Band Width Selection for High Dimensional Covariance Matrix Estimation

Qiu, Yumou and Chen, Song Xi

Iowa State University, Peking University

2014

Band Width Selection for High Dimensional Covariance Matrix Estimation

Yumou Qiu and Song Xi Chen

Iowa State University, Peking University and Iowa State University

Abstract

The banding estimator of Bickel and Levina (2008a) and its tapering version of Cai, Zhang and Zhou (2010), are important high dimensional covariance estimators. Both estimators require a band width parameter. We propose a band width selector for the banding estimator by minimizing an empirical estimate of the expected squared Frobenius norms of the estimation error matrix. The ratio consistency of the band width selector is established. We provide a lower bound for the coverage probability of the underlying band width being contained in an interval around the band width estimate. Extensions to the band width selection for the tapering estimator and threshold level selection for the thresholding covariance estimator are made. Numerical simulations and a case study on sonar spectrum data are conducted to demonstrate the proposed approaches.

Key Words and Phrases: Banding estimator; Large p, small n; Ratio-consistency; Tapering estimator; Thresholding estimator.

1 Introduction

With the advance in the modern data collection technology, data of very high dimensions are increasingly collected in scientific, social economic and financial studies, which include the microarray data, the next generation sequencing data, recordings of large networks and financial observations of large portfolios. Suppose we observe independent and identically distributed p-dimensional random variables $X_1, \cdots, X_n$ with an unknown covariance matrix $\Sigma = \mathrm{Var}(X_1)$. The covariance Σ is of great importance in multivariate analysis. The sample covariance $S_n = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)'$ is a popular and valid estimator of Σ in conventional settings where the dimension p is fixed and the sample size n is relative large. However, for high dimensional data such that $p/n \to c \in (0, \infty]$, it is known that S_n is no longer consistent; see Bai and Ying (1993), Bai, Silverstein and Yin (1998) and Johnstone (2001) for accounts of the issue.

There have been advances in constructing consistent covariance estimators for high dimensional data via the regularization methods that involve thresholding or truncation. Regularization based on the Cholesky decomposition has been considered in Wu and Pourahmadi (2003), Huang, Liu, Pourahmadi and Liu (2006) and Rothman, Levina and Zhu (2010) for estimating Σ and its inverse. Bickel and Levina (2008a) proposed banding the sample covariance S_n that truncates all sub-diagonal entries beyond certain band width to zero. Cai, Zhang and Zhou (2010) investigated a tapering estimator which can be viewed as a soft banding on the sample covariance, and demonstrated that it can attain the minimax optimal rate. For random vectors which do not have a natural ordering so that the elements of Σ do not decay as they move away from the diagonal, Bickel and Levina (2008b) proposed a thresholding estimator, which was further developed by Rothman, Levina and Zhu (2009) and Cai and Liu (2011). Regularized estimation of Σ^{-1} has also been developed in Bickel and Levina (2008a), Cai, Liu and Luo (2011) and Xue and Zou (2012).

The banding and tapering estimators require specifying the band width that defines the number of sub-diagonals which are not truncated to zero. For the thresholding estimator, a threshold level needs to be determined. Bickel and Levina (2008a,b) and Cai et al. (2010) showed that the performance of these estimators are crucially dependent on the choice of the band width or the threshold level. Bickel and Levina (2008a,b) introduced cross-validation approximations to the Frobenius risk of estimation by repeated random splitting of the sample to two segments. One segment of the sample was used to estimate Σ and the other was employed to form cross-validation scores for the band width and the threshold level selection, respectively. The conventional sample covariance was used to estimate Σ in the first segment. This can adversely affect the performance of the band width or threshold level selection due to the sample covariance's known defects under high dimensionality. For banded covariances, Qiu and Chen (2012) proposed a method to select the band width, using a by-product of their test for the bandedness of Σ. Yi and Zou (2013) proposed a band width selection for the tapering estimator by minimizing the expected squared Frobenius norm of the estimation error matrix for Gaussian distributed data.

In this paper, we employ the Frobenius risk of the banding and the tapering estimators as the objective function, and define the underlying band width as the smallest band width that minimizes the objective function. By studying the objective function under a general distributional framework, we investigate the properties of the underlying band width under a bandable covariance class that is better suited for the Frobenius norm. An estimator of the band width is proposed by minimizing a nonparametric estimator of the objective function. The use of the Frobenius norm, as Yi and Zou (2013) have noted, confers easier tractability than that based on the spectral norm. The ratio consistency of the proposed band width estimator to the underlying band width is established. We give a lower bound for the coverage probability of the underlying band width being contained in an interval around the estimated band width. Extensions to the tapering and thresholding estimators are considered.

The paper is organized as follows. The new bandable covariance class and some needed assumptions are outlined in Section 2. Section 3 defines the underlying band width and gives its properties. A ratio consistent band width estimator is constructed and its theoretical properties are investigated in Section 4. Section 5 provides an extension to the band width selection for the tapering estimator. Section 6 extends to the threshold level selection for the thresholding estimator. Simulation results and a real data analysis are presented in Sections 7 and 8, respectively. Technical proofs are provided in the Appendix and the Supplementary Material, respectively.

2 Bandable Classes and Assumptions

Let $X_1, X_2, \dots X_n$ be independent and identically distributed (IID) p-dimensional random vectors with mean μ and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$. Throughout the paper, $||\cdot||_F$, $||\cdot||_{(2,2)}$ and $||\cdot||_{(1,1)}$ denote the Frobenius, the spectral and the ℓ_1 norms of a matrix, respectively; and C with or without subscripts denotes positive constants whose value may change on different occasions. We make the following assumptions.

Assumption 1. *As $n \to \infty$, $p = p(n) \to \infty$ and $\limsup_{n\to\infty} n/p \le C < \infty$.*

Assumption 2. *(i) $X_i = \Gamma Z_i + \mu$, where Γ is a $p \times m$ matrix of constants with $m \ge p$, $\Gamma\Gamma' = \Sigma$, and $Z_1, \cdots, Z_n$ are IID m-dimensional random vectors such that $\mathrm{E}(Z_1) = 0$ and $\mathrm{Var}(Z_1) = I_m$. (ii) For $Z_1 = (Z_{11}, \dots, Z_{1m})^T$, $\{Z_{1l}\}_{l=1}^m$ are independent, and there exist finite constants Δ and ω such that $\mathrm{E}(z_{1l}^4) = 3+\Delta$ and $\mathrm{E}(z_{1l}^3) = \omega$ for $l = 1, \cdots, m$.*

Assumption 1 prescribes the mechanism governing the sample size and the dimensionality. The last part of Assumption 1 contains the "large p, small n" paradigm where p can be much larger than n, as well as the case of p and n being the same order. For the band width selection, no specific relationship between n and p is needed. However, for the threshold level selection discussed in Section 6, a restriction in the form of $\log p = o(n^{1/3})$ is required. Assumption 2 is a version of the general multivariate model employed in Bai and Saranadasa (1996) and Qiu and Chen (2012), where $\{Z_{il}\}_{l=1}^{m}$ may be viewed as the innovations of the data.

Bickel and Levina (2008a) considered the following "bandable" class of covariances:

$$\begin{aligned}\mathfrak{U}_1(\alpha, C) = \Big\{\Sigma : \max_{l_2} \sum_{|l_1-l_2|>k} |\sigma_{l_1 l_2}| \le Ck^{-\alpha} \ \text{ for all } \ k > 0 \\ \text{and } \ 0 < \varepsilon_1 \le \lambda_{min}(\Sigma) \le \lambda_{max}(\Sigma) \le \varepsilon_1^{-1}\Big\}\end{aligned} \tag{2.1}$$

for positive constants α, C and ε_1. For $p \times p$ matrix $M = (m_{l_1 l_2})_{p\times p}$, let $B_k(M) = (m_{l_1 l_2}\mathrm{I}\{|l_1 - l_2| \le k\})_{p\times p}$ be a banded version with a band width $k \in \{0, \cdots, p-1\}$. Bickel and Levina (2008a) proposed $B_k(S_n)$ as an estimator of Σ, where S_n is the sample covariance, and showed that

$$\mathrm{E}||B_k(S_n) - \Sigma||^2_{(2,2)} = O\big\{(\log(p)/n)^{\alpha/(1+\alpha)}\big\} \quad \text{if } k = \{\log(p)/n\}^{-1/(2+2\alpha)}.$$

Cai et al. (2010) considered a slightly different class

$$\begin{aligned}\mathfrak{U}_2(\alpha, C) = \Big\{\Sigma : \max_{l_2} \sum_{|l_1-l_2|>k} |\sigma_{l_1 l_2}| \le Ck^{-\alpha} \ \text{ for all } \ k > 0 \\ \text{and } \ 0 < \varepsilon_2 \le \min\{\sigma_{ll}\} \le \max\{\sigma_{ll}\} \le \varepsilon_2^{-1}\Big\}.\end{aligned} \tag{2.2}$$

They replaced the restriction on the eigenvalues in $\mathfrak{U}_1(\alpha, C)$ with those on the diagonal elements. For $\mathfrak{U}_2(\alpha, C)$, Cai et al. (2010) proposed the tapering estimator $T_k(S_n) = \Omega_T(k)\circ S_n$, where $\circ$ denotes the Hadamard product, and $\Omega_T(k) = \big(\omega_{l_1 l_2}\big)$ is the weighting matrix with

$$\omega_{l_1 l_2} := k^{-1}\{(2k - |l_1 - l_2|)_+ - (k - |l_1 - l_2|)_+\}.$$

Note that $\omega_{l_1 l_2} = 1$ for $|l_1 - l_2| \le k$, $\omega_{l_1 l_2} = 0$ for $|l_1 - l_2| \ge 2k$ and $\omega_{l_1 l_2}$ decreases linearly for $k < |l_1 - l_2| < 2k$. For easy algebraic manipulation, we use $2k$ as the effective band width rather than k as in Cai et al. (2010).

