^{(n-1)/2}}{\{(\log n - \log \log n)/n\}^{1/2} \Gamma(n/2 - 1/2)(n-1)(\pi/2)^{1/2}} + d_1 e^{-d_2 C_3 \log n} \\
 &= O(n^{-1/2}). \tag{12}
 \end{aligned}$$

Combining (8), (9), (10), (11) and (12) we have

$$E \left| \hat{\omega}(t) - \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}} \right| = O \left\{ (\log n/n)^{1/2} \right\}.$$

Together with (7) we conclude that

$$E|\hat{\omega}(t) - \omega| = O\{(\log n/n)^{1/2} \vee p^{-1/2}\}. \quad \square$$

Theorems 3, 4 and 5 can be easily proved using Proposition 1, Theorems 1 and 2. Hence we only provide the proof of Proposition 1. The following lemma is a direct result of (4.2) in [Bustoz and Ismail \(1986\)](#) and will be used frequently in the proof of Proposition 1.

Lemma 2 For any $n, k \in \mathbb{Z}^+$, we have

$$\frac{1}{\sqrt{n/2+k}} < \frac{\Gamma(n/2+k)}{\Gamma(n/2+k+1/2)} < \frac{1}{\sqrt{n/2+k-1/4}}. \tag{13}$$

Proof of Proposition 1 We prove the first statement of Proposition 1. The proof for the second statement is similar and can be found in Lemma 3.6 in the author’s PhD thesis. To prove the first statement, it suffices to show that for all n large enough, $r^2 \leq (\log n - \log \log n)/n$ implies $a_{ij}(r, g) < 0$. Denote $D = \sup_{-1 < \rho < 1} g(\rho)$. First of all decompose $a_{ij}(r, g)$ to be:

$$\begin{aligned} a_{ij}(r, g) &= \int_0^1 (1 - \rho^2)^{\frac{n}{2}} g(\rho) d\rho \\ &+ \sum_{i=1}^{\lfloor 2 \log n \rfloor} \frac{\Gamma^2(\frac{n+2i}{2}) 2^{2i} r^{2i}}{\Gamma^2(\frac{n}{2}) (2i)!} \int_0^1 (1 - \rho^2)^{\frac{n}{2}} \rho^{2i} g(\rho) d\rho \\ &+ \sum_{i=\lfloor 2 \log n \rfloor+1}^{\infty} \frac{\Gamma^2(\frac{n+2i}{2}) 2^{2i} r^{2i}}{\Gamma^2(\frac{n}{2}) (2i)!} \int_0^1 (1 - \rho^2)^{\frac{n}{2}} \rho^{2i} g(\rho) d\rho \\ &+ \sum_{i=0}^{\lfloor 2 \log n \rfloor} \frac{\Gamma^2(\frac{n+2i+1}{2}) 2^{2i+1} r^{2i+1}}{\Gamma^2(\frac{n}{2}) (2i+1)!} \int_0^1 (1 - \rho^2)^{\frac{n}{2}} \rho^{2i+1} g(\rho) d\rho \\ &+ \sum_{i=\lfloor 2 \log n \rfloor+1}^{\infty} \frac{\Gamma^2(\frac{n+2i+1}{2}) 2^{2i+1} r^{2i+1}}{\Gamma^2(\frac{n}{2}) (2i+1)!} \int_0^1 (1 - \rho^2)^{\frac{n}{2}} \rho^{2i+1} g(\rho) d\rho - 1 \\ &:= I + II + III + II' + III' - 1. \end{aligned}$$

From Lemma 2 we have

$$I \leq D \frac{\Gamma(n/2+1)\Gamma(1/2)}{\Gamma(n/2+3/2)} \leq \frac{D\sqrt{\pi}}{\sqrt{n/2+3/4}};$$

For II , when $r^2 \leq (\log n - \log \log n)/n$, we have

$$\begin{aligned} II &\leq 2D \sum_{i=1}^{\lfloor 2 \log n \rfloor} \frac{\Gamma^2(\frac{n+2i}{2}) 2^{2i} r^{2i}}{\Gamma^2(\frac{n}{2}) (2i)!} \cdot \frac{\Gamma(\frac{n}{2}+1)\Gamma(i+\frac{1}{2})}{\Gamma(\frac{n}{2}+i+\frac{3}{2})} \\ &= 2D \sum_{i=1}^{\lfloor 2 \log n \rfloor} \frac{(2r)^{2i} \sqrt{\pi} 2^{-2i}}{\Gamma(i+1)} \cdot \left(\frac{n}{2}+i-1\right) \cdots \frac{n}{2} \cdot \frac{\Gamma(\frac{n}{2}+i)}{\Gamma(\frac{n}{2}+i+\frac{1}{2})} \cdot \frac{\frac{n}{2}}{\frac{n}{2}+i+\frac{1}{2}} \\ &\leq \frac{2\sqrt{2\pi}D}{\sqrt{n-1}} \sum_{i=1}^{\lfloor 2 \log n \rfloor} \frac{r^{2i}}{i!} \left(\frac{n}{2}+2\log n\right)^i \\ &\leq \frac{2\sqrt{2\pi}D}{\sqrt{n-1}} e^{r^2(n/2+2\log n)} \\ &= O(\log^{-1/2} n). \end{aligned}$$

