Statistica Sinica: Supplement

BANDWIDTH SELECTION FOR LARGE COVARIANCE AND PRECISION MATRICES

Xuehu Zhu[1], Jian Guo[1,2], Xu Guo[3], Lixing Zhu*[3,4] and Jiasen Zheng[5]

[1] *School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China*

[2]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

[3] *Center for Statistics and Data Science, Beijing Normal University, Zhuhai, China*

[4]*Department of Mathematics, Hong Kong Baptist University, Hong Kong*

[5]*Center for Statistical Science, Tsinghua University, Beijing, China*

Supplementary Material

S1. Estimation of the Covariance matrix and its Inverse

In this section, we will discuss how to apply the estimation of the bandwidth to the estimation of the covariance matrix and the precision matrix and give the properties of the corresponding estimators.

S1.1 Regularized Covariance Matrix

One prevalent class of methodology to estimate the covariance matrix is the truncated regularization which needs to learn the banding and tapering structure of the sample covariance matrix.

S1.1.1 Banding the Sample Covariance Matrix

For a $p \times p$ matrix $M = (m_{l_1 l_2})_{p\times p}$, let $B_k(M) = (m_{l_1 l_2} \mathrm{I}\{l_1 - l_2 \mid \leq k\})_{p\times p}$ denote the banded version with bandwidth $k \in \{0, \cdots, p-1\}$. Bickel and Levina [2008] suggested $B_{\hat{K}}(S_n)$ to estimate Σ, where S_n is the sample covariance and $\hat{K}$ is an estimate of the bandwidth. When an estimate of the bandwidth is obtained, it is natural to use the banding sample covariance matrix to estimate it as well.

Consider the covariance matrix Σ belonging to the following parameter space:

$$\begin{gathered}\mathcal{U}_B(\varepsilon_0, K) = \{\Sigma : \sigma_{ij} = 0 \text{ for all } |i-j| > K \\ \text{and } \ 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0\},\end{gathered} \tag{S1.1}$$

where $\lambda_{\max}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote the maximum and minimum eigenvalues of the matrix Σ, respectively. Combining the banding estimate obtained in

Section 2.1, we give the estimate of the covariance matrix as:

$$\hat{\Sigma}_{B,\hat{K}} = B_{\hat{K}}(S_n) \text{ with } S_n = (\sigma^*_{ij})_{1\le i,j\le p} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top. \quad \text{(S1.2)}$$

The following theorem states the convergence rate of the banding covariance estimate in (S1.2).

Theorem S1.1. *Suppose that* $\mathbf{X_i}$ *are Gaussian and* $\mathcal{U}_B\left(\varepsilon_0, K\right)$ *is the class of covariance matrices defined above. If Assumption S2.1 holds, then we have that for any* $\gamma > 0$*,*

$$\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| = O_p\left(\frac{K}{\log(p\vee n)^{5\gamma/4-1/2}n^{1/2-\gamma/4}}\right)$$

uniformly on $\Sigma \in \mathcal{U}_B$*.*

Remark S1.1. This theorem shows that if $K = o\left(\log(p\vee n)^{5/4\gamma-1/2}n^{1/2-\gamma/4}\right)$, then $\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| = o_p(1)$ holds uniformly for $\Sigma \in \mathcal{U}_B$. This result gives the condition of the true bandwidth for the convergence of the banding estimator. Also, it shows that our proposed method can deal with the case where the band size K goes to infinity as the sample size goes to infinity.

S1.1.2 Tapering the Sample Covariance Matrix

To obtain an optimal estimation of the covariance matrix, Bickel and Levina [2008] and Cai et al. [2010] discussed general tapering of the covariance matrix. Cai et al. [2010] defined a tapering estimator as

$$\hat{\Sigma}_{T,K} = \left(\omega_{ij}\sigma^*_{ij}\right)_{p\times p}, \tag{S1.3}$$

where σ^*_{ij} are the elements of the estimator S_n defined in (S1.2) and the weights are

$$\omega_{ij} = \begin{cases} 1, & \text{when } |i-j| \leq K_h, \\ 2 - \frac{|i-j|}{K_h}, & \text{when } K_h < |i-j| < K, \\ 0, & \text{otherwise,} \end{cases}$$

and $K_h = K/2$ with K being even.

If the tapering order is given, the proposed method can be extended to handle the determination in this tapering structure. Recall that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent and identically distributed random vectors with mean 0 and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$, $h(k) = \frac{1}{p-k}\sum_{l=1}^{p-k}\sigma^2_{ll+k}$, $0 \leq k \leq p-1$. We assume that Σ has the structure that $\sigma_{ij} = 0$ when $|i-j| > K$, and $h(k)$ decreases as k varying from $K/2$ to K.