Cai et al. (2010) showed that for $k \sim n^{1/(1+2\alpha)}$

$$\mathrm{E}||T_k(S_n) - \Sigma||^2_{(2,2)} = O\{\log(p)/n + n^{-2\alpha/(1+2\alpha)}\},$$

which attains the minimax convergence rate over $\mathfrak{U}_2(\alpha, C)$. The banding and tapering estimators are not necessarily positive definite. One way to mitigate the problem is to obtain the spectral decomposition of the covariance estimators and replace the negative and zero eigenvalues with small positive values as suggested by Cai et al. (2010).

It is clear from the analysis in Bickel and Levina (2008a) and Cai et al. (2010) that the convergence rates of the banding and the tapering estimators are critically dependent on the band width k, whereas the band width k depends on the unknown index parameter α of the bandable classes. However, estimating the index parameter is a challenging problem.

We shall consider another "bandable" matrix class which is better suited for band width selection based on the Frobenius norm. To define the new "bandable" covariance class, let us define for $k = \{0, 1, \cdots, p-1\}$,

$$h(k) := \frac{1}{2(p-k)} \sum_{|l_1 - l_2| = k} \sigma^2_{l_1 l_2} = \frac{1}{p-k} \sum_{l=1}^{p-k} \sigma^2_{l\, l+k}$$

to be the average of the squares of the kth sub-diagonal entries.

For a fixed positive constant ν and the Δ in Assumption 2, define a covariance class

$$\begin{aligned} \mathcal{G}(\nu, q_p^0) = \big\{ \Sigma : & (i)\ \nu^{-1} \le \lambda_{min}(\Sigma) \le \lambda_{max}(\Sigma) \le \nu; \\ & \text{(ii) } h(k) = o(k^{-1}) \text{ and } \textstyle\sum_{q>k} h(q) \to 0 \text{ for } k \to \infty \text{ and } p \to \infty; \\ & \text{(iii) there exists a sequence } q_p^0 \to \infty \text{ and } q_p^0 = o(n) \text{ such that} \\ & \ nh(k) > (2 + |\Delta|)\lambda^2_{max}(\Sigma) \text{ for } k \le q_p^0 \text{ and } n \text{ large} \big\}. \end{aligned} \tag{2.3}$$

The bounded largest and smallest eigenvalues in Part (i) replicates that in $\mathfrak{U}_1(\alpha, C)$. Part (ii) of (2.3) prescribes that $h(k)$ diminishes to zero at a rate faster than k^{-1} for k large. It may be viewed as an analogue to the sparsity condition

$$\max_{l_2} \sum_{|l_1 - l_2| > k} |\sigma_{l_1 l_2}| \le C k^{-\alpha} \tag{2.4}$$

in $\mathfrak{U}_1(\alpha, C)$ and $\mathfrak{U}_2(\alpha, C)$. Note that for another covariance matrix class

$$\mathfrak{F}(\beta, M) = \{\Sigma : |\sigma_{lj}| \le M(1 + |l - j|)^{-\beta} \text{ for some } \beta > 1/2\}, \tag{2.5}$$

Part (ii) of (2.3) is satisfied. Cai et al. (2010) established the minimax convergence result for the Frobenius norm under $\mathfrak{F}(\beta, M)$ with $\beta > 1$. Hall and Jin (2010) also considered this class in their innovated higher criticism test. Part (iii) of (2.3) requires $h(k)$ to maintain a sufficient amount of "energy" for smaller band widths so that $h(k)$ is at least of order n^{-1}. We note that $h(k)$ actually starts with quite high "energy" since Part (i) implies that $nh(0) = np^{-1}\sum \sigma_{ll}^2 \to \infty$.

The reason for having the sample size n appeared in Part (iii) is because the banding estimator depends on the sample size n. As shown in the next section, the criterion function for the band width selection is based on the expected Frobenius norm of the estimation error matrix of the banding estimator, which inevitably has n involved.

The main difference between $\mathcal{G}(\nu, q_p^0)$ and $\mathfrak{U}_1(\alpha, C)$ or $\mathfrak{U}_2(\alpha, C)$ is that the sparsity in $\mathcal{G}(\nu, q_p^0)$ is written in terms of $h(k)$ whereas that in $\mathfrak{U}_1(\alpha, C)/\mathfrak{U}_2(\alpha, C)$ are defined via $\sum_{|j-l|>k} |\sigma_{jl}|$. This difference reflects the different norms employed in these studies. As we use the Frobenius norm, it is natural to define the sparsity via $h(k)$.

Two specific forms of $h(k)$, which will be referred to repeatedly, are those which decay exponentially and polynomially fast as $k \to \infty$. In the case of the exponential decay,

$$h(k) = C_p(k)\theta^{-k} \quad \text{for some } \theta > 1 \ ; \tag{2.6}$$

in the case of polynomial decay,

$$h(k) = C_p(k)k^{-\beta} \quad \text{for some } \beta > 1. \tag{2.7}$$

In both cases $\{C_p(q)\}_{q=0}^{p-1}$ are sequences bounded within $[C_1, C_2]$ for $C_1 \le C_2$. It can be shown that Part (iii) of (2.3) is satisfied under (2.6) or (2.7) with $q_p^0 = \log n/(2\log\theta)$ for the exponential decay and $q_p^0 = n^{1/(2\beta)}$ for the polynomial decay.

3 Underlying Band Width

In this section, we define the underlying band width for the matrix class $\mathcal{G}(\nu, q_p^0)$. The properties of the underlying band width are given, which provide the basics for its empirical estimation in the next section.

Consider the standardized square of Frobenius norm for $B_k(S_n) - \Sigma$,

$$p^{-1}||B_k(S_n) - \Sigma||_F^2 = p^{-1}\sum_{|l_1-l_2|\le k} (\hat\sigma_{l_1l_2} - \sigma_{l_1l_2})^2 + p^{-1}\sum_{|l_1-l_2|>k} \sigma_{l_1l_2}^2. \tag{3.1}$$

Comparing with the spectral norm, the Frobenius norm is more tractable in the context of the band width estimation. The objective function is

$$\widetilde{\mathrm{Obj}}_B(k) := p^{-1}\mathrm{E}\{||B_k(S_n) - \Sigma||_F^2\}.$$

The underlying band width is $k_B = \min\{k' | k' = \underset{0 \le k < p}{\operatorname{argmin}}\ \widetilde{\mathrm{Obj}}_B(k)\}$. As $\widetilde{\mathrm{Obj}}_B(k)$ is discrete, k_B exists and we choose the smallest minimizer in the case of multiplicity.

We now analyze the properties of k_B for $\Sigma \in \mathcal{G}(\nu, q_p^0)$. Denote $f_{l_1 l_2} = \sum_h \Gamma_{l_1 h}^2 \Gamma_{l_2 h}^2$, where $\Gamma = (\Gamma_{jl})_{p\times m}$ is defined in Assumption 2. A derivation given in Appendix shows that

$$\widetilde{\mathrm{Obj}}_B(k) = \frac{1}{np}\mathrm{tr}(\Sigma^2) + (1-n^{-1})M_n(k) + \frac{\Delta}{np}(1-n^{-1})^2 \sum_{|l_1-l_2|\le k} f_{l_1 l_2}, \tag{3.2}$$

where

$$M_n(k) = \frac{1}{p}\sum_{|l_1-l_2|>k} \sigma_{l_1 l_2}^2 + \frac{1}{np}\sum_{|l_1-l_2|\le k} \sigma_{l_1 l_1}\sigma_{l_2 l_2}. \tag{3.3}$$

As $\mathrm{tr}(\Sigma^2)/(np)$ is irrelevant to k, we only minimize

$$\mathrm{Obj}_B(k) = M_n(k) + \Delta \sum_{q\le k} R(q), \tag{3.4}$$

where $R(q) = (np)^{-1}(1-n^{-1})\sum_{|l_1-l_2|=q} f_{l_1 l_2}$. For Gaussian distributed data, $\Delta = 0$ and $\mathrm{Obj}_B(k) = M_n(k)$. The first term of $M_n(k)$ in (3.3) measures the bias caused by the banding estimation, and the second term penalizes for larger k. Therefore, $\mathrm{Obj}_B(k)$ can be viewed as a penalized risk function of the band width.

The following lemma provides the basic properties of $M_n(k)$ and $R(k)$ in $\mathrm{Obj}_B(k)$.

Lemma 1. *For $\Sigma \in \mathcal{G}(\nu, q_p^0)$,*

$$(i) \quad M_n(k) \sim k/n + p^{-1}\sum_{q>k} 2(p-q)h(q) \to 0 \ \ \textit{for } k \to \infty \ \textit{and } k = o(n)\ ;$$

$$(ii) \quad \sum_{q=0}^{p-1} R(q) \le \nu^2/n.$$

Lemma 1 and (3.4) imply that $M_n(k)$ is at least at the order k/n. Since $\sum_{q\le k} R(q) \le C/n$ for a constant C, $M_n(k)$ is the leading order of $\mathrm{Obj}_B(k)$ as $k \to \infty$.

Let $\sigma_{(1)} \le \sigma_{(2)} \le \cdots \le \sigma_{(p)}$ be the ordered diagonal elements $\{\sigma_{ll}\}_{l=1}^p$ of Σ. Define $a = 2\sigma_{(p)}^2$, $b = \sigma_{(1)}^2/2$ and

$$k_{a,n} = \min\{k : an^{-1} - h(k) > 0\} - 1 \ \text{ and } \ k_{b,n} = \max\{k : bn^{-1} - h(k) < 0\}. \tag{3.5}$$

Denote $\tilde{k}_B$ be the smallest minimizer of $M_n(k)$, and $\lfloor\cdot\rfloor$ be the integer truncation function. The following lemma provides ranges for $\tilde{k}_B$ and k_B.

