Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

11-2013

Covariance Selection by Thresholding the Sample Correlation Matrix

Binyan JIANG
Singapore Management University LARC, stats.jby@gmail.com

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

Part of the Databases and Information Systems Commons, and the Numerical Analysis and Scientific Computing Commons

Citation

JIANG, Binyan. Covariance Selection by Thresholding the Sample Correlation Matrix. (2013). *Statistics and Probability Letters*. 83, (11), 2492-2498.
Available at: https://ink.library.smu.edu.sg/sis_research/2049

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

The final version is published in Statistics and Probability Letters, 2013, 83 (11), 2492-2498.
doi:10.1016/j.spl.2013.07.008 on ScienceDirect.com

Covariance selection by thresholding the sample correlation matrix

Binyan Jiang

The Living Analytics Research Centre (LARC), Singapore Management University and Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

This article shows that when the nonzero coefficients of the population correlation matrix are all greater in absolute value than $(C_1 \log p/n)^{1/2}$ for some constant C_1, we can obtain covariance selection consistency by thresholding the sample correlation matrix. Furthermore, the rate $(\log p/n)^{1/2}$ is shown to be optimal.

Keywords: Bernstein type inequality, Covariance selection, Large correlation matrix, Large covariance matrix, Thresholding

1. Introduction

In the last decade, there has been a surge in interest in the estimation of large covariance matrices under sparsity assumptions. The number of variables p may be larger than the sample size n and the population covariance matrix is usually assumed to be sparse in that a number of the off-diagonal elements are zero.

Suppose X is a p-dimensional multivariate random vector with unknown mean $\mu = (\mu_1, \ldots, \mu_p)^T$ and covariance matrix $\Sigma_{p\times p} = (\sigma_{ij})_{p\times p}$ and let $X_1, ..., X_n$ be independent, identically distributed random observations of X. Write $\bar{X} = 1/n \sum_{i=1}^n X_i$. The sample covariance matrix is:

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

Email address: stats.jby@gmail.com (Binyan Jiang)

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

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

where $T_{ij}(\cdot)$ is a general thresholding function. For example, Bickel and Levina (2008) considered hard-thresholding $T_{ij}(s) = s\mathcal{I}(|s| > t), 1 \le i < j \le p$, 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 thresholding sample covariance matrices with more general thresholding functions possessing shrinkage properties. However, similarly to Bickel and Levina (2008), they used a thresholding parameter to threshold all the off-diagonal elements of S. Hence in this paper we shall call thresholding methods of Bickel and Levina (2008) and Rothman et al. (2009) the universal-threshold approach. Intuitively, when the diagonal elements of the population matrix are not all equal to each other, the universal-threshold approach may not be appropriate. Recently, Cai and Liu (2011) proposed an adaptive thresholding method which is applicable when the p elements in X are not homoscedastic. They used different thresholding parameters for different s_{ij} depending on the variance of s_{ij}. Similar to the universal-threshold approach, the sparsity of the resulting estimator is determined by a thresholding parameter. Other than sample covariance matrices, thresholding approaches were also applied to sample correlation matrices. El Karoui (2008) applied hard-thresholding to the off-diagonal elements of the sample correlation matrix and showed consistency under a special notion of sparsity. Cai and Liu (2011) also provided some discussion on the thresholded sample correlation matrix.

In this paper, we focus on answering a relatively simple question: when is it possible to obtain covariance selection consistency by applying the universal-threshold approach to the sample correlation matrix? Here the universal-threshold approach is applied to the sample correlation matrix instead of the sample covariance matrix so that it is adaptive to the case that p elements in X are not homoscedastic. Suppose the nonzero coefficients of the population correlation matrix are all greater in absolute value than $k(n, p)$. Here $k(n, p)$ is the separation gap between zero and those nonzero elements of the population correlation matrix. Clearly, when $k(n, p)$ is large enough, the universal-threshold approach might be able to obtain covariance selection consistency for a proper thresholding parameter. However, it is not clear how large the separation $k(n, p)$ should be. To obtain covariance selection consistency, it is shown in Section 2 that the rate of the minimal separation is exactly $(\log p/n)^{1/2}$. This result is extended in Section 3 to the case that $X_1, \ldots, X_n$ are weakly dependent.