Similarly it can be shown that $II' = O(\log^{-1/2} n)$. For III we have

$$\begin{aligned}
 III &\leq 2D \sum_{i>[2\log n]} \frac{\Gamma^2\left(\frac{n+2i}{2}\right) 2^{2i} r^{2i}}{\Gamma^2\left(\frac{n}{2}\right) (2i)!} \cdot \frac{\Gamma\left(\frac{n}{2} + 1\right) \Gamma\left(i + \frac{1}{2}\right)}{\Gamma\left(\frac{n}{2} + i + \frac{3}{2}\right)} \\
 &= 2D \sum_{i>[2\log n]} \frac{(2r)^{2i} \cdot \frac{n}{2}}{\frac{n}{2} + i + \frac{1}{2}} \cdot \frac{\Gamma\left(\frac{n}{2} + i\right)}{\Gamma\left(\frac{n}{2} + i + \frac{1}{2}\right)} \cdot \frac{\Gamma\left(i + \frac{1}{2}\right)}{\Gamma(i + 1)} \cdot \frac{\left(\frac{n}{2} + i - 1\right) \cdots \frac{n}{2}}{2i \cdots (i + 1)} \\
 &\leq 2D \sum_{i>[2\log n]} (2r)^{2i} \cdot \frac{1}{\sqrt{\frac{n}{2} + i - \frac{1}{4}}} \cdot \frac{1}{\sqrt{i + \frac{1}{4}}} \cdot \left(\frac{n}{4\log n}\right)^i \\
 &\leq \frac{2D}{\sqrt{n \log n}} \sum_{i>[2\log n]} \left(1 - \frac{\log \log n}{\log n}\right)^i \\
 &= O\left(n^{-1/2} \log^{-1/2} n\right).
 \end{aligned}$$

Similarly we have $III' = O(n^{-1/2} \log^{-1/2} n)$. Therefore, for any $r^2 \leq (\log n - \log \log n)/n$, $a_{ij}(r, g)$ converge to -1 when $n \rightarrow \infty$. □

Proof of Theorem 7 Notice that

$$\begin{aligned}
 E|\hat{\omega}_{\text{eb}}^m(\mathcal{S}_m) - \omega| &\leq E\left|\frac{1}{m} \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{a_{ij}(r_{ij},g)<0\}} - \frac{1}{m} \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{\rho_{ij}=0\}}\right| \\
 &\quad + E\left|\frac{1}{m} \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{\rho_{ij}=0\}} - \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}}\right| \\
 &\quad + E\left|\frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}} - \omega\right|. \tag{14}
 \end{aligned}$$

Similar to the bound we obtained for (8) in the proof of Theorem 2, by conditioning on \mathcal{S}_m first we have

$$E\left|\frac{1}{m} \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{a_{ij}(r_{ij},g)<0\}} - \frac{1}{m} \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{\rho_{ij}=0\}}\right| = O\left\{(\log n/n)^{1/2}\right\}. \tag{15}$$

Let E_Γ denote the conditional expectation given Γ . From (3.2) of Thompson (1997) we have

$$\begin{aligned}
 &E\left|\frac{1}{m} \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{\rho_{ij}=0\}} - \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}}\right| \\
 &= E\left\{E_\Gamma\left|\frac{1}{m} \sum_{(i,j) \in \mathcal{S}_m} \mathcal{I}_{\{\rho_{ij}=0\}} - \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}}\right|\right\}
 \end{aligned}$$

$$\begin{aligned} &\leq E \left[\frac{p(p-1) - 2m}{m\{p(p-1) - 2\}} \frac{2 \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}}}{p(p-1)} \left\{ 1 - \frac{2 \sum_{1 \leq i < j \leq p} \mathcal{I}_{\{\rho_{ij}=0\}}}{p(p-1)} \right\} \right]^{1/2} \\ &\leq \left[\frac{p(p-1) - 2m}{4m\{p(p-1) - 2\}} \right]^{1/2}. \end{aligned} \tag{16}$$

Combining (14), (7), (15) and (16) we have

$$\begin{aligned} E|\hat{\omega}_{\text{eb}}^m(\mathcal{S}_m) - \omega| &= O \left\{ (\log n/n)^{1/2} + p^{-1/2} + \left[\frac{p(p-1) - 2m}{4m\{p(p-1) - 2\}} \right]^{1/2} \right\} \\ &= O\{(\log n/n)^{1/2} \vee (m \wedge p)^{-1/2}\}. \end{aligned} \quad \square$$

Acknowledgments The author thanks Professor Wei-Liem Loh for valuable comments on the author’s PhD thesis and this paper. The author would also like to thank Professor Kenji Fukumizu, an Associate Editor and two referees for their valuable comments and suggestions.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley Series in Probability and Statistics, Wiley.

Bickel, P.J., Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, 36, 2577–2604.

Bustoz, J., Ismail, M. (1986). On gamma function inequalities. *Mathematics of Computation*, 47, 659–667.

Cai, T., Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106, 672–684.

El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, 48, 2717–2756.

Jiang, B. (2013). Covariance selection by thresholding the sample correlation matrix. *Statistics and Probability Letters*, 83, 2492–2498.

Jiang, B., Loh, W.L. (2012). On the sparsity of signals in a random sample. *Biometrika*, 99, 915–928.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99, 733–740.

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

Thompson, M. E. (1997). *Theory of sample surveys* (1st ed.). London: Chapman & Hall.