Lemma 2. *Under Assumptions 1 and 2 and for $\Sigma \in \mathcal{G}(\nu, q_p^0)$,* *(i)* $\tilde{k}_B \in [k_{a,n}, k_{b,n}]$*,* $k_{a,n} \geq q_p^0$ *and* $k_{b,n} = o(n)$*;*

(ii) $k_B \in [k_{a,n} - L, k_{b,n} + L]$ *for* $L = \lfloor 2|\Delta|\nu^4 \rfloor + 1$*.*

The lemma shows that k_B has a broader range than $\tilde{k}_B$. This is due to the uncertainty introduced by $|\Delta|R(k)$ in (3.4). The ranges given in Lemma 2 prepare for $\tilde{k}_B/k_B \to 1$ as $n \to \infty$, the key result of this section. Since $k_{a,n} \geq q_p^0 \to \infty$, it follows from Lemma 1 that $M_n(k)$ is the leading order term of $\mathrm{Obj}_B(k)$ for $k \in [k_{a,n} - L, k_{b,n} + L]$. This suggests that we can minimize $M_n(k)$ directly.

The main thrust of the paper is to minimize an empirical estimator of $M_n(k)$ to obtain an estimator of $\tilde{k}_B$, which may be viewed as a kind of M-estimation. As in the M-estimation, a condition is needed to guarantee the existence of a unique and well-separated minimum of the objective function. Since $M_n(k)$ is the leading order term of $\mathrm{Obj}_B(k)$, a condition that serves this purpose is that, for any small $\delta > 0$ and n large enough,

$$\inf_{k:|k-\tilde{k}_B|>\delta\tilde{k}_B} n\tilde{k}_B^{-1}\{M_n(k) - M_n(\tilde{k}_B)\} > C. \tag{3.6}$$

Condition (3.6) is similar to the second equation of (5.8) in van der Vaart (2000), except that (3.6) imposes a minimum rate of separation $\tilde{k}_B n^{-1}$ between $M_n(k)$ and $M_n(\tilde{k}_B)$. The latter is because that $M_n(\tilde{k}_B)$ shrinks to zero at the rate of $\tilde{k}_B/n$ as revealed by Lemma 1. The following lemma shows that under (3.6), k_B and $\tilde{k}_B$ are ratioly equivalent.

Lemma 3. *For* $\Sigma \in \mathcal{G}(\nu, q_p^0)$ *and under (3.6),* $\tilde{k}_B/k_B \to 1$ *as* $n \to \infty$*.*

As the condition (3.6) is a key condition to the M-estimation for the underlying band width, we provide two sufficient conditions to (3.6) in the Supplementary Material to show it can be satisfied if $h(k)$ decays either exponentially or polynomially.

Exponentially Decayed $h(k)$. In this case $h(k) = C(k)\theta^{-k}$ as specified in (2.6) with $\{C(k)\}_{k=0}^{p-1} \subset [C_1, C_2]$. It is shown in Appendix that $k_B \sim \log n/\log\theta$. A proof in the Supplement Material shows that (3.6) is satisfied under the exponential decay.

Polynomially Decayed $h(k)$. If $h(k)$ decays polynomially as specified in (2.7), $k_B \sim n^{1/\beta}$ as shown in the Appendix. If $\max_{k\in[k_{a,n},k_{b,n}]} |C(k) - C| \to 0$ as $n, p \to \infty$, and the diagonal elements $\{\sigma_{ll}\}_{l=1}^p$ are regulated in certain ways such that

$$\max_{k\in[k_{a,n},k_{b,n}]} p^{-1} \sum_{|l_1-l_2|=q} \sigma_{l_1l_1}\sigma_{l_2l_2} \to 2C_0 \quad \text{as } p \to \infty, \tag{3.7}$$

(3.6) is satisfied. One such situation is when all the diagonal elements are equal. If the diagonal entries differ, but are independent realizations from m super-populations, for a fixed integer m, such that $\{\sigma_{ll}\}_{l=(h-1)p_m+1}^{hp_m} \sim F_h$, where F_h is the hth super-population distribution with mean ϕ_h and finite variance for $h = 1, \cdots, m$ and $p_m = p/m$. It is shown in the Supplementary Material that (3.7) is satisfied with $C_0 = m^{-1}\sum_{h=1}^{m}\phi_h^2$.

Now let us put our analysis in the context of existing results on the banding and tapering estimation. Recall that Bickel and Levina (2008) found that if $k \sim \{\log(p)/n\}^{-1/(2\alpha+2)}$, the spectral risk of the banding estimator is $O_p\{(\log(p)/n)^{\alpha/(\alpha+1)}\}$ uniformly for $\Sigma \in \mathfrak{U}_1(\alpha, C)$. Cai et al. (2010) showed that setting $k \sim n^{1/(2\alpha+1)}$ leads to the minimax optimal rate of $O_p\{n^{-2\alpha/(2\alpha+1)} + \log(p)/n\}$ for the tapering estimator under the spectral norm for $\Sigma \in \mathfrak{U}_2(\alpha, C)$. For the Frobenius norm, they showed that the minimax rate for $\Sigma \in \mathfrak{U}_2(\alpha, C)$ is equivalent to the minimax rate for the smaller class $\mathfrak{F}(\beta, M)$ in (2.5) with $\beta > 1$, and the band width of the tapering estimator corresponding to the minimax optimal rate is $k \sim n^{1/(2\beta)}$. By inspecting their proofs, it can be shown that the banding estimator with $k \sim n^{1/(2\beta)}$ can also attain the minimax lower bound under the Frobenius norm. And the minimax rate of the banding and tapering estimators under $\mathfrak{F}(\beta, M)$ is attained at covariances with $|\sigma_{lj}| = M|l-j|^{-\beta}$. The latter model coincides with the polynomial decay model (2.7) with $h(k) = Mk^{-2\beta}$. We note that this minimax band width rate of $k \sim n^{1/(2\beta)}$ is the rate of the k_B under the polynomial decay as shown in (A.8). Since k_B minimizers the Frobenius risk, the banding estimator with the k_B should attain the minimax convergence rate under the Frobenius norm.

4 Consistent Band Width Estimator

We consider in this section estimating the band width for the banding estimator. A proposal for the tapering estimator will be given in Section 5. As outlined in the previous sections, there are two band widths k_B and $\tilde{k}_B$, which are asymptotically equivalent to each other under (3.6). However, it is easier to estimate $\tilde{k}_B$ than k_B since $M_n(k)$ is more readily estimated. Clearly, if $\Delta = 0$ as in the Gaussian case, $\text{Obj}_B(k) = M_n(k)$ which implies $k_B = \tilde{k}_B$. However, if $\Delta \neq 0$, it is difficult to estimate $\text{Obj}_B(k) = M_n(k) + \Delta\sum_{q\le k} R(q)$ due to its requiring estimating $R(k)$ and Δ.

According to (3.3), in order to estimate $M_n(k)$, we need to estimate, respectively,

$$W(k) := p^{-1}\sum_{|l_1-l_2|>k}\sigma_{l_1l_2}^2 \quad \text{and} \quad V(k) := p^{-1}\sum_{|l_1-l_2|\le k}\sigma_{l_1l_1}\sigma_{l_2l_2}.$$

Note that,

$$\sum_{|l_1-l_2|>k}\sigma_{l_1l_2}^2 = 2\sum_{q=k+1}^{p-1}(p-q)h(q) \quad \text{and} \quad \sum_{|l_1-l_2|\le k}\sigma_{l_1l_1}\sigma_{l_2l_2} = g(0)+2\sum_{q=1}^{k} g(q),$$

where $g(q) := \sum_{l=1}^{p-q}\sigma_{ll}\sigma_{l+q\,l+q}$. Define estimators of $h(q)$ and $g(q)$:

$$\begin{aligned}\hat{h}(q) &= (p-q)^{-1}\sum_{l=1}^{p-q}\Big\{\frac{1}{P_n^2}\sum_{i,j}^{*}(X_{il}X_{i\ l+q})(X_{jl}X_{j\ l+q}) - 2\frac{1}{P_n^3}\sum_{i,j,k}^{*}X_{il}X_{k\ l+q}(X_{jl}X_{j\ l+q}) \\ &\quad +\frac{1}{P_n^4}\sum_{i,j,k,m}^{*}X_{il}X_{j\ l+q}X_{kl}X_{m\ l+q}\Big\} \quad \text{and} \\ \hat{g}(q) &= \sum_{l=1}^{p-q}\Big\{\frac{1}{P_n^2}\sum_{i,j}^{*}X_{il}^2X_{j\,l+q}^2 - \frac{1}{P_n^3}\sum_{i,j,k}^{*}\big(X_{il}X_{kl}X_{j\,l+q}^2 + X_{i\,l+q}X_{k\,l+q}X_{jl}^2\big) \\ &\quad +\frac{1}{P_n^4}\sum_{i,j,k,m}^{*}X_{il}X_{j\,l+q}X_{kl}X_{m\,l+q}\Big\},\end{aligned}$$

where $\sum^{*}$ denotes summation over different subscripts and $P_n^b = n!/(n-b)!$. These two estimators are linear combinations of U-statistics of different orders with the first term being the dominating term, respectively. Let $\hat{W}(k) := 2p^{-1}\sum_{q=k+1}^{p-1}(p-q)\hat{h}(q)$ and $\hat{V}(k) := p^{-1}\{\hat{g}(0)+2\sum_{q=1}^{k}\hat{g}(q)\}$, which are unbiased estimators of $W(k)$ and $V(k)$, respectively. Then, an unbiased estimator of $M_n(k)$ is

$$\hat{M}_n(k) := \hat{W}(k) + n^{-1}\hat{V}(k). \tag{4.1}$$

As Lemmas 2 and 3 indicate $\tilde{k}_B \in [k_{a,n}, k_{b,n}]$ and $\tilde{k}_B/k_B \to 1$, k_B can be estimated by

$$\hat{k}_B = \underset{k_{1,n}\le k\le k_{2,n}}{\operatorname{argmin}}\ \hat{M}_n(k) \tag{4.2}$$

where $[k_{1,n}, k_{2,n}]$ constitutes a range for the minimization. In light of the analysis given in the previous section, we may choose $k_{1,n} = \lfloor k_{a,n}/r_1\rfloor$ and $k_{2,n} = \min\{r_2k_{b,n}, n\}$ for some positive constants r_1 and $r_2 \ge 1$. Although $k_{a,n}$ and $k_{b,n}$ are unknown, they can be estimated via $\hat{h}(q)$ and the largest and smallest marginal sample variances, $\hat{\sigma}_{(1)}$ and $\hat{\sigma}_{(p)}$, respectively. Then, the estimates of $k_{a,n}$ and $k_{b,n}$ are

$$\hat{k}_{a,n} = \min\{k : \hat{a}n^{-1} - \hat{h}(k) > 0\} - 1 \quad \text{and} \quad \hat{k}_{b,n} = \max\{k : \hat{b}n^{-1} - \hat{h}(k) < 0\},$$

where $\hat{a} = 2\hat{\sigma}_{(p)}^2$ and $\hat{b} = \hat{\sigma}_{(1)}^2/2$. Accordingly, we can choose $\hat{k}_{1,n} = \lfloor \hat{k}_{a,n}/r_1\rfloor$ and $\hat{k}_{2,n} = \min\{r_2\hat{k}_{b,n}, n\}$ upon given r_1 and $r_2 \ge 1$. In practice, we may choose $r_1 = r_2 = 2$.

