Consistency of large dimensional sample covariance matrix under weak dependence

Monika Bhattacharjee
Statistics and Mathematics Unit
Indian Statistical Institute
203 B.T. Road, Kolkata 700108
INDIA
monaiidexp.gamma@gmail.com

Arup Bose *
Statistics and Mathematics Unit
Indian Statistical Institute
203 B.T. Road, Kolkata 700108
INDIA
bosearu@gmail.com

This version January 29, 2013

Dedicated to the memory of Kesar Singh

Abstract

Convergence rates for banded and tapered estimates of large dimensional covariance matrices is known when the vector observations are independent and identically distributed. We investigate the case where the independence does not hold. Our models can accommodate suitable patterned cross covariance matrices. These estimators remain consistent in operator norm with appropriate rates of convergence under suitable class of models.

Key words and phrases. High-dimensional data, covariance matrices, cross covariances, regularization, banding, tapering, convergence rate, operator norm.

AMS 2010 Subject Classifications. Primary 62H12; Secondary 62F12, 65F35

1 Introduction

New technologies and methods in medical sciences, image processing and the internet, and many other fields of science generate data where the dimension is high and the sample size is small relative to the dimension. For example, microarray data [Dudoit et al., 2002] contains gene expression for tens of thousands of genes (rows) on a few observations (columns). Another

*Research supported by J.C. Bose National Fellowship, Dept. of Science and Technology, Govt. of India.

example is fMRI data, which measures the hemodynamic response in hundreds of thousands of voxels (rows) for only a few subjects or replicates (columns). Similarly, the Netflix movie rating data [Bennett and Lanning, 2007] contains the rating information for approximately 480,000 customers (columns) on 18,000 movies (rows). Let $X_{p\times n}$ denote the corresponding data matrix:

$$X_{p\times n} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & x_{p3} & \dots & x_{pn} \end{bmatrix}$$

where the dimension $p = p(n)$ is assumed to be increasing with the sample size n. This type of data matrix has been modeled by identical Gaussian distribution of the columns

$$C_{ip} = (x_{1i}, x_{2i}, \dots, x_{pi})' \sim \mathcal{N}(\mu, \Sigma_p) \; \forall i = 1, 2, \dots, n,$$

with mean vector $\mu \in \mathbb{R}^p$ and variance-covariance matrix $\Sigma_p = ((\sigma_{ij}))_{p\times p}$. The estimation of the large covariance matrix Σ_p is crucial for statistical inference procedures.

Usually one assumes further that the columns are independent/exchangeable. Thus, genes in microarrays, pixels in images, voxels in fMRIs and movies in Netflix movie-rating data are considered as dependent features whereas respectively the samples, repeated images, images with respect to different subjects or replications and customers are modeled to be independent.

However, this assumption has been questioned. Many have suggested that, in microarrays, the arrays are not independent (e.g., [Owen, 2005], [Efron, 2009], [Klebanov and Yakovlev, 2007], [Leek and Storey, 2008]). Latent variables such as age, gender, family history, underlying health status, measurement process, laboratory conditions may be responsible for the dependency between two patients. For the Netflix movie-rating data, a particular type of movies are likely to have similar ratings from customers having similar tastes. The latent variable time may be responsible for the dependency between two replications in fMRI data sets. Specific examples of this lack of independence can be found in [Allen and Tibshirani, 2010].

Hence, there is need for models which allow for dependence between columns. [Efron, 2009] proposed the matrix-variate normal as a model for microarrays. Mean-restricted matrix-variate normal was considered by [Allen and Tibshirani, 2010]. This distribution, denoted by $X_{p\times n} \sim \mathcal{N}_{p,n}(\nu, \mu, \Sigma_p, \Delta)$, has separate mean and covariance parameters for the rows, $\nu \in \mathcal{R}^p$, $\Sigma_p = ((\sigma_{ij}))_{p\times p}$, and the columns, $\mu \in \mathcal{R}^n$, $\Delta = ((\delta_{ij}))_{n\times n}$. If the matrix is transformed into a vector of length np, we have that $vec(X) \sim \mathcal{N}(vec(M), \Omega)$, where $M = ((\nu_i + \mu_j))_{p\times n}$, $\Omega = \Delta \otimes \Sigma_p$ and $\otimes$ is the Kronecker product between two matrices. In this model the correlation between columns is controlled without considering the effect of the components (rows); that is,

$$\frac{corr(x_{ki}, x_{lj})}{corr(x_{mi}, x_{mj})} = \frac{\delta_{kl}}{\sqrt{\delta_{kk}\delta_{ll}}} \; \forall i, j = 1, 2, \dots p \;\text{ and }\; m = 1, 2, \dots, n.$$

We will assume that $C_{ip} \sim \mathcal{N}_p(0, \Sigma_p) \; \forall i = 1, 2, \dots, n$ are identically distributed and the distribution is Gaussian with zero mean. However, we will allow dependence of appropriate nature between the columns. We call this dependence the *cross covariance strucuture*. In this paper we work under three different restrictions on cross covariance structures. In one case, the restriction is on the growth of the powers of the trace of certain matrices derived from the cross

covariance structure. In the second case, the dependence among any two columns weakens as the lag between them increases and in the third case we assume weak dependence among the last few columns. See Section 2 for details.

The existing methods to estimate Σ_p (under column independence) involves banding or tapering of the sample variance-covariance matrix. [Bickel and Levina, 2008a] proved that suitably banded and tapered estimators are both consistent in the operator norm for the sample variance-covariance matrix as long as $n^{-1} \log p \to 0$ uniformly over some fairly natural well-conditioned families of covariance matrices. They also obtained some explicit rates.

In the first case, we show that the convergence rate of the banded estimator is the same as in the *i.i.d.* case of [Bickel and Levina, 2008a] (see Theorem 3.1 of Section 3). We also provide some sufficient conditions that imply the trace condition. The other two cases do not fall under the purview of Theorem 3.1. Under appropriate conditions we obtain explicit rates of convergence for the banded estimators (see Theorems 3.2 and 3.3). In particular, for all three cases the estimators continue to remain consistent in operator norm.

Banded estimators are not necessarily positive definite. So we consider tapered estimators that preserve the positive definiteness of the sample variance covariance matrix. We obtain the rates of convergence of the tapered estimator for all three cases (see Theorems 3.4, 3.5 and 3.6). In particular the tapered estimator continues to remain consistent in operator norm under these dependent situations.

2 The model, assumptions and examples

We assume that the column variables $C_{ip} = (x_{1i}, x_{2i}, \ldots, x_{pi})'$, $1 \leq i \leq n$ are jointly normal and marginally they are identically distributed with mean zero and $\mathrm{Var}(C_{ip}) = \Sigma_p$. The problem is to estimate Σ_p as n and p both tend to ∞. In particular, the dimension p may be much larger than the sample size n. The set of all *well-conditioned* covariance matrices is defined as

$$W(\epsilon) = \left\{\Sigma_{\infty\times\infty} : 0 < \epsilon \leq \lambda_{\min}(\Sigma_p) \leq \lambda_{\max}(\Sigma_p) \leq \epsilon^{-1} < \infty\right\}$$

where

$$\begin{aligned} \Sigma_p &= p - \text{th order principal minor of } \Sigma_{\infty\times\infty}, \\ \lambda_{\max}(\Sigma_p) &= \text{largest eigenvalue of } \Sigma_p, \\ \lambda_{\min}(\Sigma_p) &= \text{smallest eigenvalue of } \Sigma_p, \end{aligned}$$

and ϵ is independent of p. [Bickel and Levina, 2008a] assumed that the columns are independent and $\Sigma_{\infty\times\infty}$ belongs to a an appropriate subclass of $\mathcal{W}(\epsilon)$. Without such an assumption consistency cannot be achieved. We consider a slightly modified subclass:

$$\mathcal{U}(\epsilon, \alpha, C) = \left\{\Sigma_{\infty\times\infty} \equiv ((\sigma_{ij})) \in W(\epsilon) \cap \mathcal{V} : \max_j \sum_{i:|i-j|>k} |\sigma_{ij}| \leq Ck^{-\alpha},\ \forall k > 0\right\}$$

where $\mathcal{V} = \left\{\Sigma_{\infty\times\infty} \equiv ((\sigma_{ij})) : \sigma_{ij} \neq 0\ \forall i, j\right\}$, i.e., we impose the condition that the entries of $\Sigma_{\infty\times\infty}$ are non-zero on the class of covariance matrices that [Bickel and Levina, 2008a] considered.