2. Covariance selection by thresholding the sample correlation matrix

Let $X_1, ..., X_n$ be independent, identically distributed p-dimensional random vectors with unknown mean $\mu = (\mu_1, \ldots, \mu_p)^T$ and covariance matrix $\Sigma_{p\times p} = (\sigma_{ij})_{p\times p}$. The sample covariance matrix is given by (1) and the sample correlation matrix is denoted as $R = (r_{ij})_{1\le i,j\le p}$ where $r_{ij} = s_{ij}/(s_{ii}s_{jj})^{1/2}, 1 \le i, j \le p$. Denote the population correlation matrix by $\Gamma = (\rho_{ij})_{p\times p}$ and define:

$$G = \{(i,j) : \rho_{ij} \neq 0, 1 \le i < j \le p\}. \tag{2}$$

For a thresholding parameter t, define:

$$\hat{G}(t) = \{(i,j) : |r_{ij}| \ge t, 1 \le i < j \le p\}. \tag{3}$$

As mentioned in Section 1, to obtain covariance selection consistency, the rate of $k(n,p)$ should be $(\log p/n)^{1/2}$ or larger. More specifically, we shall show that:

(i)When $\rho_{ij}^2 > C_1 \log p/n$ for all $(i,j) \in G$ and some constant $C_1 > 0$ large enough, a threshold t exists such that $\mathrm{pr}\{\hat{G} = G\} \to 1$;

(ii)An example exists in that $\rho_{ij}^2 = C_2 \log p/n$ for some $(i,j) \in G$ and when constant $C_2 > 0$ is small enough, $\mathrm{pr}\{\hat{G} = G\} \to 0$ for any t.

The following conditions are needed in the main results of this section:

Condition 1. Denote $X_j = (X_{1j}, ..., X_{pj})^T$, $1 \le j \le n$. there exist positive constants B and σ^2 such that for any $1 \le i \le p$,

$$E|(X_{i1} - \mu_i)^2/\sigma_{ii} - 1|^r \le 1/2\sigma^2 B^{r-2} r!, \quad r \ge 2.$$

Condition 1 is analogous to the condition for the well-known Bernstein's inequality in Bennett (1962), and it is satisfied when $(X_{i1}-\mu_i)^2/\sigma_{ii}, 1 \le i \le p$ have exponential moments. In particular, Condition 1 holds true when X follows a multivariate normal distribution.

Condition 2. $\log p/n \to 0$ as $p, n \to \infty$.

Theorem 1. *Let G and $\hat{G}(t)$ be as in (2) and (3). Suppose $|\rho_{ij}| \geq k(n,p)$ for any $(i,j) \in G$. Under Condition 1, there exists a constant $c > 0$ such that for any $0 < t < k$,*

$$\mathrm{pr}\{\hat{G}(t) = G\} \geq 1 - O[p^2 \exp\{-cn(t^2 \wedge (k-t)^2)\}],$$

where $t^2 \wedge (k-t)^2 = \min\{t^2, (k-t)^2\}$. In particular, under Condition 2, if $k(n,p) \geq (C_1 \log p/n)^{1/2}$ for some constant $C_1 > 8/c$, by choosing $t = k/2$, we have:

$$\mathrm{pr}\{\hat{G}(t) = G\} \to 1.$$

$k(n,p)$ in Theorem 1 is the separation gap between zero and those nonzero off-diagonal elements of Γ. The second statement of Theorem 1 indicates that (i) is true. Theorem 2 below indicates that (ii) is true, and hence we conclude that the rate $(\log p/n)^{1/2}$ is optimal.

Theorem 2. *Suppose $X_1, ..., X_n$ are independent, identically distributed multivariate normal random vectors with mean 0 and correlation matrix $\Gamma = I_{p\times p} + (C_2 \log p/n)^{1/2}(e_1 e_2^T + e_2 e_1^T)$ where $I_{p\times p}$ is the identity matrix and e_i is the ith column of $I_{p\times p}$. Under Condition 2, when constant $C_2 > 0$ is small enough,*

$$\mathrm{pr}\{\hat{G}(t) = G\} \to 0,$$

for any $t > 0$.