Alternatively, we can minimize $\hat{M}_n(k)$ over a more conservative interval $[0, n]$ so that

$$\hat{k}_B = \underset{0\le k\le n}{\operatorname{argmin}}\ \hat{M}_n(k), \tag{4.3}$$

by making the relationship between n and p more restrictive.

Theorem 1. *Under Assumptions 1 and 2, (3.6), if $\Sigma \in \mathcal{G}(\nu, q_p^0)$ and $(k_{b,n} - k_{a,n})/k_{a,n} \le C$, then for $\hat{k}_B$ given in (4.2), $\hat{k}_B/\tilde{k}_B \xrightarrow{p} 1$ as $n \to \infty$.*

As $\tilde{k}_B/k_B \to 1$ under (3.6), Theorem 1 implies that $\hat{k}_B$ is a ratioly consistent estimator of k_B. The same ratio consistency result can be established for the band width estimator (4.3) under Assumption 2, (3.6) and $n = O(p^{1/3})$. The latter is more restrictive than Assumption 1. That $(k_{b,n} - k_{a,n})/k_{a,n} \le C$ assumed in Theorem 1 implies that $k_{a,n}$ and $k_{b,n}$ are of the same order. Derivations leading to (A.5) and (A.8) in the Appendix show that it is satisfied under both the exponential and polynomial decays of $h(k)$.

In the following, we evaluate the estimation error of $\hat{k}_B$ to $\tilde{k}_B$ by providing a lower bound on the probability of $\tilde{k}_B$ being included in an interval around $\hat{k}_B$. To this end, we need a condition on the behavior of $M_n(k)$ in additional to (3.6).

Assumption 3. *There exist a constant $\gamma \ge 1$ and an integer $\tau \ge 1$ such that for any small $\delta > 0$, any $\tau < \eta < \delta\tilde{k}_B$ and n large enough*

$$\inf_{k\in\mathcal{J}_\eta} \{M_n(k) - M_n(\tilde{k}_B)\} \ge C\eta n^{-\gamma}, \tag{4.4}$$

where $\mathcal{J}_\eta = \{k : \eta \le |k - \tilde{k}_B| < 2\eta\} \cap [k_{a,n}, k_{b,n}]$.

While (3.6) dictates that the absolute deviation between $\tilde{k}_B$ and any k outside $(\tilde{k}_B(1-\delta), \tilde{k}_B(1+\delta))$ is at least a constant multiple of $n^{-1}\tilde{k}_B$, (4.4) prescribes that the deviation between $\tilde{k}_B$ and k inside $(\tilde{k}_B(1-\delta), \tilde{k}_B(1+\delta))$ is at least $|\tilde{k}_B - k|n^{-\gamma}$ for $\gamma \ge 1$, which is much smaller than $n^{-1}\tilde{k}_B$.

Denote $C_{1,p}(k) = \{2(p-k)\}^{-1}\sum_{|l_1-l_2|=k} \sigma_{l_1l_1}\sigma_{l_2l_2}$. In the following, we show that Assumption 3 is satisfied for both the exponential and polynomial decay of $h(k)$, whose proof is in the Supplementary Material.

Proposition 2. *For $\Sigma \in \mathcal{G}(\nu, q_p^0)$, (i) if $h(q) = C_{2,p}(q)\theta^{-q}$ for $\theta > 1$, and $\max_{q\in[k_{a,n},k_{b,n}]} |C_{i,p}(q) - C_i| \to 0$ as $n \to \infty$ for $i = 1, 2$, then Assumption 3 holds for $\tau = 1$ and $\gamma = 1$;*

(ii) if $h(q) = C_{2,p}(q)q^{-\beta}$ for $\beta > 1$ and $\max_{q\in[k_{a,n},k_{b,n}]} |C_{i,p}(q) - C_i| = o(n^{-1/\beta})$ as $n \to \infty$ for $i = 1, 2$, then Assumption 3 holds for $\tau = 1$ and $\gamma = 1 + 1/\beta$.

Theorem 2. *Under Assumptions 1, 2, 3 and (3.6), if* $\Sigma \in \mathcal{G}(\nu, q_p^0)$, $(k_{b,n} - k_{a,n})/k_{a,n} \le C$ *and* $\log(k_{2,n}) \sum_{q > k_{1,n}} h(q) = o(1)$, *then* $P(\tilde{k}_B \in [\hat{k}_B - \tau, \hat{k}_B + \tau]) = 1 - o(n^{2\gamma-1}p^{-1})$.

The proof of Theorem 2 is given in the Supplementary Material. Recall that $k_{1,n} = \lfloor k_{a,n}/r_1 \rfloor$ and $k_{2,n} = \min\{r_2 k_{b,n}, n\}$ for $r_1, r_2 \ge 1$. Derivations given in (A.7) and (A.9) show that $\log(k_{2,n}) \sum_{q > k_{1,n}} h(q) = o(1)$ under both the exponential and polynomial decays respectively for any positive constants r_1 and r_2. Since τ is usually unknown, $[\hat{k}_B - \tau, \hat{k}_B + \tau]$ is not a confidence interval of $\tilde{k}_B$. We may call it a concentration interval. Theorem 2 shows that the probability that $\tilde{k}_B$ is included in the interval converges to 1 if $n^{2\gamma-1}p^{-1}$ is bounded from infinity. For Gaussian data, $\Delta = 0$ and $k_B = \tilde{k}_B$. Hence, the concentration interval is also the one for k_B.

5 Extension to Tapering Estimation

The analysis we have made for the banding estimator can be extended to the tapering estimator of Cai et al. (2010). The underlying band width for the tapering estimator $T_k(S_n)$ can be defined via the standardized squared Frobenius norm $p^{-1}||T_k(S_n) - \Sigma||_F^2$. It can be verified that

$$\begin{aligned} & p^{-1}||T_k(S_n) - \Sigma||_F^2 \qquad (5.1) \\ = \; & p^{-1}\Big\{ \sum_{|l_1 - l_2| \le k} (\hat{\sigma}_{l_1 l_2} - \sigma_{l_1 l_2})^2 + \sum_{|l_1 - l_2| > 2k} \sigma_{l_1 l_2}^2 + \sum_{k < |l_1 - l_2| \le 2k} (\omega_{l_1 l_2}\hat{\sigma}_{l_1 l_2} - \sigma_{l_1 l_2})^2 \Big\}. \end{aligned}$$

Taking the expectation, the risk of the tapering estimation is

$$\widetilde{\mathrm{Obj}}_T(k) = p^{-1}\mathrm{E}\{||T_k(S_n) - \Sigma||_F^2\} = (np)^{-1}\mathrm{tr}(\Sigma^2) + (1 - 1/n)\,\mathrm{Obj}_T(k),$$

where

$$\begin{aligned} \mathrm{Obj}_T(k) &= N_n(k) + \Delta(np)^{-1}(1 - 1/n)\Big(\sum_{|l_1 - l_2| \le k} f_{l_1 l_2} + \sum_{k < |l_1 - l_2| \le 2k} \omega_{l_1 l_2}^2 f_{l_1 l_2} \Big) \text{ and} \\ N_n(k) &= \frac{1}{p} \sum_{|l_1 - l_2| > 2k} \sigma_{l_1 l_2}^2 + \frac{1}{np} \sum_{|l_1 - l_2| \le k} \sigma_{l_1 l_1}\sigma_{l_2 l_2} + \frac{1}{p} \sum_{k < |l_1 - l_2| \le 2k} (1 - \omega_{l_1 l_2})^2 \sigma_{l_1 l_2}^2 \\ &+ \frac{1}{np} \sum_{k < |l_1 - l_2| \le 2k} \omega_{l_1 l_2}^2 \sigma_{l_1 l_1}\sigma_{l_2 l_2}. \end{aligned}$$

The underlying band width of the tapering estimator is $k_T = \min\{k' | k' = \underset{0 \le k < p/2}{\mathrm{argmin}}\, \widetilde{\mathrm{Obj}}_T(k)\}$. Similar to the banding estimator, the minimizer of $\widetilde{\mathrm{Obj}}_T(k)$ is equivalent to that of

$\mathrm{Obj}_T(k)$. Just like $M_n(k)$ is the dominant term of $\mathrm{Obj}_B(k)$, $N_n(k)$ dominates $\mathrm{Obj}_T(k)$ and the minimization of $\mathrm{Obj}_T(k)$ can be carried out by minimizing $N_n(k)$.