In this case, we still have $h(k) = 0$ as $k > K$, and there is a skip when $k = K$. Then the same as the previous analysis of the banding estimation, we could apply a similar method to the tapering estimation. The tapering estimator-based ratios can be similarly defined as:

$$\hat{s}(k) = \frac{\hat{h}(k+1) + c_n}{\hat{h}(k) + c_n}, \quad \text{for } 0 \leq k \leq p, \tag{S1.4}$$

where c_n is recommended as $c_n = \delta \log(p \cdot n) \max\limits_{M_1 \leq k \leq M_2} |\hat{h}(k)|$. Then the bandwidth of the tapering estimator can be estimated as:

$$\hat{K} = \mathop{\arg\max}_{0 \leq k \leq p-1} \{k : \hat{s}(k) \leq \tau\}.$$

Under similar conditions as those in Theorem 2.2 of the main body, we can also obtain the consistency of the estimator $\hat{K}$, i.e. $P(\hat{K} = K) \to 1$, as $n, p \to \infty$. It seems to have the same estimators and theoretical results for the banding and tapering structures. However, the bandwidth estimation yields different covariance matrix estimates in the banding and tapering cases. That is, the covariance matrix is estimated respectively as $\hat{\Sigma}_{B,\hat{K}}$ in (S1.2) and $\hat{\Sigma}_{T,\hat{K}}$ in (S1.3). In the following we will discuss the properties of the tapering estimator.

First, define a covariance class

$$\mathcal{U}_T\left(M_0, K\right)=\left\{\Sigma: \sigma_{ij}=0 \text{ for all } |i-j|>K \text{ and } \lambda_{\max}(\Sigma) \leq M_0\right\}, \quad \text{(S1.5)}$$

with $M_0>0$. Note that the minimum eigenvalue of any covariance matrix in the parameter space $\mathcal{U}_T\left(M_0, K\right)$ is allowed to be 0, which is more general than the class (S1.1). Next we give the risk upper bound of the defined tapering estimator at the operator norm.

Theorem S1.2. *Under* ***Assumption*** *S2.2, suppose the covariance matrix Σ satisfies (S1.5). Then, the tapering estimator $\hat{\Sigma}_{T,\hat{K}}$ defined in (S1.3) satisfies that for $\alpha>0$,*

$$\|\hat{\Sigma}_{T,\hat{K}}-\Sigma\|=O_p\left(\frac{(K+\log p)^{1/2}}{\log(p \vee n)^{5\alpha/4} n^{1/2-\alpha/4}}\right)$$

uniformly on $\Sigma \in \mathcal{U}_T$. In particular, if $K=o\left(\log(p \vee n)^{5\alpha/2} n^{1-\alpha/2}\right)$, the estimator $\hat{\Sigma}_{T,\hat{K}}$ satisfies $\|\hat{\Sigma}_{T,\hat{K}}-\Sigma\|=o_p(1)$.

Remark S1.2. This tapering structure is a special case of the general tapering proposed in Bickel and Levina [2008] that replaces $\hat{\Sigma}=(\hat{\sigma}_{ij})$ with $\hat{\Sigma} * R$, where $*$ denotes Schur matrix multiplication and R is a positive definite matrix defined by $R=[g(\frac{\rho(i,j)}{\sigma})]$, $\sigma>0$. Here, $\rho(i,j)$, satisfying

$\rho(i,j) \geq 0$ and $\rho(i,i) = 0$ for all i, is a function used to characterize the distance of the (i,j) item of a matrix from the diagonal. A simple example is $\rho(i,j) = |i-j|$. g is a mapping of nonnegative real number to positive real number contenting that $g(0) = 1$ and $g(t)$ is decreasing to 0 as $t \to \infty$. Thus a major change compared to the banding hypothesis is that under the tapering hypothesis, $h(k)$ goes to zero as k tends to infinity. At the population level, it has much slower order than that under the banding hypothesis. However, since we know little about the order of the pending data, we still assume $h(k) = \frac{1}{p-k}\sum_{l=1}^{p-k}\sigma_{ll+k}^2$ has the same order as that under banding hypothesis. For the above tapering hypothesis, we can also achieve the similar bandwidth estimate. Thus, the algorithm to estimate K will be adapted the same way with the banding case. But for the estimation of the covariance matrix Σ, we also substitute $\hat{\Sigma}$ by $\hat{\Sigma} * R$ with R being related to our simplification of the tapering hypothesis. While, under this assumption, the performance for the tapering estimator relies heavily on the data. A more precise discussion of the tapering estimator may be set up in more general theory, the study is ongoing.

S1.1.3 Partial Numerical Studies

In this subsection, we assess the performance of the truncated sample covariance matrix. We estimate the bandwidth for the banding structure via the proposed methods in the main body of this paper. The tapering structure of the covariance matrix estimation often addresses the situation where the bandwidth of the covariance matrix is large, especially for the case that the true value of the covariance matrix is rather small at the truncation, i.e. $h(K) \ll 1$. To obtain a larger bandwidth for tapering structure of the sample covariance matrix, we set the smaller ridge term c_n to be $c_n = \log(p \cdot n) \max_{M_1 \leq k \leq M_2} |\hat{h}(k)|/20$. Then the covariance matrices with banding and tapering structure can be estimated by (S1.2) and (S1.3), respectively. Calculate the errors between these estimates and the true covariance matrix. Each experiment is repeated 100 times.

The data are generated from

$$\mathbf{X}_i = \Sigma^{1/2}\mathbf{Z}_i, \text{ with } \mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top,$$

where Z_{ij} are i.i.d. respectively from $N(0,1)$. Consider the truncated covariance matrix $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$ as follows.

Example 5: $\sigma_{ij} = I(i=j)+\rho|i-j|^{-(\alpha+1)}I(0 < |i-j| \leq K)$, where $\rho = 0.6$, $\alpha = 0.2$, and $K = 4, 36$, which correspond to the banding and tapering structures respectively. We design the sample size to be $n = 200, 250$ and the dimension to be $p = 60, 100, 200, 300$. The results are reported in Table 1.