3. Extension to a weakly dependent case

Denote $X_j = (X_{1j}, ..., X_{pj})^T$ and $Y_{kl}^j = \{(X_{kj}-\mu_k)/\sigma_{kk}^{1/2} + (X_{lj}-\mu_l)/\sigma_{ll}^{1/2}\}^2$, $1 \leq j \leq n, 1 \leq k, l \leq p$. In this section we assume that for any $1 \leq k, l \leq p$, $Y_{kl}^1, Y_{kl}^2, \ldots$ is a sequence of stationary, strongly mixing random variables.

For any $1 \leq k, l \leq p$, and integers $1 \leq a \leq b$, let $\mathcal{F}_{kl}^{a,b}$ denote the σ-field generated by $\{Y_{kl}^i : a \leq i \leq b\}$. Define

$$\begin{aligned}\alpha(\mathcal{F}_{kl}^{1,a}, \mathcal{F}_{kl}^{a+b,\infty}) &= \sup_{A\in\mathcal{F}_{kl}^{1,a}, B\in\mathcal{F}_{kl}^{a+b,\infty}} |\mathrm{pr}(A \cap B) - \mathrm{pr}(A)\mathrm{pr}(B)| \qquad (a, b \geq 1), \\ \alpha_{kl}(b) &= \sup_{a\geq 1} \alpha(\mathcal{F}_{kl}^{1,a}, \mathcal{F}_{kl}^{a+b,\infty}) \qquad (b \geq 1). \qquad (4)\end{aligned}$$

From the definition of strong mixing, we have $\alpha_{kl}(b) \to 0$ as $b \to \infty$. Assume now that $X_1, X_2, \ldots$ satisfy the following conditions:

CONDITION 3. (SEMIEXPONENTIAL TAIL) There exist constants $\delta > 0$, $\gamma_1 \in (0,1)$ and $K > 0$, such that

$$\sup_i E[\exp\{\delta|(X_{i1}-\mu_i)^2/\sigma_{ii} - 1|^{\gamma_1}\}] \leq K.$$

CONDITION 4. There exist positive constants γ_2, a and c such that for any $b \geq 1$ the strong mixing coefficients $\alpha_{kl}(b)$ satisfy

$$\alpha_{kl}(b) \leq a\exp(-cb^{\gamma_2}) \qquad (1 \leq k, l \leq p).$$

CONDITION 5. $\gamma < 1$ where γ is defined by $1/\gamma = 1/\gamma_1 + 1/\gamma_2$.

These three conditions are similar to the conditions in Theorem 1 of Merlevede et al. (2011). Similar to Theorem 1, we have

Theorem 3. *Let G and $\hat{G}(t)$ be as in (2) and (3). Suppose $|\rho_{ij}| \geq k(n,p)$ for any $(i,j) \in G$. Under Conditions 3 and 4 and 5, there exist constants c_1 and c_2 such that for any $0 < t < k$,*

$$\begin{aligned}\mathrm{pr}\{\hat{G}(t) = G\} \geq \quad & 1 - O[p^2 n\exp\{-c_1 n^{\gamma}(t^{\gamma} \wedge (k-t)^{\gamma})\} \\ & + p^2\exp\{-c_2 n(t^2 \wedge (k-t)^2)\}],\end{aligned}$$

where $t^2 \wedge (k-t)^2 = \min\{t^2, (k-t)^2\}$. In particular, if $\log p = O\{n^{\gamma/(2-\gamma)}\}$ and $k(n,p) \geq (C\log p/n)^{1/2}$ for some constant $C > 0$ large enough, by choosing $t = k/2$, we have:

$$\mathrm{pr}\{\hat{G}(t) = G\} \to 1.$$

4. Numerical study

In this section we provide some numerical results on covariance selection using the following three approaches: (i) Thresholding the sample correlation matrix R using a universal threshold. We shall denote this method as TR;(ii) Thresholding the sample covariance matrix S using a universal threshold as in Bickel and Levina (2008). We shall denote this method as TS; (iii) Thresholding the sample covariance matrix S using the adaptive thresholding approach as in Cai and Liu (2011). We shall denote this method as ATS. The universal threshold for the sample correlation matrix is chosen empirically using cross validation; see for example Bickel and Levina (2008). More specifically, split the n sample randomly 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 kth split, let $R_{1,k}, R_{2,k}$ be the sample correlation matrix based on the n_1 and n_2 observations respectively. The Frobenius norm of a matrix $R = (r_{ij})_{p\times p}$ is defined as $\parallel R \parallel_F^2 = \sum_{1\leq i,j\leq p} r_{ij}^2$. For a given threshold t define the thresholding operator by $T_t(R) = [r_{ij}\mathcal{I}(|r_{ij}| > t)]_{p\times p}$. We then choose t such that

$$CV(t) = \frac{1}{N}\sum_{k=1}^{N} \parallel T_t(R_{1,k}) - R_{2,k} \parallel_F^2,$$

is minimized.