We now come to the dependence structure of the columns. Let

$$\begin{aligned} \mathrm{Cov}(C_{ip}, C_{jp}) &= \Lambda_{ij} * \Sigma_p \ \ \forall i \neq j,\ 1 \leq i, j \leq n, \\ \Lambda_{ji} &= \Lambda'_{ij} \ \ \forall i \neq j,\ 1 \leq i, j \leq n \end{aligned}$$

where each Λ_{ij} is a $p \times p$ matrix and $*$ is Schur multiplication (component wise multiplication) of two matrices. If we transform the data matrix into a vector of length np, we have

$$vec(X_{p\times n}) = (C'_{1p}, C'_{2p}, \ldots, C'_{np})' \sim \mathcal{N}(0, \Delta_{np}),$$

where

$$\Delta_{np} = \begin{bmatrix} \Sigma_p & \Lambda_{12} * \Sigma_p & \Lambda_{13} * \Sigma_p & \ldots & \Lambda_{1n} * \Sigma_p \\ \Lambda'_{12} * \Sigma_p & \Sigma_p & \Lambda_{23} * \Sigma_p & \ldots & \Lambda_{2n} * \Sigma_p \\ \Lambda'_{13} * \Sigma_p & \Lambda'_{23} * \Sigma_p & \Sigma_p & \ldots & \Lambda_{3n} * \Sigma_p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Lambda'_{1n} * \Sigma_p & \Lambda'_{2n} * \Sigma_p & \Lambda'_{3n} * \Sigma_p & \ldots & \Sigma_p \end{bmatrix}.$$

The condition $\Sigma_p \in \mathcal{V}$, assures that one can recover $\{\Lambda_{ij}\}$ from the matrices Δ_{np} and Σ_p. Note that the covariance matrix of a finite order moving average process does not belong to this class. However, the covariance matrices of all ARMA processes with a non-trivial autoregressive component are in this class.

Example 2.1. Suppose

$$C_{ip} = A_p C_{(i-1)p} + Z_{ip},\ \forall i = 0, \pm 1, \pm 2, \ldots$$

where Z_{ip}'s are each p-component vector and *i.i.d.* with mean zero and $\mathrm{Var}(Z_{ip}) = \tilde{\Sigma}_p$. Also, A_p is a symmetric square matrix of order p such that $\|A_p\|_2 < 1$ and $A_p\tilde{\Sigma}_p = \tilde{\Sigma}_p A_p$ for all p. Here, $\|M\|_2 = \sqrt{\lambda_{\max}(M'M)}$. From the properties of linear operators [Bhatia, 2009], if $\|A\|_2 < 1$, then $(I - A)$ is invertible and $(I - A)^{-1} = (I + A + A^2 + \ldots)$. Let, $\Sigma_p = (I - A_p^2)^{-1}\tilde{\Sigma}_p$. Hence,

$$\Delta_{np} = \mathrm{Var}\left(vec(X_{p\times n})\right) = \left(\left(\Sigma_p A_p^{|i-j|} I(i \neq j) + \Sigma_p I(i = j)\right)\right)_{1\leq i,j\leq n}.$$

As we have mentioned earlier, if $\Sigma_p \in \mathcal{V}$ then one can express Δ_{np} as

$$\Delta_{np} = \left(\left(\Sigma_p * \Lambda_{ij} I(i \neq j) + \Sigma_p * J_p I(i = j)\right)\right)_{1\leq i,j\leq n}.$$

where J_p is a $p \times p$ matrix with all entries equal to one. It is relevant to address the issue of estimation of $\tilde{\Sigma}_p$ and A_p. This will be addressed elsewhere.

Example 2.2. Suppose, $\{Z_{ip},\ i = 0, \pm 1, \pm 2, \ldots\}$ is a sequence of p-component random vector such that $E(Z_{ip}) = 0,\ \forall i$ and $\mathrm{Cov}(Z_{ip}, Z_{jp}) = \Lambda_{|i-j|}\ \forall i, j$. Also, let Y_p be a mean zero p-component random vector such $\mathrm{Var}(Y_p) = \Sigma_p$ and independent of Z_{ip}'s. We define another sequence of p-component mean zero random vector as

$$C_{ip} = Y_p * Z_{ip},\ i = 0, 1, 2, \ldots n.$$

Clearly, we have $\Delta_{np} = \mathrm{Var}\left(vec(X_{p\times n})\right) = \left(\left(\Sigma_p * \Lambda_{|i-j|}\right)\right)_{1\leq i,j\leq n}$.

In the above examples, the correlation among the columns has the Toeplitz structure, $\Lambda_{ij} = \Lambda_{|i-j|},\ \forall 1 \le i, j \le n$.

Example 2.3. Let $\Delta_{np} = \left(\left(B_p^{i+j} I(i \neq j) + (I - B_p^2)^{-1} I(i = j)\right)\right)_{1 \le i,j \le n}$ where B_p is a symmetric $p \times p$ matrix and $\|B_p\|_2 < 1$ for all p. Then Δ_{np} is always positive semi-definite since

$$\Delta_{np} = \left(B_p\ B_p^2\ \ldots\ B_p^n\right)' \left(B_p\ B_p^2\ \ldots\ B_p^n\right) + Diag\left(I_p + \sum_{i=1}^{\infty} B_p^{2i} - B_p^{2k},\ k = 1, 2, \ldots, n\right)$$

where $Diag(A_i,\ i = 1, 2, \ldots, n)$ denotes the block-diagonal matrix with i-th diagonal block as A_i and I_p is the identity matrix of order p.

Cross covariance structures. We can separate out the covariance structure from Δ_{np} and define

$$\nabla_{np} = \begin{bmatrix} J_p & \Lambda_{12} & \Lambda_{13} & \ldots & \Lambda_{1n} \\ \Lambda_{12}' & J_p & \Lambda_{23} & \ldots & \Lambda_{2n} \\ \Lambda_{13}' & \Lambda_{23}' & J_p & \ldots & \Lambda_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Lambda_{1n}' & \Lambda_{2n}' & \Lambda_{3n}' & \ldots & J_p \end{bmatrix}$$

Broadly speaking, weak dependence between columns can be modelled by assuming that Λ_{ij} is "small" when say both indices i and j are large or when $|i - j|$ is large. In particular, for fMRI data, since time is one of the latent variables, which is responsible for the dependence, one may consider the Toeplitz (stationary) structure $\Lambda_{ij} = \Lambda_{|i-j|}$ for a suitable sequence of matrices $\{\Lambda_i\}$ and if $\Lambda_i = \Lambda^i$, then it yields the autoregressive structure of Example 2.1.

We shall need the following classes of weakly dependent cross covariance structures.

Let, $\{a_n\}_{n=1}^{\infty}$ be a sequence of non-negative integers such that $n^{-1}a_n < 1,\ \forall n \ge 1$. Weak dependence between i-th and j-th columns when $|i - j|$ is large is modelled as follows:

$$\begin{aligned} \mathcal{A}_n(a_n) &= \left\{\nabla_{np} = \left(\left(\Lambda_{|i-j|}\right)\right) : S(a_n) := \max_{a_n \le k \le n} \| \wedge_k \|_\infty = O\left(n^{-2}a_n\right)\right\}, \\ \mathcal{A}(a_n, n \ge 1) &= \left\{\nabla_{\infty\times\infty} = \left(\left(\Lambda_{|i-j|}\right)\right) : \nabla_{np} \in \mathcal{A}_n(a_n)\right\}. \end{aligned}$$

Weak dependence among the columns when i and j are large can be modelled as follows:

$$\begin{aligned} \mathcal{L}_n(a_n) &= \left\{\nabla_{np} : S'(a_n) := \max_{k \ge 1, m \ge 1} \|\Lambda_{a_n+k, a_n+k+m}\|_\infty = O\left(n^{-2}a_n\right)\right\}, \\ \mathcal{L}(a_n, n \ge 1) &= \left\{\nabla_{\infty\times\infty} : \nabla_{np} \in \mathcal{L}_n(a_n)\right\}. \end{aligned}$$

Finally, weak dependence among columns when $(i + j)$ is large can be modelled by:

$$\begin{aligned} \mathcal{H}_n(a_n) &= \left\{\nabla_{np} = \left(\left(\Lambda_{i+j} I(i \neq j) + \Lambda_0 I(i = j)\right)\right) : \max_{r \ge a_n} \|\Lambda_r\|_\infty = O\left(n^{-2}a_n\right)\right\}, \\ \mathcal{H}(a_n, n \ge 1) &= \left\{\nabla_{\infty\times\infty} = \left(\left(\Lambda_{i+j} I(i \neq j) + \Lambda_0 I(i = j)\right)\right) : \nabla_{np} \in \mathcal{H}_n(a_n)\right\}. \end{aligned}$$

Clearly, $\mathcal{H}(a_n, n \ge 1) \subset \mathcal{L}\left(\left[2^{-1}a_n\right] + 2, n \ge 1\right)$, where $[x]$ is the largest integer contained in x. So, all bounds for the latter will automatically hold for the former.

3 Results

The sample covariance matrix $\hat{\Sigma}_p = n^{-1}\sum_{i=1}^{n} C_{ip}C_{ip}'$ is the optimal estimator when p is fixed. However, if $p \to \infty$, $\hat{\Sigma}_p$ can behave badly unless it is regularized. There is a growing literature (for example, see [Bickel and Levina, 2008a], [Bickel and Levina, 2008b], [Cai et al., 2010], [Rothman et al., 2008]) on regularization. Here, we discuss two methods of regularization namely banding and tapering following [Bickel and Levina, 2008a].