Table 1: Mean and standard deviation of the estimated bandwidth, error of banding structure estimator and error of tapering structure estimator in **Example 5**.

		Covariance structure with $\rho = 0.6,\ \alpha = 0.2$			
		$K = 4$		$K = 36$	
n	p	$\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\|$	$\|\hat{\Sigma}_{T,\hat{K}} - \Sigma\|$	$\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\|$	$\|\hat{\Sigma}_{T,\hat{K}} - \Sigma\|$
200	60	0.716(0.116)	1.389(0.145)	1.780(0.077)	1.515(0.187)
200	100	0.784(0.103)	1.700(0.148)	2.475(0.045)	1.673(0.250)
200	200	0.808(0.118)	1.695(0.127)	1.817(0.058)	1.640(0.178)
200	300	0.823(0.098)	1.887(0.118)	2.619(0.027)	1.947(0.225)
250	60	0.6017(0.118)	1.133(0.115)	1.508(0.112)	1.223(0.125)
250	100	0.6544(0.104)	1.221(0.141)	1.663(0.082)	1.489(0.143)
250	200	0.7317(0.107)	1.621(0.131)	1.622(0.078)	1.594(0.131)
250	300	0.7703(0.127)	1.408(0.132)	2.134(0.028)	1.571(0.189)

The mean and standard deviations of the errors between the banding (tapering) structure estimators and the true covariance matrix are also displayed in Table 1. The results show when the true bandwidth is small, the estimator of covariance matrices with banding structure performs better than that with tapering structure. But when the true bandwidth becomes large, the covariance matric estimate with tapering structure performs better than that with banding structure.

S1.2 Regularized Precision Matrix

We borrow the method of the Cholesky decomposition from Bickel and Levina [2008] to estimate the precision matrix. Recall that Σ can be decomposed via Cholesky decomposition as:

$$\Sigma = LDL^\top,$$

where L is a lower triangular matrix and D is a diagonal matrix. Let $T = L^{-1} = (t_{ij})_{1 \le i,j \le p}$, then the precision matrix $\Omega = \Sigma^{-1}$ can be written as

$$\Omega = T^\top D^{-1} T.$$

Using the estimation of Section 3 in the main body,

$$\hat{t}_j^{(k)} = -(\chi_j^{(k)^\top} \chi_j^{(k)})^{-1} \chi_j^{(k)^\top} \chi_j, \tag{S1.6}$$

we can obtain the estimates $\hat{T}$ and $\hat{D}$ of the matrixes T and D respectively. Combining again the bandwidth estimates obtained in Section 3 of the main body, we define the following banding precision matrix estimate as:

$$\hat{\Omega}_{\hat{K}} = B_{\hat{K}}(\hat{T})^\top \hat{D}^{-1} B_{\hat{K}}(\hat{T})$$

with the estimated bandwidth $\hat{K}$. Note that this estimate is not the same as $B_{\hat{K}}(S_n)^{-1}$, which is always not well-defined at $p > n$.

Define the following class:

$$\begin{aligned} \mathcal{U}_B^{-1}(\varepsilon_0, K) = \{ & \Sigma : t_{ij} = 0 \text{ for all } |i-j| > K \\ & \text{and } 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0 \}. \end{aligned} \tag{S1.7}$$

The following result establishes the convergence rate of the banding precision estimates.

Theorem S1.3. *Uniformly for $\Sigma \in \mathcal{U}^{-1}(\varepsilon_0, K)$, if $\mathbf{X}_i$ are Gausssian and $n^{-1}\log p = o(1)$, we have for any $\beta > 0$,*

$$\|\hat{\Omega}_{\hat{K}} - \Omega\| = O_p\left(K^2 \frac{(\log p)^{2-\beta/4}}{n^{1-\beta/4}}\right).$$

In particular, if $K^2 = o\left(\frac{n^{1-\beta/4}}{(\log p)^{2-\beta/4}}\right)$, then $\|\hat{\Omega}_{\hat{K}} - \Omega\| = o_p(1)$.

S2. Proof of main results

We first recall the assumptions:

Assumption S2.1. $\log p = o(n^{1/5})$, as $\min\{n, p\} \to \infty$.

Assumption S2.2. Assume Σ is a positive definite matrix, let $\mathbf{Z}_k = \Sigma^{-1/2}\mathbf{X}_k$. Variables $X_{il},\ 1 \leq l \leq p$ and $\mathbf{Z}_k$'s are sub-Gaussian vectors with

$\sup_{1\leq l\leq p} \|X_{kl}\|_{\psi_2} < K_0$ and $E(\exp(\alpha^\top \mathbf{Z}_k)) \leq \exp(K_z^2\|\alpha\|^2)$ for some constants $0 < K_0,\ K_z < \infty$.

Then several lemmas are presented and the proofs of the theorems are given.