In this simulation, we set $n = 100$, $p = 50, 100, 200$ and we consider the following two models:

Model 4.1 (Homoscedastic). 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 4.2 (Heteroscedastic). Following Cai and Liu (2011), let $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$, where $\sigma_{ij} = (1 - |i - j|/10)_+$ if $1 \leq i, j \leq p/2$, $\sigma_{ii} = 4$ if $p/2 + 1 \leq i \leq p$, and $\sigma_{ij} = 0$ otherwise.

For each case, the simulation is repeated 100 times. Denote the sparsity, i.e., the proportion of zero elements in the off-diagonal of Γ as ω. Let $\hat{R} = (\hat{r}_{ij})_{p\times p}$ be the thresholded sample correlation matrix obtained using the TR method or the thresholded covariance matrix obtained using the TS or ATS method. Denote the proportion of zero elements in the off-diagonal of $\hat{R}$ as $\hat{\omega}$. The mean and its standard deviation (sd) of the following quantities are computed over 100 replications: (i) L_1-loss: $|\hat{\omega} - \omega|$; (ii) Error1 $= \#\{(i, j) : 1 \leq i < j \leq p, \rho_{ij} = 0, \hat{r}_{ij} \neq 0\}$; (iii) Error2 $= \#\{(i, j) : 1 \leq i < j \leq p, \rho_{ij} \neq 0, \hat{r}_{ij} = 0\}$. Error1 is the number of times of classifying a zero element to a nonzero element while Error2 is the number of times of classifying a nonzero element to a zero element. Hence Error1+Error2 is the total misclassification error. Simulation results are given in Tables 1 and 2.

From Tables 1 and 2 we can see that the TR approach generally has smaller L_1-loss values under Models 4.1 and 4.2, indicating the TR approach can better estimate the sparsity of Σ than the other two approaches under models 4.1 and 4.2. In addition, the TR approach has smaller Error1+Error2

Table 1: Simulation results under Model 4.1 over 100 replications.

$p = 50$	TR	TS	ATS
L_1-loss(sd)	0.024(0.002)	0.037(0.003)	0.044(0.004)
Error1(sd)	42.4(2.0)	54.9(3.8)	67.0(4.4)
Error2(sd)	40.2(2.5)	43.5(3.0)	35.4(2.8)
$p = 100$			
L_1-loss(sd)	0.026(0.003)	0.046(0.005)	0.051(0.005)
Error1(sd)	182.9(9.2)	261.3(20.4)	276.1(16.1)
Error2(sd)	193.8(15.3)	203.5(20.2)	193.9(22.3)
$p = 200$			
L_1-loss(sd)	0.019(0.002)	0.038(0.004)	0.041(0.001)
Error1(sd)	700.8(29.2)	897.3(49.3)	1109.4(56.8)
Error2(sd)	755.6(39.0)	872.9(76.0)	689.0(62.1)

values, indicating that in terms of covariance selection, the TR approach can have smaller misclassification errors.

Appendix A. Technical details

For $1 \le j, k \le p, 1 \le i \le n$, denote $t_{jk} = n\sigma_{jj}^{1/2}\sigma_{kk}^{1/2}/s_{jj}^{1/2}s_{kk}^{1/2}$, $Z_{ji} = (X_{ji} - \mu_j)/\sigma_{jj}^{1/2}$ and $Z_{ji}^* = (X_{ji} - \bar{X}_j)/\sigma_{jj}^{1/2}$ where $\bar{X}_j$ is the j-th element of $\bar{X}$.