We will first provide a trace condition under which the banded estimator has the same convergence rate as we have in the case of *i.i.d.* column variables (see Theorem 3.1). Then we derive the appropriate convergence rate when the cross covariance structures are in the classes described in the previous section (see Theorems 3.2 and 3.3). Then we will consider regularization by tapering to preserve the positive definiteness and derive appropriate convergence rates for tapered estimators (see Theorems 3.4, 3.5 and 3.6).

Banding. For any square matrix $M = ((m_{ij}))$ of order p and for any $0 \le k \le p$, we define

$$B_k(M) = \left(\left(m_{ij}1\left(|i-j| \le k\right)\right)\right).$$

$B_k(M)$ is called the k-banded version of M, and k is called the banding parameter.

Let, $\rho_{jk} = \sigma_{jk}(\sigma_{jj}\sigma_{kk})^{-\frac{1}{2}}$ and Γ_+^{jk} and Γ_-^{jk} be two $(n \times n)$ matrices defined by

$$\Gamma_{\pm}^{jk}(p,q) = \begin{cases} \frac{\Lambda_{pq}(jj) \pm \left(\Lambda_{pq}(jk)+\Lambda_{pq}(kj)\right)\rho_{jk}+\Lambda_{pq}(kk)}{2\left(1\pm\rho_{jk}\right)}, & p \ne q \\ 1, \qquad p = q, \end{cases}$$

$1 \le j,k \le p$. Then we obtain the same rate as in *i.i.d.* sample derived by [Bickel and Levina, 2008a].

Theorem 3.1. *Let $X_{p\times n} \sim \mathcal{N}_{np}(0, \Delta_{np})$ for all n and $\Sigma_{\infty\times\infty} \in \mathcal{U}(\epsilon, \alpha, C)$. If for some $M > 0$,*

$$\sup_{n,j,k} n^{-1}\mathrm{Tr}\left((\Gamma_{\pm}^{jk})^r\right) \le M^r \tag{3.1}$$

then for $k_n \asymp (n^{-1}\log p)^{\frac{-1}{2(\alpha+1)}}$, we have $\|B_{k_n}(\hat{\Sigma}_p) - \Sigma_p\|_2 = O_P[(n^{-1}\log p)^{\frac{\alpha}{2(\alpha+1)}}]$.

Remark 3.1. *Suppose, $\{x_k\}$ is a sequence of real numbers such that $x_k = x_{-k}$ for all k and $T_n = ((x_{i-j}))_{1\le i,j\le n}$. If $\|\Lambda_{ij}\|_\infty \le x_{i-j}$ $\forall i \ne j$, $1 \le i,j \le n$, then*

$$\mathrm{Tr}\left(\left(\Gamma_{\pm}^{jk}\right)^r\right) \le \mathrm{Tr}\left((T_n)^r\right) \quad \forall 1 \le j,k \le p,\ r = 1,2,\ldots.$$

So, in this case we can explore if there exists a constant $M > 0$ exist such that

$$\sup_n n^{-1}\,\mathrm{Tr}\left((T_n)^r\right) \le M^r, \quad \forall r \ge 1. \tag{3.2}$$

Using Lemma 4.1, one can show that if $\sum|x_k| < \infty$, then (3.2) holds.

Now consider a function $f : [0, 2\pi] \to \mathbb{R}$ which is non-negative, symmetric (about π) and square integrable but is unbounded. Let $\{\hat{f}_k\}$ be the Fourier coefficients of f. Then they are all real and $\hat{f}_k = \hat{f}_{-k}$ for all k. Let $x_k = |\hat{f}_k|$. Let $T_n = ((x_{i-j}))_{1\le i,j\le n}$ and $T_{f,n} = ((\hat{f}_{i-j}))_{1\le i,j\le n}$. Then

$$n^{-1}\,\mathrm{Tr}\left(\left(T_{f,n}\right)^r\right) \le n^{-1}\,\mathrm{Tr}\left((T_n)^r\right) \quad \forall r.$$

Now suppose (3.2) holds for T_n. Then it also holds for $T_{f,n}$. Hence by Lemma 4.2, f is almost everywhere bounded, which is a contradiction. So (3.2) cannot hold for this choice of $\{x_k\}$.

Remark 3.2. *Let,* $\Lambda_{ij} = 0, \ \forall |i-j| > k$. *Then* (3.1) *will hold if*

$$\sum_{l=1}^{k} \left(\sup_{|i-j|=l} ||\Lambda_{ij}||_\infty \right) < \infty.$$

As a special case if $\Lambda_{ij} = \Lambda_{|i-j|} \ \forall i, j$ *and* $\Lambda_r = 0 \ \forall r > k$, *Then* (3.1) *will hold if*

$$||\Lambda_r||_\infty < \infty, \ \forall 1 \leq r \leq k.$$

Now, if $\nabla_{\infty\times\infty} \in \mathcal{L}(a_n, \ n \geq 1)$ or $\mathcal{A}(a_n, \ n \geq 1)$, then we cannot say whether (3.1) will hold or not. In these classes, we do not have any control over Λ_{ij} for $\min(i, j) < a_n$ or $|i-j| < a_n$ respectively and moreover $a_n \to \infty$. As the following theorems show, we have a slower rate of convergence for the two classes.

Theorem 3.2. *Suppose* $X_{p\times n} \sim N_{np}(0, \triangle_{np})$. *If* $\sum_{\infty\times\infty} \in \mathcal{U}(\epsilon, \alpha, c)$ *and* $\nabla_{\infty\times\infty} \in \mathcal{L}(l_n, \ n \geq 1)$ *for some non-decreasing sequence* $\{l_n\}_{n\geq 1}$ *of non-negative integers such that* $n^{-1} l_n \log p \to 0$ *as* $n \to \infty$ *and* $\liminf n^{-1} l_n^2 \log p > 0$. *Then with* $k_n \asymp \left(n^{-1} l_n \log p\right)^{-\frac{1}{1+\alpha}}$,

$$||B_{k_n}(\hat{\Sigma}_p) - \Sigma_p||_2 = O_P\left((n^{-1} l_n \log p)^{\frac{\alpha}{\alpha+1}}\right).$$

Remark 3.3. *Theorem 3.2 will not be applicable in case the sequence* $\{l_n\}$ *is bounded above. This is because if* $n^{-1} l_n \log p \to 0$ *then* $n^{-1} \log p \to 0$ *and hence* $n^{-1} l_n^2 \log p \to 0$ *.However, when* $\{l_n\}$ *is bounded by K,* $C_{(K+1)p}$, $C_{(K+2)p}, \ldots$ *will be i.i.d. sample and we can construct the estimator on the basis of this i.i.d. sample.*

Theorem 3.3. *Suppose* $X_{p\times n}$, $Vec(X_{p\times n}) \sim \mathcal{N}_{np}(0, \triangle_{np})$. *If* $\sum_{\infty\times\infty} \in \mathcal{U}(\epsilon, \alpha, C)$ *and* $\nabla_{\infty\times\infty} \in \mathcal{A}(a_n, n \geq 1)$ *for some non-decreasing sequence* $\{a_n\}_{n\geq 1}$ *of non-negative integers such that* $a_n \sqrt{n^{-1} \log p} \to 0$ *and* $a_n^{-1} \sqrt{n \log p} \to \infty$ *as* $n \to \infty$. *Then with* $k_n \asymp \left(a_n n^{-\frac{1}{2}} \sqrt{\log p}\right)^{-\frac{1}{1+\alpha}}$,

$$||B_{k_n}(\hat{\Sigma}_p) - \Sigma_p||_2 = O_P\left(\left(a_n \sqrt{n^{-1} \log p}\right)^{\frac{\alpha}{\alpha+1}}\right).$$

Remark 3.4. *If the sequence* $\{a_n\}$ *is bounded above, then the rate of convergence will reduce to the convergence rate for i.i.d. sample as mentioned in Theorem 3.1 .*

Tapering. Positive definiteness is a desirable property for estimators of any variance-covariance matrix and the banded version of $\hat{\Sigma}_p$ may not be a positive definite matrix [Bickel and Levina, 2008a]. As the Schur product of two positive definite matrices is always positive definite, one can consider the Schur product of $\hat{\Sigma}_p$ with an appropriate positive definite matrix to preserve the positive definiteness This leads to the idea of tapering.

Let, $g : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+ \cup \{0\}$ be a continuous, non-increasing function such that $g(0) = 1$, $\int_0^\infty g(x) < \infty$ and $1 - g(x) = O(x^\gamma)$ for some $\nu \geq 1$ in some neighborhood of zero.