Lemma S2.1. *Under Assumption S2.2, we have for any* $1 \leq i \leq n$

$$\sup_{1\leq l\leq p} P(|X_{il}| > t) \leq C_1 \exp\left(-\frac{t^2}{K_1^2}\right),$$

for some constant $C_1 > 0$ *and* K_1 *only depends on* K_0*.*

Proof. Recall that X_{il} $1 \leq i \leq n$ are independent centered sub-Gaussian random variables, and $K_0 > \max_i \|X_{il}\|\psi_2$. By an application of Lemma 5.5 in Vershynin [2010], for any $t > 0$, $0 \leq l \leq p-1$, we have

$$P(|X_{il}| > t) \leq e \cdot \exp\left(-\frac{t^2}{K_1^2}\right),$$

where $K_1 > 0$ is an absolute constant only depending on K_0. By maximizing l on both sides, one can get

$$\sup_{1\leq l\leq p} P(|X_{il}| > t) \leq e \cdot \exp\left(-\frac{t^2}{K_1^2}\right).$$

Let $C_1 = e$, this completes the proof. $\square$

Lemma S2.2. *Let*

$$\Phi_p(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{p-q}\sum_{l=1}^{p-q}(X_{il}X_{il+q})(X_{jl}X_{jl+q}) - \frac{1}{p-q}\sum_{l=1}^{p-q}\sigma_{ll+q}^2,$$

for $0 \leq q \leq p-1$*. Under Assumption S2.2, we have, for any* $1 \leq i \neq j \leq n$ *and* $t > 0$*,*

$$\max_{0\leq q\leq p-1} P(|\Phi_p(\mathbf{X}_i, \mathbf{X}_j)| > t) \leq Cp \cdot \exp\left(\frac{-\min\{t, t^{1/2}\}}{K_2}\right). \qquad \text{(S2.8)}$$

Proof. Step 1. Let $Q_{ij,q} = \frac{1}{p-q}\sum\limits_{l=1}^{p-q}(X_{il}X_{il+q})(X_{jl}X_{jl+q})$. We have, for any $t > 0$,

$$\begin{aligned} P(|Q_{ij,q}| > t) &\leq P\left(\bigcup_{1\leq l\leq p-q}\{|X_{il}X_{il+q}X_{jl}X_{jl+q}| \geq t\}\right) \\ &\leq (p-q)\sup_{1\leq l\leq p} P\left(|X_{il}X_{il+q}X_{jl}X_{jl+q}| \geq t\right) \qquad \text{(S2.9)} \\ &\leq 4(p-q)\sup_{1\leq l\leq p} P(|X_{il}| > t^{1/4}). \end{aligned}$$

By applying Lemma S2.1, there exists $C_1 > 0$ only depending on K_0 such that for any $0 < t^{1/4} < 1$

$$\sup_{1\leq l\leq p} P\left(|X_{il}| > t^{1/4}\right) \leq C_1 \exp\left(-\frac{t^{1/2}}{K_1}\right),$$

which implies, together with (S2.9),

$$\begin{aligned} P(|Q_{ij,q}| > t) &\leq 4(p-q)C_1 \exp(-\frac{t^{1/2}}{K_1}), \\ \max_{0\leq q\leq p-1} P(|Q_{ij,q}| > t) &\leq 4p \cdot C_1 \exp\left(-\frac{t^{1/2}}{K_1}\right). \end{aligned} \tag{S2.10}$$

Step 2. Combining the conclusion (S2.10) of Step 1. and the formula (5.14)-(5.15) in Vershynin [2010], there exists a constant M_1 such that

$$\|Q_{ij,q}\|_{\psi_1} \leq M_1.$$

By Remark 5.18 'Centering' of Vershynin [2010],

$$\|Q_{ij,q} - E(Q_{ij,q})\|_{\psi_1} \leq 2\|Q_{ij,q}\|_{\psi_1} \leq 2M_1.$$

Similarly, following the formula (5.14)-(5.15) in Vershynin [2010], there exists a constant M_2 such that

$$\max_{0\leq q\leq p-1} P\left(|Q_{ij,q} - E(Q_{ij,q})\,| > t\right) \leq 4p \cdot C_1 \exp\left(-\frac{\min\{t, t^{1/2}\}}{M_2}\right).$$

Notice that $\Phi_q(\mathbf{X}_i, \mathbf{X}_j) = Q_{ij,q} - E(Q_{ij,q})$. Letting $4C_1 = C, K_2 = M_2$, we conclude that for any $1 \leq i \neq j \leq n$ and $t > 0$,

$$\max_{0\leq q\leq p-1} P\left(|\Phi_p(\mathbf{X}_i, \mathbf{X}_j)| > t\right) \leq Cp \cdot \exp\left(-\frac{\min\{t, t^{1/2}\}}{K_2}\right),$$

where K_2 depends on K_0. □

Now we recall an inequality on U-statistics in Theorem A of Serfling [1980]. Let $Y_i, Y_2, \cdots, Y_n$ are $i.i.d$ variables. If $a < h(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_m}) < b$ for any $1 \leq i_1 < \cdots < i_m \leq n$. Then the U-statistic

$$F_n = \frac{1}{A_n^m} \sum_{1\leq i_1<\cdots<i_m\leq n} h(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_m})$$

satisfies that for any $t > 0$,

$$P(F_n - E(F_n) \geq t) \leq \exp\left(-\frac{nt^2}{(b-a)^2}\right). \tag{S2.11}$$

Lemma S2.3. *Under Assumptions S2.1 and S2.2,we have*

$$P\left(\sup_{0\le q\le p-1}|\tilde{h}(q)-h(q)|\ge Cq_n\right)=o_p(1),$$

where $q_n = O\left(\sqrt{\frac{\{\log(p\vee n)\}^5}{n}}\right)$.