Denote ω_q to be the tapering weight for $|l_1 - l_2| = q$. Utilizing $\hat{h}(q)$ and $\hat{g}(q)$ in the previous section, we define $\widetilde{W}(k) := 2p^{-1}\sum_{q=k+1}^{2k}(1-\omega_q)^2(p-q)\hat{h}(q)$ and $\widetilde{V}(k) := 2p^{-1}\sum_{q=k+1}^{2k}\omega_q^2\hat{g}(q)$. An unbiased estimator of $N_n(k)$ is

$$\hat{N}_n(k) := \hat{W}(2k) + \widetilde{W}(k) + n^{-1}\{\hat{V}(k) + \widetilde{V}(k)\}, \tag{5.2}$$

where $\hat{W}(2k)$ and $\hat{V}(k)$ are estimators used in the estimation of $M_n(k)$ for the banding estimation. The proposed estimator for k_T is

$$\hat{k}_T = \underset{0\le 2k\le n}{\mathrm{argmin}}\ \hat{N}_n(k) \tag{5.3}$$

by noting that the tapering estimator used $2k$ as the effective band width. Denote $\tilde{k}_T$ to be the smallest minimizer of $N_n(k)$. An analysis on the band widths k_T and $\tilde{k}_T$ may be carried out in a similar fashion to what we have done for k_B and $\tilde{k}_B$ for the banding estimator. The ratio convergence of $\hat{k}_T$ to k_T may be established under certain conditions. We will evaluate the empirical performance of $\hat{k}_T$ in the simulations and the case study in Sections 7 and 8.

6 Extension to Thresholding Estimation

Both the banding and tapering estimators require the variables in X having a natural ordering such that the correlation decays as two variables are further apart. For covariances not satisfying such ordering, Bickel and Levina (2008b) proposed the thresholding estimator under the following covariance class:

$$\mathfrak{V}(q, c_0(p), M) = \left\{\Sigma : \sigma_{l_1 l_1} \le M,\ \sum_{l_2=1}^{p} |\sigma_{l_1 l_2}|^q \le c_0(p), \quad \text{for all } l_1\right\} \tag{6.1}$$

for a $q \in (0,1)$ and some positive function $c_0(p)$. For any $p \times p$ matrix $M = (m_{l_1 l_2})_{p\times p}$, the thresholding operator is

$$D_s(M) = (m_{l_1 l_2}\mathrm{I}\{|m_{l_1 l_2}| \ge s\})_{p\times p}$$

with a threshold level s. Bickel and Levina (2008b) proposed $D_{t_n}(S_n)$ as an estimator of Σ, where $t_n = \sqrt{2t(\log p)/n}$ for a positive threshold parameter t, and showed that, if $(\log p)/n = o(1)$,

$$||D_{t_n}(S_n) - \Sigma||_{(2,2)} = O\big\{c_0(p)(\log(p)/n)^{(1-q)/2}\big\}. \tag{6.2}$$

See Rothman, Levina and Zhu (2009) and Cai and Liu (2011) for related studies.

The Frobenius risk function for the thresholding estimator can be explicitly expressed, as shown in the following proposition. Let $\phi(\cdot)$ and $\bar{\Phi}(\cdot)$ be the standard normal density and upper tail probability functions, respectively, and $\widetilde{\mathrm{Obj}}_D(t,\Sigma) = \mathrm{E}\{||D_{t_n}(S_n)-\Sigma||_F^2\}$.

Proposition 3. *Suppose* $\log p = o(n^{1/3})$ *and for any* $1 \leq l \leq p$, *there exists a positive constant* H_l *such that* $\mathrm{E}\big[\exp\{t(X_{1l}-\mu_l)^2\}\big] < \infty$ *when* $|t| < H_l$, *then, for any* $\Sigma = (\sigma_{l_1 l_2})_{p\times p}$, $\widetilde{\mathrm{Obj}}_D(t,\Sigma) = \mathrm{Obj}_D(t,\Sigma)\big(1+o(1)\big)$, *where*

$$\begin{aligned}\mathrm{Obj}_D(t,\Sigma) = \sum_{l_1,l_2=1}^{p} \Bigg\{ & \frac{g_{l_1l_2}^2}{n}[\eta_{l_1l_2}^{(1)}\phi(\eta_{l_1l_2}^{(1)}) + \bar{\Phi}(\eta_{l_1l_2}^{(1)}) + \eta_{l_1l_2}^{(2)}\phi(\eta_{l_1l_2}^{(2)}) + \bar{\Phi}(\eta_{l_1l_2}^{(2)})] \\ & + \sigma_{l_1l_2}^2[\bar{\Phi}(-\eta_{l_1l_2}^{(1)}) - \bar{\Phi}(\eta_{l_1l_2}^{(2)})]\Bigg\},\end{aligned} \tag{6.3}$$

$\eta_{l_1l_2}^{(1)} = \sqrt{n}(t_n - \sigma_{l_1l_2})/g_{l_1l_2}$, $\eta_{l_1l_2}^{(2)} = \sqrt{n}(t_n + \sigma_{l_1l_2})/g_{l_1l_2}$ *and* $g_{l_1l_2}^2 = \mathrm{Var}\{(X_{1l_1}-\mu_{l_1})(X_{1l_2} - \mu_{l_2})\}$.

The proof is given in the Supplementary Material. The sub-gaussian condition in the theorem is required to utilize the moderate deviation results. However, if a standardization is used so that $s_{ij} = \sum_{l=1}^{n}(X_{li}-\mu_i)(X_{lj}-\mu_j)/n$ is used to estimate the underlying marginal variance as in Cai and Liu (2011), the sub-Gaussian assumption can be relaxed. The standardization allows moderate deviation results for self-normalized statistics, which requires less assumption as shown in Jing, Shao and Wang (2003).

From Proposition 3, it is seen that $\mathrm{Obj}_D(t,\Sigma)$ is the leading order term of $\widetilde{\mathrm{Obj}}_D(t,\Sigma)$. We use $\mathrm{Obj}_D(t,\Sigma)$ as a substitute of $\widetilde{\mathrm{Obj}}_D(t,\Sigma)$. Under Assumption 2, it can be shown that $g_{l_1l_2}^2 = \sigma_{l_1l_1}\sigma_{l_2l_2} + \sigma_{l_1l_2}^2 + \Delta f_{l_1l_2}$. For simplicity, we focus on the normally distributed data in this section such that $\Delta = 0$ and $g_{l_1l_2}^2 = \sigma_{l_1l_1}\sigma_{l_2l_2} + \sigma_{l_1l_2}^2$. Therefore, in order to estimate $g_{l_1l_2}^2$, it is suffice to estimate $\sigma_{l_1l_2}$.

Note that $\eta_{l_1l_2}^{(1)}$, $\eta_{l_1l_2}^{(2)}$ and t_n are continuous and differentiable functions. So, $\mathrm{Obj}_D(t,\Sigma)$ is continuous and differentiable with respect to t. Therefore, the minimum of $\mathrm{Obj}_D(t,\Sigma)$ exists on any closed interval $[0,B]$ for $B>0$. Define the underlying threshold level as

$$t_0(\Sigma) = \underset{t\in[0,B]}{\arg\min}\ \mathrm{Obj}_D(t,\Sigma) \tag{6.4}$$

Before we present an algorithm to find an estimate of $t_0(\Sigma)$, we review the cross validation (CV) approach proposed in Bickel and Levina (2008b), which was designed to approximate the Frobenius risk $\widetilde{\mathrm{Obj}}_D(t,\Sigma)$. They proposed splitting the original sample into two groups of size n_1 and n_2 randomly for N times. In the vth split, let

S_1^v and S_2^v be the sample covariances based on the two sub-samples, respectively. The estimated Frobenius risk with respect to t is

$$\hat{R}_D(t) = \frac{1}{N}\sum_{v=1}^{N} ||D_{t_n}(S_1^v) - S_2^v||_F^2 \tag{6.5}$$

and the estimated threshold level is

$$\hat{t}_{BL} = \arg\min_{t\in[0,B]} \hat{R}_D(t). \tag{6.6}$$

Similar approach has been used in Bickel and Levina (2008a) to select the band width for the banding estimator, and in Cai and Liu (2011) for the adaptive thresholding estimator. Due to the inconsistence of S_2^v, $\hat{R}_D(t)$ is unreliable for $\widetilde{\text{Obj}}_D(t,\Sigma)$, which may result in unstable threshold selection as revealed in our simulation study.

We propose an iterative procedure for selecting the threshold level t which makes use of the derived expressions for the Frobenius risk in Proposition 3. We use $\widehat{\text{Obj}}_D(t, D_{\hat{t}_{n,BL}}(S_n))$ for $\hat{t}_{n,BL} = \sqrt{2\hat{t}_{BL}(\log p)/n}$ as an initial estimate of $\text{Obj}_D(t,\Sigma)$ where $D_{\hat{t}_{n,BL}}(S_n)$ is the thresholding estimator of Σ with the Bickel and Levina's threshold selector $\hat{t}_{BL}$. In the computation of $\widehat{\text{Obj}}_D(t, D_{\hat{t}_{n,BL}}(S_n))$, all the $g_{l_1l_2}$, $\eta_{l_1l_2}^{(1)}$ and $\eta_{l_1l_2}^{(2)}$ appeared in (6.3) are replaced by their estimates implied under $D_{\hat{t}_{n,BL}}(S_n)$. Then, the selected threshold level in the first iteration is

$$\hat{t}_1 = \arg\min_{t\in[0,B]} \widehat{\text{Obj}}_D(t, D_{\hat{t}_{n,BL}}(S_n)), \tag{6.7}$$

which may be viewed as a refinement of Bickel and Levina's approach.

Having acquired the $\hat{t}_{h-1}$ for a $h \geq 1$, the hth iterative threshold estimator is

$$\hat{t}_h = \arg\min_{t\in[0,B]} \widehat{\text{Obj}}_D(t, D_{\hat{t}_{n,h-1}}(S_n)), \tag{6.8}$$

where $\hat{t}_{n,h-1} = \sqrt{2\hat{t}_{h-1}(\log p)/n}$. Simulations given in the next section demonstrate that the algorithm tends to converge within five iterations and had superior performance over Bickel and Levina's CV method.