Now, we define

$$\begin{aligned}
A_n &= \{1, 2, 3, \ldots, p(n)\}, \ \forall n \geq 1, \\
r_{\sigma_n}(i, j) &= g\left(\frac{|i-j|}{\sigma_n}\right) \ \forall \ 1 \leq i, j \leq p \ \text{ and for some } \sigma_n > 0, \\
R_n &= \left((r_{\sigma_n}(i, j))\right)_{i,j\in A_n}, \ \forall n \geq 1, \\
R_{\sigma_n}(A) &= A * R_n, \ \forall n \geq 1 \ \text{ and for any matrix } A.
\end{aligned}$$

Also, g is such that R_n is positive definite. One such choice is $g(x) = e^{-|x|}$.

Theorem 3.4. *Under the conditions of Theorem 3.1, if* $\sigma_n \asymp \left(n^{-1}\log p\right)^{-\frac{1}{2(1+\gamma)}\left[\frac{\gamma}{1+\alpha}+1\right]}$ *then*

$$\|R_{\sigma_n}(\hat{\Sigma}_p) - \Sigma_p\|_2 = O_P\Big[\Big(n^{-1}\log p\Big)^{\frac{\gamma\alpha}{2(1+\alpha)(1+\gamma)}}\Big].$$

Theorem 3.5. *Under the conditions of Theorem 3.2, if* $\sigma_n \asymp \left(n^{-1}l_n\log p\right)^{-\frac{1}{2(1+\gamma)}\left[\frac{\gamma}{1+\alpha}+1\right]}$ *then*

$$\|R_{\sigma_n}(\hat{\Sigma}_p) - \Sigma_p\|_2 = O_P\Big[\Big(l_n n^{-1}\log p\Big)^{\frac{\gamma\alpha}{(1+\alpha)(1+\gamma)}}\Big].$$

Theorem 3.6. *Under the conditions of Theorem 3.3, if* $\sigma_n \asymp \left(a_n n^{-1/2}\sqrt{\log p}\right)^{-\frac{1}{2(1+\gamma)}\left[\frac{\gamma}{1+\alpha}+1\right]}$ *then*

$$\|R_{\sigma_n}(\hat{\Sigma}_p) - \Sigma_p\|_2 = O_P\Big[\Big(a_n\sqrt{n^{-1}\log p}\Big)^{\frac{\gamma\alpha}{(1+\alpha)(1+\gamma)}}\Big].$$

Remark 3.5. *Note that the rate of convergence of these estimators is weaker than that of banded estimators. As* $\gamma \to \infty$*, the function* g *tends to* 1 *faster as* $x \to 0$ *and the rates tend to the rates of the banded estimators.*

Remark 3.6. *All the rate results hold if* $\{\Lambda_{ij}\}$ *depend on* n*.*

4 Proofs

We first define some matrix norms and inequalities. For any matrix $M = ((m_{ij}))_{p\times p}$

$$\begin{aligned}
\|M\|_{(1,1)} &= \sup\{\|Mx\|_1 : \|x\|_1 = 1\} = \max_j \sum_i |m_{ij}|,\\
\|M\|_{(\infty,\infty)} &= \sup\{\|Mx\|_\infty : \|x\|_\infty = 1\} = \max_i \sum_j |m_{ij}|,\\
\|M\|_\infty &= \max_{j,i} |m_{ij}|
\end{aligned}$$

where $\|\ \|_1$ and $\|\ \|_\infty$ are usual norms in l_1 and l_∞ respectively, i.e., $\|x\|_1 = \sum_{i=1}^p |x_i|$ and $\|x\|_\infty = \max_j |x_j|$. The following inequalities hold between the above norms (see [Golub and Van Loan, 1996], pp. 83 and 84):

$$\|M\|_2 \le \left[\|M\|_{(1,1)}\|M\|_{(\infty,\infty)}\right]^{\frac{1}{2}} = \|M\|_{(1,1)} \ \text{(for symmetric } M\text{)}, \tag{4.1}$$

$$\|M\|_2 \le p\|M\|_\infty. \tag{4.2}$$

4.1 Lemmas

The next two Lemmas were needed in Remark 3.1 in Section 3.

Lemma 4.1. *Suppose* $\{x_k\}$ *is a sequence of real numbers such that* $x_k = x_{-k}$ *for all* k*,* $T_n = ((x_{i-j}))_{1\le i,j\le n}$ *for all* n *and* $\sum_{k=-\infty}^{\infty}|x_k| < \infty$*. Then*

$$\sup_n n^{-1}\operatorname{Tr}(T_n^m) \le \Big(\sum_{k=-\infty}^{\infty}|x_k|\Big)^m,\ m = 0, 1\ldots$$

Proof. For all n and m, we have

$$\begin{aligned}
\mathrm{Tr}(T_n^m) &= \left(\sum_{i,j_1,\ldots,j_{m-1}=0}^{n-1} x_{i-j_1} x_{j_1-j_2} \ldots x_{j_{m-1}-i} \right) \\
&\leq n \left(\sum_{k_1,k_2,\ldots,k_{m-1}=-(n-1)}^{n-1} x_{k_1} x_{k_2} \ldots x_{k_{m-1}} x_{(-\sum_{j=1}^{m-1} k_j)} \right) \\
&\leq n \left(\sum_{k=-\infty}^{\infty} |x_k| \right)^m .
\end{aligned}$$

This completes the proof. □

Suppose, $g : [0, 2\pi] \to \mathbb{R}$ is a square integrable function. Then the Fourier coefficients of g are defined as

$$\hat{g}(k) = (2\pi)^{-1} \int_0^{2\pi} g(x) e^{-ikx} dx, \quad k = 0, \pm 1, \ldots$$

If g is symmetric (about π), then $\hat{g}(k) = \hat{g}(-k)\ \forall k$. Let $T_{g,n}$ be the Toeplitz matrix defined by

$$T_{g,n} = ((\hat{g}\,(i-j)))_{1 \leq i,j \leq n}\,.$$

Lemma 4.2. *Suppose g is non-negative symmetric and square integrable and there exists $M > 0$ such that,*

$$\sup_n n^{-1} \mathrm{Tr}\left(T_{g,n}^k\right) \leq M^k, \ k = 1, 2, \ldots \tag{4.3}$$

Then g is almost everywhere bounded.

Proof. The proof is an application of Szegö's theorem [Grenander and Szegö, 1984]. Let X_n be a random variable such that

$$P\,(X_n = \lambda_{in}) = n^{-1}, \quad i = 1, \ldots, n$$

where $\{\lambda_{1n}, \ldots, \lambda_{nn}\}$ are all the eigenvalues of $T_{g,n}$. By Szegö's theorem $X_n \xrightarrow{\mathcal{D}} g(\mathcal{U})$ where $\mathcal{U}$ is a random variable distributed uniformly on $[0, 2\pi]$. Now from inequality (4.3) for all n,

$$EX_n^k = n^{-1} \sum_{i=1}^{n} \lambda_{in}^k = n^{-1} \mathrm{Tr}\left(T_{g,n}^k\right) \leq M^k, k = 1, \ldots \tag{4.4}$$

Now, observe that

$$x' T_{g,n} x = \sum_{k,j=1}^{n} x_k x_j \frac{1}{2\pi} \int_0^{2\pi} e^{-i(k-j)x} g(x) dx = \frac{1}{2\pi} \int_0^{2\pi} g(x) |\sum_{k=1}^{n} x_k e^{-ikx}|^2 dx \geq 0.$$

So, $T_{g,n}$ is non-negative definite, that is, X_n is non-negative. Thus (4.4) implies that $\{X_n^k\}$ is uniformly integrable for all $k = 1, 2, \ldots$. As a consequence

$$EX_n^k \to E(g(\mathcal{U}))^k, \quad k = 1, 2 \ldots$$

and using (4.4), $E(g(\mathcal{U}))^k \leq M^k, k = 1, \ldots$. From this it is immediate that g is almost everywhere bounded. □

The following Lemma is needed in the proof of Theorem3.1.

Lemma 4.3. *Suppose A is $k \times k$ positive definite matrix and I is the identity matrix. Then*

$$\int_{\mathbb{R}^k} e^{-\frac{1}{2}y'(A-2itI)y}dy = (2\pi)^{\frac{k}{2}}(\det(A-2itI))^{-\frac{1}{2}},\ t \in \mathbb{R}.$$

Proof. Let $\lambda > 0$ be the minimum eigenvalue of A. Define f and g as

$$\begin{aligned} g(z) &= (2\pi)^{\frac{k}{2}}\left[\det(A-2ZI)\right]^{-\frac{1}{2}},\ \mathcal{R}e\ z < \lambda, \\ f(z) &= \int_{-\infty}^{\infty} e^{-y'(A-2zI)y}dy,\ \mathcal{R}e\ z < \lambda. \end{aligned}$$

Note that both g and f are well defined. It is easy to check by direct integration that if $Z = x \in (-\infty, \lambda)$, then $f(x) = g(x)$. It is also easy to check that both f and g are analytic functions on $\{z : \mathcal{R}e\ z < \lambda\}$. Since they agree on $\{z : z = x \in (-\infty, \lambda)\}$, they must be identical functions. Hence $f(it) = g(it),\ t \in \mathbb{R}$ and the proof is complete. □

The following Lemma is used in the proofs of Theorems 3.2 and 3.3.