Proof. First, we assume for any $i \neq j$, $E(\Phi_q(\mathbf{X}_i, \mathbf{X}_j)) = 0$. For any $0 \le q \le p-1$, we have

$$\begin{aligned}\tilde{h}(q)-h(q) &= \frac{1}{A_n^2}\sum_{i,j}^{*}\frac{1}{p-q}\sum_{l=1}^{p-q}(X_{il}X_{il+q})(X_{jl}X_{jl+q})-\frac{1}{p-q}\sum_{l=1}^{p-q}\sigma_{ll+q}^2\\ &= \frac{1}{A_n^2}\sum_{i,j}^{*}\left(\frac{1}{p-q}\sum_{l=1}^{p-q}X_{il}X_{il+q}X_{jl}X_{jl+q}-\frac{1}{p-q}\sum_{l=1}^{p-q}\sigma_{ll+q}^2\right)\\ &\equiv: \frac{1}{A_n^2}\sum_{i,j}^{*}\Phi_q(\mathbf{X}_i,\mathbf{X}_j).\end{aligned}$$

Define the event $A_{n,t} = \left\{\max_{0\le q\le p-1, 1\le i,j\le n}|\Phi_q(\mathbf{X}_i,\mathbf{X}_j)| > t\right\}$ for any $t > 0$ and let $t_0 = (5K_1)^2\{\log(p\vee n)\}^2$. As $\mathbf{X}_i,\ 1\le i\le n$ are *i.i.d* variables, together with the conclusion (S2.8) in Lemma S2.2, we get

$$\begin{aligned}P(A_{n,t_0}) &\le pn^2\left\{\max_{0\le q\le p-1,1\le i,j\le n}|\Phi_q(\mathbf{X}_i,\mathbf{X}_j)|>t_0\right\}\\ &\le Cn^2p^2\cdot\exp\left(-\min\{t_0,t_0^{1/2}\}/K_1\right) \qquad \text{(S2.12)}\\ &\le C(p\vee n)^{-1},\end{aligned}$$

and

$$
\begin{aligned}
&P\left(\max_{0\le q\le p-1}|\tilde{h}(q)-h(q)|\ge t\right)\\
&\le P\left(\left\{\max_{0\le q\le p-1}|\tilde{h}(q)-h(q)|\ge t\right\}\cap A_{n,t_0}^c\right)+P(A_{n,t_0}).
\end{aligned}
$$

Define $\tilde{\Phi}_q(\mathbf{X}_i,\mathbf{X}_j)=\Phi_q(\mathbf{X}_i,\mathbf{X}_j)I(|\Phi_q(\mathbf{X}_i,\mathbf{X}_j)|<t_0)$ and

$$
F_n=\frac{1}{A_n^2}\sum_{i,j}^{*}\left\{\tilde{\Phi}_q(\mathbf{X}_i,\mathbf{X}_j)-E(\tilde{\Phi}_q(\mathbf{X}_i,\mathbf{X}_j))\right\},
$$

for $0\le q\le p-1,\ 1\le i,j\le n$. It is obvious that F_n is an U-statistic with bounded kernel function in the interval $[-t_0,t_0]$. Taking $t=2\sqrt{2}\sqrt{\frac{\log p}{n}}t_0$, and applying the inequality (S2.11), we obtain

$$
\begin{aligned}
&P\left(\left\{\max_{0\le q\le p-1}|\tilde{h}(q)-h(q)|\ge t\right\}\cap A_{n,t_0}^c\right)\\
&=P\left(\left\{\max_{0\le q\le p-1}|F_n|\ge t\right\}\cap A_{n,t_0}^c\right)\\
&\le P\left(\sup_{0\le q\le p-1}|F_n|\ge 2\sqrt{2}\sqrt{\frac{\log p}{n}}t_0\right)\\
&\le p\sup_{0\le q\le p-1}P\left(|F_n|\ge 2\sqrt{2}\sqrt{\frac{\log p}{n}}t_0\right)\\
&\le \exp(-\log p)\\
&=p^{-1}.
\end{aligned}
\qquad\text{(S2.13)}
$$

Therefore, combining the formula (S2.12) and (S2.13) and letting $\tilde{C} = C+1$, we derive that

$$P\left(\sup_{0\le q\le p-1} |\tilde{h}(q) - h(q)| \ge 2\sqrt{2}\sqrt{\frac{\log p}{n}} t_0\right) \le \tilde{C}(p \vee n)^{-1},$$

where $\sqrt{\frac{\log p}{n}} t_0 = O\left(\sqrt{\frac{\{\log(p\vee n)\}^5}{n}}\right)$. The proof is done. □

S2.1 The Proof of Theorem 2.1

To prove that, as $\min(n,p) \to 0$,

$$P\left(\max_{0\le q\le p-1} |\hat{h}(q) - h(q)| > Cq_n\right) = o(1).$$

Proof. It can be seen easily that for any t,

$$\begin{aligned} P\left(\max_{0\le q\le p-1} |\hat{h}(q) - h(q)| > t\right) \le & P\left(\max_{0\le q\le p-1} |\hat{h}(q) - \tilde{h}(q)| > t/2\right) \\ & + P\left(\max_{0\le q\le p-1} |\tilde{h}(q) - h(q)| \ge t/2\right). \end{aligned}$$