Lemma 1. *Under Condition 1, for any $1 \le j, k \le p$ and $0 < x \le K$, there exists a constant $f > 0$, depending on K, B and σ^2 only, such that*

$$\text{pr}\Big[\Big|\sum_{i=1}^{n}\Big\{(Z_{ji}^* + Z_{ki}^*)^2 - 2(1+\rho_{jk})\Big\}\Big| \ge nx\Big] \le \exp(-fnx^2); \tag{A.1}$$

$$\text{pr}\Big[\Big|\sum_{i=1}^{n}\Big\{(Z_{ji}^* - Z_{ki}^*)^2 - 2(1-\rho_{jk})\Big\}\Big| \ge nx\Big] \le \exp(-fnx^2). \tag{A.2}$$

Proof. Define $\bar{Z}_j = 1/n\sum_{i=1}^{n} Z_{ji}$. By verifying the condition of the Bernstein's inequality using Condition 1, see (7) of Bennett (1962), we have:

$$\text{pr}(\bar{Z}_j^2 \ge x) \le \exp(-d_1nx^2), \tag{A.3}$$

Table 2: Simulation results under Model 4.2 over 100 replications.

$p = 50$	TR	TS	ATS
L_1-loss(sd)	0.035(0.001)	0.057(0.003)	0.047(0.001)
Error1(sd)	2.4(0.4)	44.0(3.7)	0.6(0.2)
Error2(sd)	44.9(1.0)	103.6(1.9)	57.9(1.3)
$p = 100$			
L_1-loss(sd)	0.023(<0.001)	0.030(0.001)	0.030(0.001)
Error1(sd)	4.7(0.7)	126.6(5.2)	2.6(0.4)
Error2(sd)	117.9(1.3)	273.2(2.9)	148.8(2.5)
$p = 200$			
L_1-loss(sd)	0.014(<0.001)	0.011(0.001)	0.018(<0.001)
Error1(sd)	4.5(0.4)	588.6(20.6)	4.2(0.7)
Error2(sd)	289.2(2.5)	557.0(5.7)	353.0(6.1)

$$\text{pr}\Big[\Big|\sum_{i=1}^{n}\Big\{(Z_{ji}+Z_{ki})^2 - 2(1+\rho_{jk})\Big\}\Big| \geq nx\Big] \leq \exp(-d_2 n x^2), \qquad \text{(A.4)}$$

for some positive constants d_1, d_2 depending on K, B and σ^2 only. From (A.3), (A.4) and the following fact

$$\sum_{i=1}^{n}\Big(Z_{ji}^{*}+Z_{ki}^{*}\Big)^2 = \sum_{i=1}^{n}\Big(Z_{ji}+Z_{ki}\Big)^2 - n\bar{Z}_j^2 - n\bar{Z}_k^2 - 2n\bar{Z}_j\bar{Z}_k,$$

we conclude (A.1) holds. (A.2) can be proved similarly.

The following proposition is a key in proving Theorems 1, 2 and 3. It gives a Bernstein type inequality for elements of the sample correlation matrix.

Propositon 1. *Under Condition 1, for any $0 < v \leq 2$ and $1 \leq j, k \leq p$, there exist constants $d_1 > 0$ and $d_2 > 0$, depending on B and σ^2 only, such that*

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

Proof. When $\rho_{jk} = \pm 1$, LHS of the inequality equals zero, and so the inequality holds. Now we consider the case: $-1 < \rho_{jk} < 1$.

$$\begin{aligned} \mathrm{pr}(|r_{jk} - \rho_{jk}| \geq v) &= \mathrm{pr}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ji}^* Z_{ki}^* t_{jk} - \rho_{jk}\Big| \geq v\Big) \\ &\leq \mathrm{pr}\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ji}^* Z_{ki}^* (t_{jk} - 1)\Big| \geq \frac{v}{2}\Big\} \\ &\quad + \mathrm{pr}\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} (Z_{ji}^* Z_{ki}^* - \rho_{jk})\Big| \geq \frac{v}{2}\Big\}. \qquad (A.5) \end{aligned}$$