Lemma 4.4. *(See [Bhatia, 2009]) Let $\mathcal{H}$ be a Hilbert space and $B(\mathcal{H})$ be the set of all bounded liner operators on $\mathcal{H}$. Let $A, B \in \mathcal{B}(\mathcal{H})$. If A is invertible and $\|A-B\|_2 \leq \frac{1}{\|A^{-1}\|_2}$, then $\|B^{-1}-A^{-1}\|_2 \leq \frac{\|A^{-1}\|_2^2\|A-B\|_2}{1-\|A^{-1}\|_2\|A-B\|_2}$.*

The next Lemma is needed to prove Theorems 3.1 – 3.3.

Lemma 4.5. *[Saulis and Statulevičius, 1991] Suppose $E\xi = 0$, there exist $\gamma \geq 0$, $H > 0$ and $\bar{\triangle} > 0$ such that $|\Gamma_k(\xi)| \leq \left(\frac{k!}{2}\right)^{1+\gamma}\frac{H}{\bar{\triangle}^{k-2}},\ k = 2, 3, 4, \ldots,$ where $|\Gamma_k(\xi)| = |\frac{d^k}{dt^k}(\log E(e^{it\xi}))|_{t=0}$. Then for all $x \geq 0,\ P[\pm\xi \geq x] \leq e^{-\frac{x^2}{2}\left(H+x\bar{\triangle}^{\frac{-1}{2\gamma+1}}\right)^{-\frac{2\gamma+1}{\gamma+1}}}$.*
In particular, for $\xi = (\chi_n^2 - n),\ \gamma = 0, H = 4n$ and $\bar{\triangle} = \frac{1}{2}$. Hence, $P\left(|\chi_n^2 - n| \geq x\right) \leq e^{-\frac{x^2}{4(2n+x)}}$.

4.2 Proof of Theorem 3.2

First note that by (4.1), $\|B_{k_n}(\hat{\Sigma}_p) - \Sigma_p\|_2 \leq \|B_{k_n}(\hat{\Sigma}_p) - \Sigma_p\|_{(1,1)}$.
By triangle inequality, $\|B_{k_n}(\hat{\Sigma}_p) - \Sigma_p\|_{(1,1)} \leq \|B_{k_n}(\hat{\Sigma}_p) - B_{k_n}(\Sigma_p)\|_{(1,1)} + \|B_{k_n}(\Sigma_p) - \Sigma_p\|_{(1,1)}$.
Since $\sum_{\infty\times\infty} \in \mathcal{U}(\epsilon, \alpha, c)$, we have $\|B_{k_n}(\Sigma_p) - \Sigma_p\|_{(1,1)} = O_P(k_n^{-\alpha}) = O_P\left(\left(l_n n^{-1}\log p\right)^{\frac{\alpha}{\alpha+1}}\right)$.

Also, $P\Big[\|B_{k_n}(\hat{\Sigma}_p) - B_{k_n}(\Sigma_p)\|_\infty \geq t\Big] \leq \sum_{|j-k|\leq k_n} P\Big[|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}| \geq t\Big]$. We will find a uniform upper bound for these probabilities. Let,

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sigma_{jj}}},\ \rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}},\ 1 \leq j,\ k \leq p,\ 1 \leq i \leq n. \tag{4.5}$$

Now,

$$\begin{aligned}
P\Big[|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \sigma_{jk}| \geq t\Big] &= P\Big[|\sum_{i=1}^{n} z_{ij}z_{ik} - n\rho_{jk}| \geq \frac{nt}{\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&\leq P\Big[|\sum_{i=1}^{l_n} z_{ij}z_{ik} - l_n\rho_{jk}| \geq \frac{nt}{2\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&\quad + P\Big[|\sum_{i=l_n+1}^{n} z_{ij}z_{ik} - (n-l_n)\rho_{jk}| \geq \frac{nt}{2\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&\leq l_n P\Big[|z_{1j}z_{1k} - \rho_{jk}| \geq \frac{nt}{2l_n\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&\quad + P\Big[|\sum_{i=l_n+1}^{n} z_{ij}z_{ik} - (n-l_n)\rho_{jk}| \geq \frac{nt}{2\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&= l_n T_1 + T_2 \text{ (say)},
\end{aligned}$$

where

$$\begin{aligned}
T_1 &= P\Big[|\{(z_{1j}+z_{1k})^2 - 2(1+\rho_{jk})\} - \{(z_{1j}-z_{1k})^2 - 2(1-\rho_{jk})\}| \geq \frac{2nt}{l_n\sqrt{\sigma_{jj}\sigma kk}}\Big] \\
&\leq P\Big[|(z_{1j}+z_{1k})^2 - 2(1+\rho_{jk})| \geq \frac{nt}{l_n\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&\quad + P\Big[|(z_{1j}-z_{1k})^2 - 2(1-\rho_{jk})| \geq \frac{nt}{l_n\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&= P\Big[|\frac{(z_{1j}+z_{1k})^2}{2(1+\rho_{ij})} - 1| \geq \frac{nt}{2(1+\rho_{jk})l_n\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&\quad + P\Big[|\frac{(z_{1j}-z_{1k})^2}{2(1-\rho_{jk})} - 1| \geq \frac{nt}{2(1-\rho_{jk})l_n\sqrt{\sigma_{jj}\sigma_{kk}}}\Big].
\end{aligned}$$

Similarly,

$$\begin{aligned}
T_2 &\leq P\Big[|\sum_{i=l_n+1}^{n} \frac{(z_{ij}+z_{ik})^2}{2(1+\rho_{jk})} - (n-l_n)| \geq \frac{nt}{2(1+\rho_{jk})\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\
&\quad + P\Big[|\sum_{i=l_n+1}^{n} \frac{(z_{ij}-z_{ik})^2}{2(1-\rho_{jk})} - (n-l_n)| \geq \frac{nt}{2(1-\rho_{ik})\sqrt{\sigma_{jj}\sigma_{kk}}}\Big].
\end{aligned}$$

Note that

$$\begin{aligned}
(1+|\rho_{jk}|)(\sigma_{jj}\sigma_{kk})^{\frac{1}{2}} &= (\sigma_{jj}\sigma_{kk})^{\frac{1}{2}} + |\sigma_{jk}| \leq 2(\frac{\sigma_{jj}}{2} + \frac{\sigma_{kk}}{2} + \frac{|\sigma_{jk}|}{2}) \\
&= 2\frac{U'\sum_p U}{U'U} \leq 2\max_{y\neq 0}\frac{y'\sum_p y}{y'y} = 2\lambda_{\max}(\Sigma_p) \leq 2\epsilon^{-1} \qquad (4.6)
\end{aligned}$$

where $U = (0, 0, \ldots, \frac{1}{\sqrt{2}}, \ldots, a\frac{1}{\sqrt{2}}, \ldots, 0)'$ and $a = \begin{cases} 1, & \sigma_{jk} > 0 \\ -1, & \sigma_{jk} < 0 \end{cases}$.

Now, if we choose $t = Mn^{-1}l_n \log p$, then for sufficiently large n and $C_1, C_2 > 0$, using the density of χ^2_1,

$$l_n T_1 \leq 2l_n P\left[|\chi^2_1 - 1| \geq \frac{nt}{4l_n\epsilon^{-1}}\right] \leq 2l_n C_1 e^{-C_2 M \log p} \quad . \tag{4.7}$$

Let, $t = Ml_n n^{-1} \log p \to 0$, Then for some constant $C_3 > 0$, from Lemma 4.5

$$p\left[|\chi^2_{(n-l_n)} - (n - l_n)| \geq \frac{nt}{8\epsilon^{-1}}\right] \leq e^{-C_3 \frac{l_n^2}{n}(\log p)^2 M^2}. \tag{4.8}$$

Let

$$U_i^{jk} = \frac{z_{ij} + z_{ik}}{\sqrt{2(1+\rho_{jk})}},\ V_i^{jk} = \frac{z_{ij} - z_{ik}}{\sqrt{2(1-\rho_{jk})}},\ 1 \leq i \leq n,\ i \leq j,k \leq p. \tag{4.9}$$