Similarly, we have

$$
\begin{aligned}
&P\left(\max_{0\le q\le p-1}|\hat{h}(q)-\tilde{h}(q)|>t/2\right)\\
=&P\Bigg(\max_{0\le q\le p-1}\Bigg|\frac{1}{p-q}\sum_{l=1}^{p-q}\big\{-\frac{2}{A_n^3}\sum_{i,j,k}^{*}X_{il}X_{kl+q}(X_{jl}X_{jl+q})\\
&+\frac{1}{A_n^4}\sum_{i,j,k,m}^{*}X_{il}X_{jl+q}X_{kl}X_{ml+q}\big\}\Bigg|>t\Bigg)\\
\le&P\left(\max_{0\le q\le p-1}|F_{n1}|>\frac{t}{8}\right)+P\left(\max_{0\le q\le p-1}|F_{n2}|>\frac{t}{4}\right),
\end{aligned}
$$

where F_{n1} and F_{n2} have the following forms:

$$
\begin{aligned}
F_{n1} &= \frac{1}{A_n^3}\sum_{i,j,k}^{*}\frac{1}{p-q}\sum_{l=1}^{p-q}X_{il}X_{kl+q}X_{jl}X_{jl+q},\\
F_{n2} &= \frac{1}{A_n^4}\sum_{i,j,k,m}^{*}\frac{1}{p-q}\sum_{l=1}^{p-q}X_{il}X_{jl+q}X_{kl}X_{ml+q}.
\end{aligned}
$$

Notice that F_{n1} and F_{n2} have the same type as F_n and are both U-statistics. By using the similar arguments as the previous proof for Lemma S2.3, there exist constants $C_1,\ C_2$ such that

$$
\begin{aligned}
P\left(\sup_{0\le q\le p-1}|F_{n1}-EF_{n1}|\ge C_1q_n\right)&=o_p(1),\\
P\left(\sup_{0\le q\le p-1}|F_{n2}-EF_{n2}|\ge C_2q_n\right)&=o_p(1).
\end{aligned}
$$

Note that $EF_{n1} = EF_{n2} = 0$. Letting $C = \max\{8C_1, 4C_2\}$, we have

$$\begin{aligned} P\left(\max_{0\le q\le p-1} |\hat{h}(q) - \tilde{h}(q)| > Cq_n\right) &\le P\left(\max_{0\le q\le p-1} |F_{n1}| > \frac{Cq_n}{8}\right) \\ &\quad + P\left(\max_{0\le q\le p-1} |F_{n2}| > \frac{Cq_n}{4}\right) \\ &= o_p(1). \end{aligned} \tag{S2.14}$$

The results in S2.14 and Lemma S2.3 lead to the conclusion of this theorem.

□

S2.2 The Proof of Theorem 2.2

Proof. Under Assumptions S2.1 and S2.2, Theorem 2.1 indicates that the following inequality holds with a probability tending to 1:

$$\max_{0\le k\le p-1} |\hat{h}(k) - h(k)| < C_0 q_n. \tag{S2.15}$$

Thus, we have that with a probability tending to 1,

$$h(k) - C_0 q_n \le \hat{h}(k) \le h(k) + C_0 q_n, \ \forall\ 0 \le k \le p-1.$$

This implies that

$$-C_0 q_n \le \min_{K+1 \le k \le p-1} \hat{h}(k) \le \max_{K+1 \le k \le p-1} \hat{h}(k) \le C_0 q_n.$$

Now we turn to compute the VCC objective function. From the definition of $\hat{s}_k$ in (2.2), we can see that when $k = K$, the following inequality holds:

$$\frac{-C_0 q_n + c_n}{h(K) + C_0 q_n + c_n} \le \hat{s}(K) = \frac{\hat{h}(K+1) + c_n}{\hat{h}(K) + c_n} \le \frac{C_0 q_n + c_n}{h(K) - C_0 q_n + c_n}.$$

Due to the conditions $c_n \to 0$, $c_n/h(K) \to 0$ and $q_n/c_n = o(1)$, we have in probability

$$\hat{s}(K) \to 0.$$

Additionally, we have the following inequality in probability:

$$\min_{k>K} \hat{s}(k) \ge \frac{\min_{k>K} \hat{h}(k+1) + c_n}{\max_{k>K} \hat{h}(k) + c_n} \ge \frac{-C_0 q_n + c_n}{C_0 q_n + c_n} \to 1,$$

which is also due to the conditions $c_n \to 0$, $c_n/h(K) \to 0$ and $q_n/c_n = o(1)$.

Therefore, we can conclude $P(\hat{K} = K) \to 1$, as $n, p \to \infty$. □

S2.3 The Proof of Theorem 3.1

Proof. As $\mathbf{X}_i$ follow a normal distribution, given $\chi_j^{(M)}$, $\hat{t}_{j,j-k}^{(M)}$ has the conditional normal distribution whose mean is $t_{j,j-k}$. Let $\tilde{t}(j) = \hat{t}_{j,j-k}^{(M)} - t_{j,j-k}$. It is easy to obtain that $\tilde{t}(j)$ for $k+1 \leq j \leq p$ are sub-Gaussian variables. By applying Lemma 5 of Vershynin [2010], we can derive that there exist constants C_2 and K_0 such that