From Lemma 1 and the following equality

$$\begin{aligned} \sum_{i=1}^{n}\Big(Z_{ji}^* Z_{ki}^* - \rho_{jk}\Big) &= \frac{1}{4}\Big[\sum_{i=1}^{n}\Big\{\Big(Z_{ji}^* + Z_{ki}^*\Big)^2 - 2\Big(1 + \rho_{jk}\Big)\Big\} \\ &\quad - \sum_{i=1}^{n}\Big\{\Big(Z_{ji}^* - Z_{ki}^*\Big)^2 - 2\Big(1 - \rho_{jk}\Big)\Big\}\Big], \end{aligned}$$

we conclude that there exists a constant $f_1 > 0$, depending on B and σ^2 only, such that,

$$\mathrm{pr}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}(Z_{ji}^* Z_{ki}^* - \rho_{jk}\Big)\Big| \geq \frac{v}{2}\Big) \leq \exp(-f_1 n v^2). \qquad (A.6)$$

On the other hand, let $a = \frac{v}{2(|\rho_{jk}|+v)}$, we have:

$$\begin{aligned} \mathrm{pr}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} Z_{ji}^* Z_{ki}^* (t_{jk} - 1)\Big| \geq \frac{v}{2}\Big) &\leq \mathrm{pr}\Big(\Big|\sum_{i=1}^{n} Z_{ji}^* Z_{ki}^*\Big| \geq \frac{nv}{2a}\Big) \\ &\quad + \mathrm{pr}(|t_{jk} - 1| > a). \qquad (A.7) \end{aligned}$$

Similar to (A.6), there exists a constant $f_2 > 0$ such that,

$$\begin{aligned} \mathrm{pr}\Big(\Big|\sum_{i=1}^{n} Z_{ji}^* Z_{ki}^*\Big| \geq \frac{nv}{2a}\Big) &\leq \mathrm{pr}\Big(\Big|\sum_{i=1}^{n} Z_{ji}^* Z_{ki}^* - n\rho_{jk}\Big| \geq nv\Big) \\ &\leq \exp(-f_2 n v^2). \qquad (A.8) \end{aligned}$$

$$
\begin{aligned}
\mathrm{pr}(|t_{jk}-1|>a) &= \mathrm{pr}\Big\{\sum_{i=1}^n Z_{ji}^{*2}\sum_{i=1}^n Z_{ki}^{*2} < \frac{n^2}{(1+a)^2}\Big\} \\
&\quad +\mathrm{pr}\Big\{\sum_{i=1}^n Z_{ji}^{*2}\sum_{i=1}^n Z_{ki}^{*2} > \frac{n^2}{(1-a)^2}\Big\} \\
&\le \mathrm{pr}\Big(\sum_{i=1}^n Z_{ji}^{*2} < \frac{n}{1+a}\Big) + \mathrm{pr}\Big(\sum_{i=1}^n Z_{ki}^{*2} < \frac{n}{1+a}\Big) \\
&\quad +\mathrm{pr}\Big(\sum_{i=1}^n Z_{ji}^{*2} > \frac{n}{1-a}\Big) + \mathrm{pr}\Big(\sum_{i=1}^n Z_{ki}^{*2} > \frac{n}{1-a}\Big) \\
&\le \mathrm{pr}\Big\{\Big|\sum_{i=1}^n \Big(Z_{ji}^{*2}-1\Big)\Big| > \frac{an}{1+a}\Big\} \\
&\quad +\mathrm{pr}\Big\{\Big|\sum_{i=1}^n \Big(Z_{ki}^{*2}-1\Big)\Big| > \frac{an}{1+a}\Big\}.
\end{aligned}
$$

Again by Lemma 1, there exists a constant $f_3 > 0$ independent of n such that,

$$
\mathrm{pr}(|t_{jk}-1|>a) \le 2\exp\{-f_3 n a^2/(1+a)^2\} \le 2\exp(-f_3 n v^2/64). \quad \text{(A.9)}
$$

The theorem is proved by combining (A.5), (A.6), (A.7), (A.8) and (A.9).