Now,

$$(z_{l_n+1,j}, z_{l_n+1,k}, z_{l_n+2,j}, z_{l_n+2,k}, \ldots z_{n,j}, z_{n,k})' \sim N_{2(n-l_n)}(0, \triangle^{jkl_n})$$

where,

$$\triangle^{jkl_n} = \begin{pmatrix} R_{jk} * \Lambda_{l_n+1,l_n+1,jk} & R_{jk} * \Lambda_{l_n+1,l_n+2,jk} & \ldots & R_{jk} * \Lambda_{l_n+1,n,jk} \\ R_{jk} * \Lambda_{l_n+1,l_n+2,jk} & R_{jk} * \Lambda_{l_n+2,l_n+2,jk} & \ldots & R_{jk} * \Lambda_{l_n+2,n,jk} \\ \vdots & \vdots & \vdots & \vdots \\ R_{jk} * \Lambda_{l_n+1,n,jk} & R_{jk} * \Lambda_{l_n+2,n,jk} & \ldots & R_{jk} * \Lambda_{n,n,jk} \end{pmatrix}$$

$$\Lambda_{l_n+p,l_n+q,jk} = \begin{pmatrix} \Lambda_{l_n+p,l_n+q}(j,j) & \Lambda_{l_n+p,l_n+q}(j,k) \\ \Lambda_{l_n+p,l_n+q}(k,j) & \Lambda_{l_n+q,l_n+q}(k,k) \end{pmatrix}$$

for all $1 \leq p,\ q \leq n - l_n$ and $R_{jk} = \begin{pmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{pmatrix},\ 1 \leq j,k \leq p$. Hence,

$$U^{jkl_n} = (U^{jk}_{l_n+1}, U^{jk}_{l_n+2}, \ldots U^{jk}_n)' \sim N_{(n-l_n)}(0, \Gamma_+^{jkl_n}),\ 1 \leq j,\ k \leq p$$

$$V^{jkl_n} = (V^{jk}_{l_n+1}, V^{jk}_{l_n+2}, \ldots, V^{jk}_n)' \sim N_{(n-l_n)}(0, \Gamma_-^{jkl_n}),\ 1 \leq j,\ k \leq p$$

where $\Gamma_+^{jkl_n}$ and $\Gamma_-^{jkl_n}$ both are $(n - l_n)$ order matrix given by

$$\Gamma_\pm^{jkl_n}(p,q) = \begin{cases} \frac{\Lambda_{l_n+p,l_n+q}(j,j) \pm \left(\Lambda_{l_n+p,l_n+q}(j,k) + \Lambda_{l_n+p,l_n+q}(k,j)\right)\rho_{jk} + \Lambda_{l_n+p,l_n+q}(k,k)}{2(1\pm\rho_{jk})}, & p \neq q \\ 1, & p = q. \end{cases}$$

Then

$$\begin{aligned} T_2 &\leq P\left[|\left(U^{jkl_n}\right)'\left(U^{jkl_n}\right) - (n-l_n)| \geq \frac{nt}{4\epsilon^{-1}}\right] \\ &\quad + P\left[|\left(V^{jkl_n}\right)'\left(V^{jkl_n}\right) - (n-l_n)| \geq \frac{nt}{4\epsilon^{-1}}\right] \\ &\leq P\left[|\left(U^{jkl_n}\right)'\left(\Gamma_+^{jkl_n}\right)^{-1}\left(U^{jkl_n}\right) - (n-l_n)| \geq \frac{nt}{8\epsilon^{-1}}\right] \\ &\quad + P\left[|\left(V^{jkl_n}\right)'\left(\Gamma_-^{jkl_n}\right)^{-1}\left(V^{jkl_n}\right) - (n-l_n)| \geq \frac{nt}{8\epsilon^{-1}}\right] \end{aligned}$$

$$
\begin{aligned}
&+P\left[|\left(U^{jkl_n}\right)'\left(I-\left(\Gamma_+^{jkl_n}\right)^{-1}\right)\left(U^{jkl_n}\right)|\geq\frac{nt}{8\epsilon^{-1}}\right]\\
&+P\left[|\left(V^{jkl_n}\right)'\left(I-\left(\Gamma_-^{jkl_n}\right)^{-1}\right)\left(V^{jkl_n}\right)|\geq\frac{nt}{8\epsilon^{-1}}\right]\\
\leq\ &+2P\left[|\chi^2_{(n-l_n)}-(n-l_n)|\geq\frac{nt}{8\epsilon^{-1}}\right]\\
&+P\left[|\left(U^{jkl_n}\right)'\left(I-\left(\Gamma_+^{jkl_n}\right)^{-1}\right)\left(U^{jkl_n}\right)|\geq\frac{nt}{8\epsilon^{-1}}\right]\\
&+P\left[|\left(V^{jkl_n}\right)'\left(I-\left(\Gamma_-^{jkl_n}\right)^{-1}\right)\left(V^{jkl_n}\right)|\geq\frac{nt}{8\epsilon^{-1}}\right].
\end{aligned}
$$

Let, $I-(\Gamma_+^{jkl_n})^{-1}=C^{jkl_n}$ (say). Then, for some constant $C_4, C_5>0$

$$
\begin{aligned}
P\left[|\left(U^{jkl_n}\right)'\left(C^{jkl_n}\right)\left(U^{jkl_n}\right)|\geq\frac{nt}{8\epsilon^{-1}}\right] &\leq P\left[\max_{U\neq 0}\frac{|\left(U'C^{jkl_n}U\right)|}{U'U}\left(U^{jkl_n}\right)'\left(U^{jkl_n}\right)\geq\frac{nt}{8\epsilon^{-1}}\right]\\
&= P\left[\sqrt{\lambda_{\max}\left(C^{jkl_n}\right)'\left(C^{jkl_n}\right)}\left(U^{jkl_n}\right)'\left(U^{jkl_n}\right)\geq\frac{nt}{8\epsilon^{-1}}\right]\\
&= P\left[||C^{jkl_n}||_2\left(U^{jkl_n}\right)'\left(U^{jkl_n}\right)\geq\frac{nt}{8\epsilon^{-1}}\right]\\
&\leq nP\left[||C^{jkl_n}||_2\,\chi_1^2\geq\frac{t}{8\epsilon^{-1}}\right]\\
&\leq nC_4e^{-C_5\frac{t}{||C^{jkl_n}||_2}}.
\end{aligned}
$$

Moreover, by (4.2), $||A^{jkl_n}||_2\leq n||A^{jkl_n}||_\infty\leq nS'(l_n)$. Since $\nabla_{\infty\times\infty}\in\mathcal{L}(l_n,\ n\geq 1)$, we have $||A^{jkl_n}||_2=o(1)$ and $||A^{jkl_n}||_2\leq 1$ for sufficiently large n and for some constant $C_6>0$, $||C^{jkl_n}||_2\leq C_6||A^{jkl_n}||_2\leq nC_6S'(l_n)$. Hence, putting $t=Ml_nn^{-1}\log p$, for some constant $C_7,\ C_8>0$,

$$
P\left[|(U^{jkl_n})'(I-(\Gamma_+^{jkl_n})^{-1})(U^{jkl_n})|\geq\frac{nt}{8\epsilon^{-1}}\right]\leq nC_4e^{-C_7\frac{t}{nS'(l_n)}}\leq nC_4e^{-C_8M\log p}. \tag{4.10}
$$

Similar bound holds for V^{jkl_n}. From (4.7) to (4.10), for sufficiently large n,

$$
\begin{aligned}
P\left[|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}-\sigma_{jk}|\geq Mn^{-1}l_n\log p\right] &\leq 2l_nC_1e^{-C_2M\log p}+2e^{-C_3M^2\frac{l_n^2}{n}(\log p)^2}+2nC_4e^{-C_8M(\log p)}\\
&= 2l_nC_1p^{-C_2M}+2e^{-C_3M^2\frac{l_n^2}{n}(\log p)^2}+2C_4np^{-C_8M}.
\end{aligned}
$$

Now, for some constant $C>0$,

$$
P\left[||B_{k_n}(\hat{\Sigma}_p)-B_{k_n}(\Sigma_p)||_\infty\geq Mn^{-1}l_n\log p\right]\leq C\left[p^{3-C_2M}+p^2e^{-C_3\frac{M^2l_n^2}{n}(\log p)^2}+p^{3-C_8M}\right].
$$

If $M>\frac{3}{C_2}\vee\frac{3}{C_8}$, then $p^{3-C_2M}+p^{3-C_8M}\to 0$. The logarithm of the 2nd term is

$$
2\log p-C_3M^2n^{-1}l_n^2(\log p)^2=\log p\left[2-C_3M^2n^{-1}l_n^2(\log p)\right].
$$

Now if $\liminf l_n^2n^{-1}\log p>0$ then it is bounded away from zero by S (say).
So, if $M>\frac{3}{c_2}\vee\frac{3}{C_8}\vee\sqrt{\frac{2}{C_3S}}$, then the 2nd term also tends to zero. This completes the proof. □