$$P(|\tilde{t}(j)| > t) \leq C_2 \exp\left(-\frac{nt^2}{K_0^2}\right).$$

Let $t = C_2' K_0 \sqrt{\frac{\log p}{n}}$ with $C_2' > \sqrt{3}$. We have

$$\begin{aligned}
P\left(\max_{0\leq k\leq M} |\hat{l}(k) - l(k)| \geq t\right) &\leq p \cdot \max_{0\leq k\leq M} P\left(|\hat{l}(k) - l(k)| \geq t\right) \\
&\leq p \cdot \max_{0\leq k\leq M} P\left(\frac{1}{p-k}\left|\sum_{j=k+1}^{p} (|\hat{t}_{j,j-k}^{(M)}| - |t_{j,j-k}|)\right| \geq t\right) \\
&\leq p \cdot \max_{0\leq k\leq M} P\left(\frac{1}{p-k}\left|\sum_{j=k+1}^{p} \tilde{t}(j)\right| \geq t\right) \\
&\leq p^2 \cdot \max_{0\leq k\leq M} P(|\tilde{t}(j)| > t) \\
&\leq C_2 \exp\left(-\frac{nt^2}{K_0^2} + 2\log p\right) \\
&\leq C_2 p^{-1}.
\end{aligned}$$

Letting $C_1 = C_2' K_0$, the proof is completed. □

S2.4 The Proof of Theorem 3.2

Proof. We follow the similar arguments of proving Theorem 2.2 to prove this theorem. Under Assumptions S2.1 and S2.2, Theorem 3.2 indicates that with a probability tending to 1, we have the following inequality as:

$$\max_{0\le k\le M} |\hat{l}(k) - l(k)| < C_0\gamma_n, \tag{S2.16}$$

where $\gamma_n = O(\sqrt{\log p/n})$. The above implies that with a probability tending to 1,

$$l(K) - C_0\gamma_n \le \hat{l}(K) \le l(K) + C_0\gamma_n,$$

$$-C_0\gamma_n \le \min_{K+1\le k\le p-1} \hat{l}(k) \le \max_{K+1\le k\le p-1} \hat{l}(k) \le C_0\gamma_n.$$

Combining the definition of $\hat{r}_k$ in (3.3), we have the following inequality:

$$\frac{-C_0\gamma_n + \tilde{c}_n}{l(K) + C_0\gamma_n + \tilde{c}_n} \le \frac{\hat{l}(K+1) + \tilde{c}_n}{\hat{l}(K) + \tilde{c}_n} \le \frac{C_0\gamma_n + \tilde{c}_n}{l(K) - C_0\gamma_n + \tilde{c}_n}.$$

As $\tilde{c}_n \to 0$, $\tilde{c}_n/l(K) \to 0$ and $\tilde{c}_n\sqrt{n/\log p} \to \infty$, we have in probability

$$\hat{r}(K) = \frac{\hat{l}(K+1) + \tilde{c}_n}{\hat{l}(K) + \tilde{c}_n} \to 0.$$

On the other hand, we derive:

$$\min_{k>K} \hat{r}(k) \geq \frac{\min\limits_{k>K} \hat{l}(k+1) + c_n}{\max\limits_{k>K} \hat{l}(k) + c_n} \geq \frac{-C_0\gamma_n + c_n}{C_0\gamma_n + c_n} \to 1,$$

due to $\tilde{c}_n \to 0$, $\tilde{c}_n / l(K) \to 0$ and $\tilde{c}_n \sqrt{n/\log p} \to \infty$.

Therefore, as $n, p \to \infty$, $\hat{K} = K$ with a probability going to 1. □

S2.5 The Proof of Theorem S1.1

Proof. To get the order of $\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\|$, we calculate the following probability value:

$$\begin{aligned} P(\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| > t) =& P(\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| > t | \hat{K} = K) P(\hat{K} = K) \\ &+ P(\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| > t | \hat{K} \neq K) P(\hat{K} \neq K) \\ \leq& P(\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| > t | \hat{K} = K) + P(\hat{K} \neq K) \\ \leq& \frac{E\|\hat{\Sigma}_{B,K} - \Sigma\|^2}{t^2} + P(\hat{K} \neq K). \end{aligned} \tag{S2.17}$$

By the proof of Theorem 1 in Bickel and Levina [2008] and noticing that $\sigma_{ij} = 0$ for all $|i - j| > K$, we have

$$E\|\hat{\Sigma}_{B,K} - \Sigma\|^2 \leq C \left(K \frac{\log p}{n} + K \left(\frac{\log p}{n} \right)^{1/2} \right)^2 \leq C \left(K^2 \frac{\log p}{n} \right). \tag{S2.18}$$

The last inequality is obtained as $\frac{\log p}{n} \to 0$. Combining (S2.17) and (S2.20) and letting $t = K\log(p \vee n)^{1/2-5\gamma/4}n^{\gamma/4-1/2}$ for any $\gamma > 0$, we have

$$P(\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| > t) \leq C\left(\sqrt{\frac{\log(p \vee n)^5}{n}}\right)^{\gamma} + P(\hat{K} \neq K).$$

Noting that $\sqrt{\frac{\log(p\vee n)^5}{n}} = q_n \to 0$ and $P(\hat{K} \neq K) = o_p(1)$, it follows that

$$P(\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| > t) = o_p(1).$$

Thus, we conclude that $\|\hat{\Sigma}_{B,\hat{K}} - \Sigma\| = O_p\left(K\log(p \vee n)^{1/2-5\gamma/4}n^{\gamma/4-1/2}\right)$. This completes the proof. □

S2.6 The Proof of Theorem S1.2

Proof. Similarly, we have

$$P(\|\hat{\Sigma}_{T,\hat{K}} - \Sigma\| > t) \leq \frac{E\|\hat{\Sigma}_{T,K} - \Sigma\|^2}{t^2} + P(\hat{K} \neq K). \tag{S2.19}$$