7 Simulation Results

We report results of simulation studies which were designed to evaluate the empirical performance of the proposed band width and threshold estimators for the banding, tapering and thresholding covariance estimators. We also compared with the cross-validation estimator of Bickel and Levina (2008a,b) and SURE of Yi and Zou (2013).

IID p-dimensional random vectors were generated according to

$$X_i = \Sigma^{\frac{1}{2}} Z_i, \tag{7.1}$$

where $Z_i = (Z_{i1}, \cdots, Z_{ip})'$ and the innovations $\{Z_{ij}\}_{j=1}^{p}$ were IID from (i) $N(0,1)$ and (ii) the standardized t-distribution with degree of freedom 5 (t_5) so that they have zero mean and unit variance. For the tapering estimation, we compared the proposed band width estimator with SURE of Yi and Zou (2013) for $N(0,1)$, and the standardized $Gamma(1, 0.5)$, $Gamma(0.5, 1)$, $Gamma(0.3, 1)$ and $Gamma(0.1, 1)$ distributed innovations, which correspond to the excess kurtosis Δ being 0, 6, 12, 20 and 60, respectively.

Two designs of covariance structures for $\Sigma = (\sigma_{l_1 l_2})_{p\times p}$ were considered

$$\begin{aligned} &\text{(A):} \quad \sigma_{l_1 l_2} = \theta^{-|l_1 - l_2|} \ \text{ for } \theta > 1; \\ &\text{(B):} \quad \sigma_{l_1 l_2} = \mathbb{I}(l_1 = l_2) + \xi |l_1 - l_2|^{-\beta} \mathbb{I}(l_1 \neq l_2) \ \text{ for } \xi \in (0,1) \text{ and } \beta > 1, \end{aligned} \tag{7.2}$$

which prescribe the exponential and polynomial decay, respectively. In the simulation, we chose $\theta = 0.7^{-1}, 0.9^{-1}$, and $\xi = 0.5$ and $\beta = 1.5$, respectively.

We also considered a covariance structure to confirm the discussion made regarding the unequal diagonal entries associated with the polynomial decay in Section 3. Specifically, let $\{\sigma_{ll}\}_{l=(h-1)p'+1}^{hp'} \overset{iid}{\sim} \chi_h^2$ for $h = 1, \cdots, 10$, and $p' = p/10$. Let $\Lambda = \text{diag}(\sigma_{11}^{1/2}, \cdots, \sigma_{pp}^{1/2})$. The third design (Design (C)) of Σ was

$$\begin{aligned} &\Sigma = \Lambda \Psi \Lambda \ \text{ and } \Psi = (\rho_{l_1 l_2}) \text{ with} \\ &\rho_{l_1 l_2} = \mathbb{I}(l_1 = l_2) + 0.5|l_1 - l_2|^{-1.5} \mathbb{I}(l_1 \neq l_2). \end{aligned} \tag{7.3}$$

The random generation of the diagonal elements made the column series $\{X_{i1}, \cdots, X_{ip}\}$ under Design (C) non-stationary. Similar design was considered in Cai et al. (2013).

When evaluating the thresholding estimator, the normally distributed data were generated for the covariance structure (A) in (7.2) with $\theta = 0.7^{-1}$ and 0.9^{-1}, as well as a block diagonal covariance (Design (D)):

$$\begin{aligned} &\Sigma_{p\times p} = \text{diag}(\Sigma^{(1)}_{p/2\times p/2}, \Sigma^{(2)}_{p/2\times p/2}) \ \text{ where } \Sigma^{(1)} \text{ and } \Sigma^{(1)} \text{ follow} \\ &\text{structure (A) with } \theta = 0.3^{-1} \text{ and } 0.9^{-1}, \text{ respectively.} \end{aligned} \tag{7.4}$$

To mimic the "large p, small n" paradigm, we chose $n = 40, 60$ and $p = 40, 200, 400$ and 1000, respectively. We considered the more conservative band width estimator in

(4.3) that has a wider span of search region. For the banding estimation, comparison has been made with the cross-validation approach of Bickel and Levina (2008a,b). Similar to (6.5), the empirically estimated Frobenius risk with respect to the band width k is

$$\hat{R}(k) = \frac{1}{N} \sum_{v=1}^{N} ||B_k(\hat{\Sigma}_1^v) - \hat{\Sigma}_2^v||_F^2 \tag{7.5}$$

and the band width estimator is $\hat{k}_{BL} = \arg\min_k \hat{R}(k)$. According to Bickel and Levina (2008b), we chose $n_1 = n(1 - 1/\log n)$ and the number of random splits $N = 50$. We choose $B = 2.5$ in (6.6), (6.7) and (6.8) in the algorithm for the threshold levels. All the simulation results reported in this section were based on 500 replications.

Tables 1, 2 and 3 report averages and standard deviations of the proposed band width estimators for both the banding and the tapering estimation, and those of Bickel and Levina (2008a) (BL)'s CV band width estimator, under both the Gaussian and the standardized t_5 innovations with the covariances (A), (B) and (C) in (7.2) and (7.3).

It is observed from Tables 1 and 2 that the proposed band width had smaller bias and standard deviation than those of the Bickel and Levina's CV estimators for almost all the cases in the simulations. The bias and standard deviation of the proposed band width selector were consistently less than 0.5 for larger p, which may be viewed as confirmatory to the finding in Theorem 2 that the underlying band widths are within $\mathcal{O}_1 = [\hat{k}_B - 1, \hat{k}_B + 1]$ with overwhelming probability. It is also observed that as p was increased, both the bias and the standard deviation of the proposed band width estimator were reduced. This was not necessarily the case for the CV band width selector.

Comparing the results of the band widths for the banding and the tapering estimators in Table 1 and 3 under Design (A), we found that the underlying k_B and k_T were more responsive to the increase of the sample size n than to the increase of the dimension p. This may be understood by the fact that the penalty term $(np)^{-1} \sum_{|l_1 - l_2| \le k} \sigma_{l_1 l_1} \sigma_{l_2 l_2}$ in the objective function decreases as n is increased. Although there is a division of p in the penalty term, it is absorbed as part of the averaging process. As a result, the underlying band widths were not sensitive to p upon given a particular covariance design. Under both the standardized normal and t_5 innovations, it was found that $k_B = \tilde{k}_B$ and $k_T = \tilde{k}_T$ for all the (p, n) combinations under the covariance Designs (A)-(C). This was not necessarily the case for more skewed data, for instance the standardized $Gamma(0.1, 1)$ innovation (Figure 3).

Table 4 reports the average and the standard deviations of the selected threshold levels by the proposed iterative approach and Bickel and Levina (2008b)'s CV method.

It shows that the selected threshold level from the first iteration was already better than the CV method for having smaller bias and being less variable. The second iteration improved those of the first significantly, and the improvement continued as the iteration went. A convergence was largely established within five iterations.

In addition to evaluate the performance of the band width estimation, we also computed the estimation loss for Σ with the estimated band widths, and Bickel and Levina's (BL) as well as Cai and Yuan (2012)'s (CY) adaptive blocking estimation. Let $\hat{\Sigma}_{\hat{k}_B}$ and $\hat{\Sigma}_{\hat{k}_T}$ be the banding and the tapering estimators with the proposed band width selection, respectively; and $\hat{\Sigma}_{\hat{k}_{BL}}$ and $\hat{\Sigma}_{CY}$ be the banding estimator with BL's band width selection and Cai and Yuan's adaptive blocking estimation, respectively. For each of the covariance estimators, say $\hat{\Sigma}$, we gathered the spectral loss $||\hat{\Sigma} - \Sigma||_{(2,2)}$ and the Frobenius loss $||\hat{\Sigma} - \Sigma||_F$. Figure 1 displays the box plots of the estimation losses under Design (A) with $\theta = 0.7^{-1}$, Design (B) with $\xi = 0.5$ and $\beta = 1.5$ and the Gaussian innovations.

We observe from Figures 1 that under the spectral norm, the estimation losses of $\hat{\Sigma}_{\hat{k}_{BL}}$ encountered large variance under both the spectral and Frobenius norms, which was likely caused by the large variation of the BL's band width estimator shown in Tables 1 and 2. The estimation errors of $\hat{\Sigma}_{CY}$ were quite large in terms of the Frobenius norm. While its relative performance was improved under the spectral norm, the errors were still larger than those of the banding and tapering estimators with the proposed band width selection methods under the covariance Designs (A) and (B). We observe a significant advantage of the covariance estimation with the proposed band width selection method. In particular, the losses of the banding and the tapering estimators with the proposed band widths were substantially less than those of $\hat{\Sigma}_{CY}$ and $\hat{\Sigma}_{\hat{k}_{BL}}$. Although $\hat{\Sigma}_{\hat{k}_{BL}}$'s median loss was less than that of $\hat{\Sigma}_{CY}$ in most cases, it was much more variable. In contrast, the banding and the tapering estimation with the proposed band widths had the smallest medians and variation. We also observe that the estimation loss of the tapering estimator was smaller than that of the banding estimator under Design (A). This is due to that the $h(k)$ function decays gradually as the band width k was increased. Therefore, the tapering estimator fits these covariance structures better than the banding estimator. However, under Design (B), the advantage of the tapering estimator over the banding estimator was much reduced.

We also compared the proposed method (4.3) with the fixed and the change-point methods of Qiu and Chen (2012) designed for banded covariances. We considered $\hat{k}_{0.5,0.06}$

for the fixed estimator, and the change-point estimator was applied on band widths whose p-values for the banded test were larger than 10^{-10}. Figure 2 reports the bias and standard deviation of these three methods for covariance design (A) with the Gaussian distributed innovation. The covariance design prescribes an exponentially decaying off-diagonal with the speed of the decay controlled by θ. And the covariance under this regime is not banded but bandable. We observe that the performance of the proposed estimators was much more accurate than those of Qiu and Chen (2012), with much smaller bias and standard deviation. For the covariance design with $\theta = 0.7^{-1}$, both the fixed and the change-point estimators over-estimated the underlying band width. For the covariance design with $\theta = 0.9^{-1}$, we found dramatic under-estimation and over-estimation for the fixed and the change-point estimators, respectively. The inferior performance of Qiu and Chen's methods confirms that they are not suitable for "bandable" covariances.