4.3 Proof of Theorem 3.3

As before for some constant $C > 0$ we have

$$\begin{aligned} \|B_{k_n}(\hat{\Sigma}_p) - \Sigma_p\|_2 &\leq \|B_{k_n}(\Sigma_p) - \Sigma_p\|_{(1,1)} + Ck_n\|B_{k_n}(\hat{\Sigma}_p) - B_{k_n}(\Sigma_p)\|_\infty, \quad (4.11) \\ \|B_{k_n}(\Sigma_p) - \Sigma_p\|_{(1,1)} &= O_P\left(k_n^{-\alpha}\right) \quad (4.12) \\ P\Big[\|B_{k_n}\hat{\Sigma}_p) - B_{k_n}(\Sigma_p)\|_\infty \geq t\Big] &\leq \sum_{|j-k|\leq k_n} P\Big[|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}| \geq t\Big]. \quad (4.13) \end{aligned}$$

and we will again find a uniform upper bound for these probabilities. Let,

$$\begin{aligned} A_{r,a_n} &= \{i \in \mathbb{Z}^+ \cup \{0\} : (ia_n + r) \leq n\}, \quad (4.14) \\ C_{r,a_n} &= \text{cardinality of } A_{r,a_n},\ 1 \leq r \leq a_n \quad (4.15) \end{aligned}$$

and $z_{ij},\ \rho_{jk}$ are as in (4.5). Now,

$$\begin{aligned} P\Big[|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}| \geq t\Big] &\leq \sum_{r=1}^{a_n} P\Big[|\sum_{i\in A_{r,a_n}} \{z_{(ia_n+r)j}z_{(ia_n+r)k} - \rho_{jk}\}| \geq \frac{nt}{a_n\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\ &\leq \sum_{r=1}^{a_n} P\Big[|\sum_{i\in A_{r,a_n}} \frac{(z_{(ia_n+r)j} + z_{(ia_n+r)k})^2}{2(1+\rho_{jk})} - C_{r,a_n}| \geq \frac{nt}{a_n(1+\rho_{jk})\sqrt{\sigma_{jj}\sigma_{kk}}}\Big] \\ &\quad + \sum_{r=1}^{a_n} P\Big[|\sum_{i\in A_{r,a_n}} \frac{(z_{(ia_n+r)j} - z_{(ia_n+r)k})^2}{2(1-\rho_{jk})} - C_{r,a_n}| \geq \frac{nt}{a_n(1-\rho_{jk})\sqrt{\sigma_{jj}\sigma_{kk}}}\Big]. \end{aligned}$$

Let,

$$\begin{aligned} U_i^{jkra_n} &= \frac{z_{(ia_n+r)j} + z_{(ia_n+r)k}}{\sqrt{2(1+\rho_{jk})}}, \quad V_i^{jkra_n} = \frac{z_{(ia_n+r)j} - z_{(ia_n+r)k}}{\sqrt{2(1-\rho_{jk})}},\ 1 \leq j,k \leq p,\ i \in A_{r,a_n},\ 1 \leq r \leq a_n \\ U^{jkra_n} &= Vec\left(U_i^{jkra_n} : i \in A_{r,a_n}\right), V^{jkra_n} = Vec\left(V_i^{jkra_n} : i \in A_{r,a_n}\right),\ 1 \leq j,k \leq p,\ 1 \leq r \leq a_n. \end{aligned}$$

Now, applying (4.6) we have,

$$\begin{aligned} P\Big[|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}| \geq t\Big] &\leq \sum_{r=1}^{a_n} P\Big[|(U^{jkra_n})'(U^{jkra_n}) - C_{r,a_n}| \geq \frac{nt}{2a_n\epsilon^{-1}}\Big] \\ &\quad + \sum_{r=1}^{a_n} P\Big[|(V^{jkra_n})'(V^{jkra_n}) - C_{r,a_n}| \geq \frac{nt}{2a_n\epsilon^{-1}}\Big]. \end{aligned}$$

Also, we have $Vec\left((z_{(ia_n+r)j}, z_{(ia_n+r)k})_{i\in A_{r,a_n}}\right) \sim \mathcal{N}_{2C_{r,a_n}}(0, \triangle^{jkra_n})\ \ \forall 1 \leq r \leq a_n, 1 \leq j,k \leq p$, where

$$\triangle^{jkra_n} = \begin{bmatrix} \Lambda_{0.jk} * R_{jk} & \Lambda_{a_n.jk} * R_{jk} & \Lambda_{2a_n.jk} * R_{jk} & \dots & \Lambda_{(C_{r,a_n}-1)a_n.jk} * R_{jk} \\ \Lambda_{a_n.jk} * R_{jk} & \Lambda_{0.jk} * R_{jk} & \Lambda_{a_n.jk} * R_{jk} & \dots & \Lambda_{(C_{r,a_n}-2)a_n.jk} * R_{jk} \\ \Lambda_{2a_n.jk} * R_{jk} & \Lambda_{a_n.jk} * R_{jk} & \Lambda_{0.jk} * R_{jk} & \dots & \Lambda_{(C_{r,a_n}-3)a_n.jk} * R_{jk} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Lambda_{(C_{r,a_n}-1)a_n.jk} * R_{jk} & \Lambda_{(C_{r,a_n}-2)a_n.jk} * R_{jk} & \Lambda_{(C_{r,a_n}-3)a_n.jk} * R_{jk} & \dots & \Lambda_{0.jk} * R_{jk} \end{bmatrix},$$

$$\Lambda_{ba_n.jk} = \begin{bmatrix} \Lambda_{ba_n}(j,j) & \Lambda_{ba_n}(j,k) \\ \Lambda_{ba_n}(k,j) & \Lambda_{ba_n}(k,k) \end{bmatrix} \qquad and \qquad R_{jk} = \begin{bmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{bmatrix}.$$

Let,

$$\Gamma_{\pm}^{jkra_n}(p,q) = \begin{cases} \frac{\Lambda_{|p-q|a_n}(j,j)+\left(\Lambda_{|p-q|a_n}(j,k)\pm\Lambda_{|p-q|a_n}(k,j)\right)\rho_{jk}+\Lambda_{|p-q|a_n}(k,k)}{2\left(1\pm\rho_{jk}\right)} & , p \neq q \\ 1, \quad p = q \end{cases}$$

$$\Gamma_{\pm}^{jkra_n} = ((\,\Gamma_{\pm}^{jkra_n}(p,q)\,)), \;\; C_{\pm}^{jkra_n} = I - (\Gamma_{\pm}^{jkra_n})^{-1},$$

Hence, $U^{jkra_n} \sim \mathcal{N}_{C_{r,a_n}}(0, \Gamma_{+}^{jkra_n})$, $V^{jkra_n} \sim \mathcal{N}_{C_{r,a_n}}(0, \Gamma_{-}^{jkra_n})$. Therefore,

$$\begin{aligned} P\Big[|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \sigma_{jk}| \geq t\Big] \;\leq\; & 2a_n P\Big[|\chi^2_{C_{r,a_n}} - C_{r,a_n}| \geq \frac{nt}{4a_n\epsilon^{-1}}\Big] \\ & + \sum_{r=1}^{a_n} P\Big[|U^{jkra_n})' C_{+}^{jkra_n}(U^{jkra_n})| \geq \frac{nt}{4a_n\epsilon^{-1}}\Big] \\ & + \sum_{r=1}^{a_n} P\Big[|(V^{jkra_n})' C_{-}^{jkra_n}(V^{jkra_n})| \geq \frac{nt}{4a_n\epsilon^{-1}}\Big]. \end{aligned} \tag{4.16}$$

Again, for $t = Ma_n(n^{-1}\log p)^{\frac{1}{2}}$ and some $C_1 > 0$, as $n^{-1}\log p \to 0$ we have

$$P\Big[|\chi^2_{C_{r,a_n}} - C_{r,a_n}| \geq \frac{nt}{4a_n\epsilon^{-1}}\Big] \leq e^{-C_1 M \log p} \tag{4.17}$$

Now, as before, for some $C_2, C_2' > 0$,

$$P\Big[|(U^{jkra_n})' C_{+}^{jkra_n}(U^{jkra_n})| \geq \frac{nt}{4a_n\epsilon^{-1}}\Big] \leq nC_2 \exp\{-C_2' t a_n^{-1} \|C_{+}^{jka_n}\|_2^{-1}\}.$$

Since $\nabla_{\infty\times\infty} \in \mathcal{A}(a_n, n \geq 1)$, we have $\|C_{+}^{jkra_n}\|_2 \leq 1$ and hence $\|C_{+}^{jkra_n}\|_2 \leq C_3 n S(a_n)$ for some $C_3 > 0$. Therefore, putting $t = Ma_n(n^{-1}\log p)^{\frac{1}{2}}$, we have, for some constant $C_4 > 0$,

$$P[|(U^{jkra_n})' C_{+}^{jkra_n}(U^{jkra_n})| \geq (4a_n)^{-1} nt\epsilon] \leq nC_2\, exp\{-C_4 M {a_n}^{-1}\sqrt{n\log p}\}.$$