By the proof of Theorem 2 in Cai et al. [2010] and the fact that $\sigma_{ij} = 0$ for all $|i - j| > K$, we have

$$E\|\hat{\Sigma}_{T,K} - \Sigma\|^2 \leq C\frac{K + \log p}{n}. \tag{S2.20}$$

Letting $t = \frac{(K+\log p)^{1/2}}{\log(p\vee n)^{5\alpha/4}n^{1/2-\alpha/4}}$ for any $\alpha > 0$, it follows that

$$P(\|\hat{\Sigma}_{T,\hat{K}} - \Sigma\| > t) \leq C\left(\sqrt{\frac{\log(p \vee n)^5}{n}}\right)^{\alpha} + P(\hat{K} \neq K) = o_p(1).$$

Then the result holds. □

S2.7 The Proof of Theorem S1.3

Proof. Let

$$\hat{\mathbf{X}}_j = \sum_{t=1}^{j-1} a_{jt}\mathbf{X}_t = \mathbf{Z}_j^T\mathbf{a}_j$$

where $\mathbf{Z}_j = (\mathbf{X}_1, \ldots, \mathbf{X}_{j-1})^T$ and $\mathbf{a}_j = (a_{j1}, \ldots,\ a_{j,j-1})^T$ the coefficients. Each vector $\mathbf{a}_j^T$ can be computed as

$$\mathbf{a}_j = (\mathrm{Var}\,(\mathbf{Z}_j))^{-1}\,\mathrm{Cov}\,(\mathbf{X}_j, \mathbf{Z}_j)\,.$$

Let $\varepsilon_j = \mathbf{X}_j - \hat{\mathbf{X}}_j, d_j^2 = \mathrm{Var}\,(\varepsilon_j)$ and let $D = \mathrm{diag}\left(d_1^2, \ldots, d_p^2\right)$. By Cholesky decomposition, we can have

$$\Sigma_p = (I - A)^{-1}D\left[(I - A)^{-1}\right]^T,$$
$$\Sigma_p^{-1} = (I - A)^T D^{-1}(I - A).$$

Let $A_k = B_k(A)$ and D_k be the diagonal matrix containing the corresponding residual variances. Moreover, by $\tilde{A}_k$ and $\tilde{D}_k$, we denote the empirical versions of A_k and D_k. Therefore, we have

$$\hat{\Omega}_{\hat{K}} - \Omega_{\hat{K}} = \left(I - \tilde{A}_{\hat{K}}\right) \tilde{D}_{\hat{K}}^{-1} \left(I - \tilde{A}_{\hat{K}}\right)^T - \left(I - A_{\hat{K}}\right) D_{\hat{K}}^{-1} \left(I - A_{\hat{K}}\right)^T .$$

Define $A^{(1)} = \left[A^{(3)}\right]^T = I - \tilde{A}_{\hat{K}}, B^{(1)} = \left[B^{(3)}\right]^T = I - A_{\hat{K}}, A^{(2)} = \tilde{D}_{\hat{K}}^{-1}, B^{(2)} = D_{\hat{K}}^{-1}$, and then we have

$$\begin{aligned}
&\|A^{(1)}A^{(2)}A^{(3)} - B^{(1)}B^{(2)}B^{(3)}\| \\
\leq &\sum_{j=1}^{3} \|A^{(j)} - B^{(j)}\| \prod_{k \neq j} \|B^{(k)}\| + \sum_{j=1}^{3} \|B^{(j)}\| \prod_{k \neq j} \|A^{(k)} - B^{(k)}\| \\
&+ \prod_{j=1}^{3} \|A^{(j)} - B^{(j)}\| .
\end{aligned}$$

By the Lemma A.2 and Lemma A.3 in Bickel and Levina [2008], we have

$$\begin{aligned}
\sum_{j=1}^{3} \|A^{(j)} - B^{(j)}\| \prod_{k \neq j} \|B^{(k)}\| &= O_p\left(\left(\frac{\log p}{n}\right)^{1/2} \frac{(K^2 \log p)^2}{n}\right), \\
\sum_{j=1}^{3} \|B^{(j)}\| \prod_{k \neq j} \|A^{(k)} - B^{(k)}\| &= O_p\left(K^2 \frac{(\log p)^2}{n} \left(1 + \left(\frac{\log p}{n}\right)^{1/2}\right)\right), \\
\prod_{j=1}^{3} \|A^{(j)} - B^{(j)}\| &= O_p\left(\left(K^2 \frac{(\log p)^2}{n} + \left(\frac{\log p}{n}\right)^{1/2}\right)\right).
\end{aligned}$$

Noting that $\frac{\log p}{n} \to 0$, we have $\|\hat{\Omega}_{\hat{K}} - \Omega_{\hat{K}}\| = O_p\left(K^2 \frac{(\log p)^2}{n}\right)$. By similar analysis with the proof of Theorem S1.1 and choosing $t = K^2 \frac{(\log p)^{2-\beta/4}}{n^{1-\beta/4}}$, we can get

$$P(\|\hat{\Omega}_{\hat{K}} - \Omega\| > t) = o_p(1).$$

This completes the proof. □

References

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

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.

R. J. Serfling. *Approximation Theorems of Mathematic Statistics*. 1980.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.