The relative performance of the proposed band width selection for the tapering estimator to that of the SURE of Yi and Zou (2013) is displayed in Figure 3. The figure plots the differences in the absolute bias and the standard deviation between the SURE and the proposed band width selection under covariance Design (A) with $\theta = 0.7^{-1}$. The comparison was made under the Gaussian innovation ($\Delta = 0$), and the standardized Gamma innovations with $\Delta = 6, 12, 20$ and 60. We recall that Δ measures the excessive kurtosis over that of the Gaussian. We observed that the performance of SURE and the proposed were largely comparable for smaller Δ and larger n ($n = 60$). As Δ got larger so that the data deviate more from the Gaussian, the performance of SURE was adversely affected. The standard deviation and the bias of the proposed band width estimates were largely stable with respect to the changing Δ. It is noted that SURE is proposed under Gaussianity whereas the proposed band width estimation is largely nonparametric. This was the reason that the proposed method outperformed SURE for the Gamma distributed innovations.

8 Empirical Study

In this section, we reported an empirical study on a sonar spectrum data set by conducting the banding and tapering covariance estimation with the proposed band width selection methods. Gorman and Sejnowski (1988a and 1988b) and Yi and Zou (2013) had analyzed the same data, which are publicly available at the University of California

Irvine Machine Learning Repository. The data set collects the so-called sonar returns which are the amplitudes of bouncing signals off an object, essentially the return signal strength over time. The sonar returns were collected from bouncing signals off a metal cylinder and a cylindrically shaped rock, respectively positioned on a sandy ocean floor. The data set contains 208 returns, 111 of them from the metal cylinder and 97 from the rock. A data preprocessing based on the Fourier transform was applied to obtain the spectral envelope for each sonar return, and each spectral envelope composed of 60 numerical readings in the range 0.0 to 1.0, with each reading representing the energy within a particular frequency band. Hence, the data dimension $p = 60$, and there were two samples of sizes 111 and 97 respectively.

Gorman and Sejnowski (1988) analyzed the data set by the neural network, aiming to classify sonar targets to two groups. Yi and Zou (2013) found that there was a quite obvious decay among entries of the sample covariance along the off-diagonals. They estimated the covariance matrices for the metal and the rock groups by their SURE-tuned tapering estimation method. Their analysis suggested the effective band width of the tapering estimator to be 34 for the rock group.

We consider estimating the covariance matrices by the banding and tapering estimators with the proposed band width selection. The estimated $h(k)$ for the rock and metal groups are displayed in the upper panel of Figure 4, from which we see that $h(k)$ decays rapidly as the band width k increases, indicating potential bandable structure of the covariance. The estimated Frobenius loss $\hat{M}_n(k)$ and $\hat{N}_n(k)$ for both groups are displayed in the two lower panels of Figure 4 for both the banding and tapering estimators, respectively. These graphs showed that the band widths which minimize the Frobenius losses of the banding estimation were 26 and 37 for the rock group and metal group, respectively. These were quite different from the estimates of 35 and 44 for the two groups prescribed by the CV method of Bickel and Levina. For the tapering estimation, the proposed approach selected band widths of 17 and 25 for the two groups, and hence the effective band widths were 34 and 50 for the two groups, respectively. This respected the ordering that k_B is between k_T and $2k_T$. Although the SURE method produced similar band width estimates of 16 and 25 for the two groups, the CV method for the tapering estimation gave band widths 28 and 26, respectively. These again were sharply different from the band width estimates using the proposed method.

9 Discussion

Cai and Yuan (CY) (2012) proposed an adaptive covariance estimator through a block thresholding approach for the normally distributed data with the covariance matrix class $\mathfrak{U}_1(\alpha, C)$. They showed that such adaptive estimator can achieve the minimax convergence rate under the spectral norm. The approach of Cai and Yuan (2012) is "data-driven" up to the initial block size k_0 and a thresholding parameter λ, which were set to be $\lfloor \log p \rfloor$ and 6, respectively. The initial block size k_0 functions similarly as the band width in the banding and tapering estimation. While fixing the initial block size k_0 attains simplicity, it may be less responsive to the different underlying covariance structures.

The block thresholding estimator can attain the minimax rate of convergence, so can the tapering estimator of Cai et al. (2010). It is important and assuring to have minimax properties. However, the minimax rate tends to be less sensitive when the matrix class under consideration is large, for instance the $\mathfrak{U}_1(\alpha, C)$ class. As shown in Section 3, the rates of the underlying band width k_B, which minimizes the Frobenius risk for the banding estimation are quite responsive to the different forms of sparsity of Σ. Specifically, the exponential and polynomial decays lead to different rates for k_B. This responsive feature can produce less estimation error. Our simulation study showed that the banding and the tapering estimators with the proposed band widths outperformed the block thresholding estimator consistently under the Frobenius norm for all three covariance designs used in the simulation, which was also the case under the spectral norm for the covariance designs (A) and (B). For the third design of covariance (Design (C)), the performance of the CY's estimator was comparable to those of the banding and tapering estimators.

It can be shown that the banding estimation can also reach the minimax convergence rate under the Frobenius norm at k_B, the underlying band width that minimizes the Frobenius risk. Under the matrix class considered in Theorem 1, the difference between $\mathrm{Obj}_B(\hat{k}_B)$ and $\mathrm{Obj}_B(k_B)$ is negligible comparing to $\mathrm{Obj}_B(k_B)$, as revealed by Corollary 1 in the Supplementary Material. This leads to the belief that the banding estimation with the estimated band width $\hat{k}_B$ should also attain the minimax rate under the Frobenius norm for the matrix class $\mathcal{G}(\nu, q_p^0)$. Confirming this theoretically would be an interesting future research topic, given the limited space available for this paper.

Yi and Zou (2013) considered the band width selection for the tapering estimator for Gaussian distributed data. The proposed method is nonparametric so it is more widely applicable, which may explains the better performance of the proposed method for the case of the Gamma distributed innovations.

Appendix

Derivation of (3.2). Without loss of generality, we assume $\mu = 0$. The first term on the right hand side of (3.1) can be decomposed as

$$\sum_{|l_1-l_2|\le k} (\hat{\sigma}_{l_1l_2} - \sigma_{l_1l_2})^2 = A_1 + A_3 - 2A_2, \tag{A.1}$$

where $A_1 = \frac{1}{n^2}\sum_{|l_1-l_2|\le k}\sum_{i,j} X_{il_1}X_{il_2}X_{jl_1}X_{jl_2} - \frac{2}{n}\sum_{|l_1-l_2|\le k}\sum_i X_{il_1}X_{il_2}\sigma_{l_1l_2} + \sum_{|l_1-l_2|\le k}\sigma^2_{l_1l_2}$, $A_2 = \sum_{|l_1-l_2|\le k}\left(\frac{1}{n}\sum_{i=1}^n X_{il_1}X_{il_2} - \sigma_{l_1l_2}\right)\bar{X}_{l_1}\bar{X}_{l_2}$ and $A_3 = \sum_{|l_1-l_2|\le k}(\bar{X}_{l_1}\bar{X}_{l_2})^2$. For the first term in A_1, from Assumption 2, we have

$$\begin{aligned}
& \mathrm{E}\sum_{|l_1-l_2|\le k}\sum_{i,j} X_{il_1}X_{il_2}X_{jl_1}X_{jl_2} \\
= & \sum_{|l_1-l_2|\le k}\{\sum^{*}_{i,j}\mathrm{E}(X_{il_1}X_{il_2})\mathrm{E}(X_{jl_1}X_{jl_2}) + \sum_i \mathrm{E}(X^2_{il_1}X^2_{il_2})\} \\
= & \sum_{|l_1-l_2|\le k}\{n(n+1)\sigma^2_{l_1l_2} + \Delta n f_{l_1l_1l_2l_2} + n\sigma_{l_1l_1}\sigma_{l_2l_2}\}.
\end{aligned}$$

Note that $\mathrm{E}n^{-1}\big(\sum_i X_{il_1}X_{il_2}\big) = \sigma_{l_1l_2}$. By combining the three parts together,

$$\mathrm{E}(A_1) = n^{-1}\sum_{|l_1-l_2|\le k}(\sigma^2_{l_1l_2} + \sigma_{l_1l_1}\sigma_{l_2l_2} + \Delta f_{l_1l_2}).$$

Similarly, for A_2 and A_3, we have that

$$\begin{aligned}
\mathrm{E}(A_2) &= n^{-2}\sum_{|l_1-l_2|\le k}(\sigma^2_{l_1l_2} + \sigma_{l_1l_1}\sigma_{l_2l_2} + \Delta f_{l_1l_2}) \quad \text{and} \\
\mathrm{E}(A_3) &= n^{-2}\sum_{|l_1-l_2|\le k}(2\sigma^2_{l_1l_2} + \sigma_{l_1l_1}\sigma_{l_2l_2} + \Delta n^{-1} f_{l_1l_2}).
\end{aligned} \tag{A.2}$$