Proof of Theorem 1. Theorem 1 follows from Proposition 1 and the following inequality.

$$
\begin{aligned}
\mathrm{pr}\{\hat{G}(t)\neq G\} &\le p^2\mathrm{pr}\{|r_{ij}-\rho_{ij}|\ge|k-t|,(i,j)\in G^c\} \\
&\quad +p^2\mathrm{pr}\{|r_{ij}|\ge t,(i,j)\in G^c\}.
\end{aligned}
$$

Proof of Theorem 2. Without loss of generality assume that p is an odd number. It suffices to show that when constant C_2 is small enough,

$$
\mathrm{pr}\{|r_{12}| < \max(|r_{34}|,|r_{56}|\ldots,|r_{p-1,p}|)\}\to 1.
$$

Since $X_1\ldots,X_n$'s are multivariate normal and $\Gamma = I_{p\times p}+(C_2\log p/n)^{1/2}(e_1e_2^T+e_2e_1^T)$, we know that $r_{12},r_{34},\ldots,r_{p-1,p}$ are independent. Therefore it suffices to show that

$$
\mathrm{pr}(|r_{12}-\rho_{12}|\le\rho_{12})\{1-\mathrm{pr}(|r_{34}|\le 2\rho_{12},\ldots,|r_{p-1,p}|\le 2\rho_{12})\}\to 1. \quad \text{(A.10)}
$$

We shall prove (A.10) by showing:

$$\mathrm{pr}(|r_{12}-\rho_{12}|\le\rho_{12})\to 1, \tag{A.11}$$

and

$$\mathrm{pr}(|r_{34}|\le 2\rho_{12},\ldots,|r_{p-1,p}|\le 2\rho_{12})=\mathrm{pr}(|r_{34}|\le 2\rho_{12})^{(p-2)/2}\to 0. \tag{A.12}$$

By Markov's inequality, we have

$$\mathrm{pr}(|r_{12}-\rho_{12}|\le\rho_{12})\ge 1-E(r_{12}-\rho_{12})^2/\rho_{12}^2.$$

Together with the following observation from Kendall (1960),

$$E(r_{12}-\rho_{12})^2=O(1/n),$$

we conclude that (A.11) is true.

Now using the density function of r_{34}, see for example Anderson (2003), and (4.2) of Bustoz and Ismail (1986) we have

$$\begin{aligned}
\mathrm{pr}(|r_{34}|\le 2\rho_{12}) &= 1-2\int_{2\rho_{12}}^{1}\frac{\Gamma(n/2-1/2)}{\Gamma(n/2-1)\,\pi^{1/2}}(1-r^2)^{(n-4)/2}\mathrm{d}r\\
&\le 1-\frac{2}{\pi^{1/2}}\int_{2\rho_{12}}^{3\rho_{12}}(n/2-2)^{1/2}(1-r^2)^{n/2}\mathrm{d}r\\
&\le 1-\frac{2^{1/2}(n-4)^{1/2}}{\pi^{1/2}}\rho_{12}(1-9\rho_{12}^2)^{n/2}\\
&\to 1-\frac{(2C_2)^{1/2}\log^{1/2}p}{\pi^{1/2}p^{9C_2/2}}.
\end{aligned}$$

Therefore there exists a constant $0<c<(2C_2/\pi)^{1/2}$ such that when p,n are large enough,

$$\begin{aligned}
\mathrm{pr}(|r_{34}|\le 2\rho_{12})^{(p-2)/2} &\le \left\{1-\left(c\log p/p^{9C_2}\right)^{1/2}\right\}^{p/2-1}\\
&\to \exp\{-(c\log p/p^{9C_2-2})^{1/2}/2\}
\end{aligned}$$

which will tends to 0 as long as $C_2<2/9$. Hence we conclude that (A.12) is true when C_2 is small enough.

Proof of Theorem 3. Under Conditions 3, 4 and 5, by Theorem 1 in Merlevede et al. (2011) we can obtain a Bernstein type inequality for the sample correlation coefficients as in Proposition 1. The rest of the proof is similar to the proof of Theorem 1.

Acknowledgements

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 a referee for valuable comments and suggestions.

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, 3rd ed. Wiley Series in Probability and Statistics. Wiley, New York.

Bennett, G., 1962. Inequalities for the sum of independent random variables. Journal of the American Statistical Association 48, 33-45.

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

Kendall, M.G., 1960. The evergreen correlation coefficient. em Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling, 274-277. Stanford University Press, Stanford.

Merlevede, F., Peligrad, M., Rios, E. 2011. A Bernstein type inequality and moderate deviations for weakly dependent sequences. Probability Theory and Related Fields 151, 435-474.

Rothman, A.J., Levina, E., Zhu, J., 2008. Generalized thresholding of large covariance matrices. Journal of the American Statistical Association 104, 177-186.