Similarly, for some constant $C_2, C_4 > 0$,

$$P[|(V^{jkra_n})' C_{-}^{jkra_n}(V^{jkra_n})| \geq (4a_n)^{-1} nt\epsilon] \leq nC_2\, exp\{-C_4 M {a_n}^{-1}\sqrt{n\log p}\}.$$

Hence, by (4.16) and (4.17) we have

$$P\Big[|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \sigma_{jk}| \geq t\Big] \;\leq\; 2a_n e^{-C_1 M\log p} + 2na_n C_2 e^{-C_4 M\frac{\sqrt{n\log p}}{a_n}}.$$

Therefore, for some constant $C_5 > 0$

$$P\Big[\|B_{k_n}(\hat{\Sigma}_p) - B_{k_n}(\Sigma_p)\|_\infty \geq t\Big] \leq C_5[p^{3-C_1 M} + p^4 C_2\, \mathrm{e}^{-C_4 M\frac{\sqrt{n\log p}}{a_n}}].$$

Clearly, the first term $\to 0$ as $n \to \infty$ if $M > \frac{3}{C_1}$. Now, since $a_n\sqrt{n^{-1}\log p} \to 0$ and $a_n^{-1}\sqrt{n\log p} \to \infty$, we have, for some constant $C_6, C_7 > 0$,

$$p^4 e^{-C_4 M\frac{\sqrt{n\log p}}{a_n}} \;=\; e^{C_6\frac{\sqrt{n\log p}}{a_n}\left(a_n\sqrt{\frac{\log p}{n}} - C_7 M\right)} \to 0.$$

Hence, $\|B_{k_n}(\hat{\Sigma}_p) - B_{k_n}(\Sigma_p t)\|_\infty = O_P(a_n\sqrt{n^{-1}\log p})$. □

4.4 Proof of Theorem 3.1

As before we have (4.11), (4.12) and (4.13). Let,

$$U = \left(U_1^{jk}, U_2^{jk}, \ldots, U_n^{jk}\right)', \quad V = \left(V_1^{jk}, V_2^{jk}, \ldots, V_n^{jk}\right)'$$

where U_i^{jk} and V_i^{jk} are as in (4.9). Then, we have

$$P\left[|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \sigma_{jk}| \geq t\right] \leq P\left[|U'U - n| \geq \frac{nt}{2\epsilon^{-1}}\right] + P\left[|V'V - n| \geq \frac{nt}{2\epsilon^{-1}}\right].$$

Now, using Lemma 4.3 we have $E(e^{itU'U}) = \left[\det(I - 2it\Gamma_+^{jk})\right]^{-\frac{1}{2}}$. Hence,

$$\frac{d^r}{dt^r}\log E(e^{itU'U}) = -\frac{1}{2}\sum_{p=1}^{n}\frac{d^r}{dt^r}\log(1 - 2it\lambda_p^{jk})$$

where $\lambda_p^{jk},\ 1 \leq p \leq n$, are eigenvalues of Γ_+^{jk}. So, we have $|\Gamma_k(U'U - n)| = \frac{1}{2}\sum_{p=1}^{n}(r-1)\,2^r\left(\lambda_p^{jk}\right)^r$. Now, from (3.1), we have $|\Gamma_k(U'U)| \leq \left(\frac{r!}{2}\right)\frac{4nM^2}{\left(\frac{1}{2M}\right)^{r-2}}$. Hence, using Lemma 4.5, for some $C_1 > 0$ and $t = M'(n^{-1}\log p)^{\frac{1}{2}}$, $P\left[|U'U - n| \geq \frac{nt}{2\epsilon^{-1}}\right] \leq e^{-C_1 M'^2 \log p}$. Hence, for some $C_2,\ C_3 > 0$,

$$P\left[||B_{k_n}\left(\hat{\Sigma}_p\right) - B_{k_n}\left(\Sigma_p\right)||_\infty \geq t\right] \leq C_2 p^2 e^{C_3 M'^2 \log p} \to 0$$

for $M' > \sqrt{\frac{2}{C_3}}$. This completes the proof. □

4.5 Proof of Theorems 3.4– 3.6

By (4.1) and triangular inequality.

$$||R_{\sigma_n}\left(\hat{\Sigma}_p\right) - \Sigma_p||_2 \leq ||R_{\sigma_n}\left(\hat{\Sigma}_p\right) - R_{\sigma_n}\left(\Sigma_p\right)||_{(1,1)} + ||R_{\sigma_n}\left(\Sigma_p\right) - \Sigma_p||_{(1,1)}.$$

Now, for some constant $C_1 > 0$,

$$||R_{\sigma_n}\left(\hat{\Sigma}_p\right) - R_{\sigma_n}\left(\Sigma_p\right)||_{(1,1)} \leq ||\hat{\Sigma}_p - \Sigma_p||_\infty\left(2\sum_{l=0}^{p} g\left(\frac{l}{\sigma_n}\right)\right) \leq \sigma_n||\hat{\Sigma}_p - \Sigma_p||_\infty\left[C_1\int_0^\infty g(x)dx\right].$$

As before we have $||\hat{\Sigma}_p - \Sigma_p||_\infty = \begin{cases} O_P\left(n^{-1/2}\sqrt{\log p}\right), & \text{in Theorem 3.4} \\ O_P\left(l_n n^{-1}\log p\right), & \text{in Theorem 3.5} \\ O_P\left(a_n n^{-1/2}\sqrt{logp}\right). & \text{in Theorem 3.6.} \end{cases}$

Again, by triangular inequality

$$\begin{aligned} ||R_{\sigma_n}(\Sigma_p) - \Sigma_p||_{(1,1)} \quad \leq \quad & ||R_{\sigma_n}(\Sigma_p) - B_{k_n}[R_{\sigma_n}(\Sigma_p)]||_{(1,1)} + ||B_{k_n}[R_{\sigma_n}(\Sigma_p)] - B_{k_n}(\Sigma_p)||_{(1,1)} \\ & + ||B_{k_n}(\Sigma_p) - \Sigma_p||_{(1,1)} \end{aligned}$$

and we have $\|B_{k_n}(\Sigma_p) - \Sigma_p\|_{(1,1)} = O(k_n^{-\alpha})$ and $\|R_{\sigma_n}(\Sigma_p) - B_{k_n}[R_{\sigma_n}(\Sigma_p)]\|_{(1,1)} = O(k_n^{-\alpha})$. Now, for some constant $C_2, C_3 > 0$, as σ_{ij} 's are bounded, for sufficiently large n,

$$\begin{aligned}\|B_{k_n}[R_{\sigma_n}(\Sigma_p)] - B_{k_n}(\Sigma_p)\|_{(1,1)} &\leq \max_i \sum_{j:|i-j|\leq k_n} \left(1 - g\left(\frac{|i-j|}{\sigma_n}\right)\right)|\sigma_{ij}| \\ &\leq C_2 \sum_{l=-k_n}^{k_n} \left(1 - g\left(\frac{l}{\sigma_n}\right)\right) \leq C_3 \left(\frac{k_n}{\sigma_n}\right)^{\gamma} k_n.\end{aligned}$$

Now, consider $k_n = \begin{cases} \left(n^{-1}\log p\right)^{-\frac{\gamma}{2(1+\gamma)(1+\alpha)}}, & \text{for Theorem 3.4} \\ \left(l_n n^{-1}\log p\right)^{-\frac{\gamma}{(1+\gamma)(1+\alpha)}}, & \text{for Theorem 3.5} \\ \left(a_n n^{-1/2}\sqrt{\log p}\right)^{-\frac{\gamma}{(1+\gamma)(1+\alpha)}}, & \text{for Theorem 3.6 .} \end{cases}$

This completes the proof. □

References

G. I. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Stat.*, 4(2):764–790, 2010.

J. Bennett and S. Lanning. The netflix prize. *Proceedings of KDD Cup and Workshop 2007, San Jose, California, USA*, 2007.

R. Bhatia. *Notes on Functional Analysis*. Hindustan Book Agency, India, 2009.

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008a.

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6): 2577–2604, 2008b. doi: 10.1214/08-AOS600.

T. T. Cai, C. H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumor using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002. doi: 10.1198/016214502753479248.

B. Efron. Are a set of microarrays independent of each other? *Ann. Appl. Stat.*, 3(3):922–942, 2009.

G. H. Golub and C. F. Van Loan. *Matrix Computations, 3rd ed.* The Johns Hopkins University Press, Baltimore, 1996.

U. Grenander and G. Szegö. *Toeplitz forms and their applications*. Chelsea Publishing Co., New York, second edition, 1984.

L. Klebanov and A. Yakovlev. Diverse correlation structures in gene expression data and their utility in improving statistical inference. *Ann. Appl. Stat.*, 1(2):538–559, 2007.

J. Leek and J. Storey. General framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008. doi: 10.1073/pnas.0808709105.

A. B. Owen. Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(3):411–426, 2005.

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

L. Saulis and V. A. Statulevičius. *Limit Theorems for Large Deviations*. Kluwer Academic Publishers, Dordrecht, 1991.