Substituting these into (A.1), we have from (3.1) that

$$\begin{aligned}
\widetilde{\mathrm{Obj}}_B(k) =& \frac{1}{np}\sum_{|l_1-l_2|\le k}(\sigma^2_{l_1l_2} + \sigma_{l_1l_1}\sigma_{l_2l_2} + \Delta f_{l_1l_2}) + \frac{1}{p}\sum_{|l_1-l_2|>k}\sigma^2_{l_1l_2} \\
& - \frac{1}{n^2p}\sum_{|l_1-l_2|\le k}\{\sigma_{l_1l_1}\sigma_{l_2l_2} + (2-n^{-1})\Delta f_{l_1l_2}\} \\
=& \frac{1}{np}\mathrm{tr}(\Sigma^2) + (1-n^{-1})M_n(k) + \frac{\Delta}{np}(1-n^{-1})^2\sum_{|l_1-l_2|\le k} f_{l_1l_2},
\end{aligned} \tag{A.3}$$

which leads to (3.2) with $M_n(k)$ being defined in (3.3). □

Rate of k_B under exponential decay sub-class. Suppose $h(q) = C(q)\theta^{-q}$ for $\theta > 1$ and $\{C(q)\}_{q=0}^{p-1} \in [C_1, C_2]$. Consider two equations:

$$a/n = C_1\theta^{-k} \quad \text{and} \quad b/n = C_2\theta^{-k}$$

which represent interceptions of two horizontal lines at a/n and b/n to the lower and upper bound functions of $h(k)$, respectively. The solutions for k are, respectively,

$$s_{a,n} = (\log n - \log a + \log C_1)/\log\theta \;\text{ and }\; s_{b,n} = (\log n - \log b + \log C_2)/\log\theta.$$

Note that for $q \leq s_{a,n}$, $a/n - h(q) \leq a/n - C_1\theta^{-q} \leq a/n - C_1\theta^{-s_{a,n}} = 0$. So, we have $k_{a,n} \geq s_{a,n}$. Similarly, for $q \geq s_{b,n}$, $b/n - h(q) \geq b/n - C_2\theta^{-s_{b,n}} = 0$, which implies that $k_{b,n} < s_{b,n}$. Therefore,

$$\tilde{k}_B,\; k_B \sim \log(n)/\log(\theta) \;\text{ and} \tag{A.4}$$

$$\frac{k_{b,n} - k_{a,n}}{k_{a,n}} \leq \frac{s_{b,n} - s_{a,n}}{s_{a,n}} = \frac{\log\{(aC_2)/(bC_1)\}}{\log n - \log a + \log C_1} \to 0. \tag{A.5}$$

Also, note that

$$h(k) \leq C_2\theta^{-k} \leq C_2\theta^{-s_{a,n}} = aC_2/(C_1 n) \tag{A.6}$$

for any $k \in [k_{a,n}, k_{b,n}]$. For any constant $r_1 \geq 1$ and $k_{1,n} = k_{a,n}/r_1$,

$$\sum_{q>k_{1,n}} h(q) \leq \sum_{q>s_{a,n}/r_1} C_2\theta^{-q} \leq C\theta^{-\frac{\log n}{r_1 \log\theta}} = Cn^{-1/r_1}. \;\square \tag{A.7}$$

Rate of k_B under polynomial decay sub-class. Suppose $h(q) = C(q)q^{-\beta}$ for $\beta > 1$ and $\{C(q)\}_{q=0}^{p-1} \in [C_1, C_2]$. Similar to the exponential decay sub-class, consider the equations: $a/n = C_1 q^{-\beta}$ and $b/n = C_2 q^{-\beta}$. And, their solutions are $s_{a,n} = (C_1 n/a)^{1/\beta}$ and $s_{b,n} = (C_2 n/b)^{1/\beta}$, respectively. Therefore, we have

$$\tilde{k}_B,\; k_B \sim n^{1/\beta} \;\text{ and }\; (k_{b,n} - k_{a,n})/k_{a,n} \leq \tilde{C} \tag{A.8}$$

for a positive constant $\tilde{C}$, and

$$\sum_{q>k_{1,n}} h(q) \leq \sum_{q>s_{a,n}/r_1} C_2 q^{-\beta} \leq C_2\{(C_1 n/a)^{1/\beta} r_1^{-1}\}^{1-\beta} = Cn^{(1-\beta)/\beta}, \tag{A.9}$$

for any constant $r_1 \geq 1$. □

To prove Theorem 1, first, we intend to calculate the variance of $(p-q)\hat{h}(q)$ and $\hat{g}(q)$. To this end, we introduce some notations. For $q = 0, \cdots, p-1$, define

$$\begin{aligned}
F_{1,q} &= \frac{1}{P_n^2} \sum_{l=1}^{p-q} \sum_{i,j}^{*} (X_{il}X_{i\,l+q})(X_{jl}X_{j\,l+q}),\\
F_{2,q} &= \frac{1}{P_n^3} \sum_{l=1}^{p-q} \sum_{i,j,k}^{*} X_{il}X_{k\,l+q}(X_{jl}X_{j\,l+q}),\\
F_{3,q} &= G_{3,q} = \frac{1}{P_n^4} \sum_{l=1}^{p-q} \sum_{i,j,k,m}^{*} X_{il}X_{j\,l+q}X_{kl}X_{m\,l+q},\\
G_{1,q} &= \frac{1}{P_n^2} \sum_{l=1}^{p-q} \sum_{i,j}^{*} X_{il}^2 X_{j\,l+q}^2 \quad \text{and}\\
G_{2,q} &= \frac{1}{P_n^3} \sum_{l=1}^{p-q} \sum_{i,j,k}^{*} \big(X_{il}X_{kl}X_{j\,l+q}^2 + X_{i\,l+q}X_{k\,l+q}X_{jl}^2\big).
\end{aligned}$$

Then, $\hat{W}(k) = 2p^{-1}\sum_{q=k+1}^{p-1}(F_{1,q} - 2F_{2,q} + F_{3,q})$ and $\hat{V}(k) = p^{-1}\{G_{1,0} - G_{2,0} + G_{3,0} + 2\sum_{q=1}^{k}(G_{1,q} - G_{2,q} + G_{3,q})\}$. The following lemma presents the variances of $F_{i,q}$ and $G_{i,q}$ for $i = 1, 2, 3$, whose proof can be found in the Supplementary Material.

Lemma A1. *Under Assumptions 2, if* $\lambda_{\max}(\Sigma) \le C < \infty$, *for any* $q = 0, \cdots, p-1$,

(i) $\mathrm{Var}(F_{1,q}) = O\{ph(q)n^{-1} + pn^{-2}\}$, $\mathrm{Var}(F_{2,q}) = O\{ph(q)n^{-2} + pn^{-3}\}$ *and* $\mathrm{Var}(F_{3,q}) = \mathrm{Var}(G_{3,q}) = O(pn^{-4})$;

(ii) $\mathrm{Var}(G_{1,q}) = O(pn^{-1})$ *and* $\mathrm{Var}(G_{2,q}) = O(pn^{-2})$.

Proof of Theorem 1. Let $S_0 = [k_{1,n}, k_{2,n}]$. For any $\delta > 0$ and every n, define $S_1 = \{k : |k - \tilde{k}_B| \ge \delta\tilde{k}_B\} \cap S_0$. Then, if $\hat{k}_B \in S_{1,n}$, we have $\sup_{k\in S_1}\{\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k)\} \ge 0$. It follows that,

$$P(|\hat{k}_B - \tilde{k}_B| \ge \delta\tilde{k}_B) = P(\hat{k}_B \in S_1) \le P[\sup_{k\in S_1}\{\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k)\} \ge 0].$$

For the term on the right side of the inequality, noting by (3.6), we have $\inf_{k\in S_1}\{M_n(k) - M_n(\tilde{k}_B)\} \ge C\tilde{k}_B n^{-1}$. Hence,

$$\begin{aligned}
&P[\sup_{k\in S_1}\{\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k)\} \ge 0]\\
\le &P[\sup_{k\in S_1}\{\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k) + M_n(k) - M_n(\tilde{k}_B)\} \ge C\tilde{k}_B n^{-1}].
\end{aligned} \tag{A.10}$$

Note that $\mathrm{E}\{\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k)\} = M_n(\tilde{k}_B) - M_n(k)$ and

$$\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k) = \frac{2}{np}\sum_{i=1}^{3}\sum_{q=k+1}^{\tilde{k}_B} G_{i,q} - \frac{2}{p}\sum_{i=1}^{3}\sum_{q=k+1}^{\tilde{k}_B} F_{i,q} \quad \text{for } k < \tilde{k}_B, \text{ and}$$

$$\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k) = \frac{2}{p}\sum_{i=1}^{3}\sum_{q=\tilde{k}_B+1}^{k} F_{i,q} - \frac{2}{np}\sum_{i=1}^{3}\sum_{q=\tilde{k}_B+1}^{k} G_{i,q} \quad \text{for } k > \tilde{k}_B.$$

By Lemma A1, it follows that

$$\begin{aligned}\mathrm{Var}\{\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k)\} \leq& C|k-\tilde{k}_B| \sum_{q\in[\tilde{k}_B,k]} \{(pn)^{-1}h(q) + p^{-1}n^{-2}\} \\ =& C\{|k-\tilde{k}_B|(pn)^{-1}o(1) + (k-\tilde{k}_B)^2p^{-1}n^{-2}\}.\end{aligned}$$

Therefore, by Chebyshev's inequality, the probability on the right side of (A.10) can be bounded by a constant times

$$\sum_{k\in S_1} \frac{(pn)^{-1}\{|k-\tilde{k}_B|o(1) + (k-\tilde{k}_B)^2n^{-1}\}}{\tilde{k}_B^2 n^{-2}} \leq C\sum_{k\in S_1}\{n(p\tilde{k}_B)^{-1}o(1) + p^{-1}\},$$

where the inequality above comes from the condition $(k_{b,n} - k_{a,n})/k_{a,n} \leq C$. Note that $|S_1| \leq C(k_{b,n} - k_{a,n})$ for a positive constant C. It follows that

$$\begin{aligned}& P[\sup_{k\in S_1}\{\hat{M}_n(\tilde{k}_B) - \hat{M}_n(k)\} \geq 0] \\ \leq\;& C(k_{b,n} - k_{a,n})\{n(p\tilde{k}_B)^{-1}o(1) + p^{-1}\} = O\{np^{-1}o(1) + k_{b,n}p^{-1}\}.\end{aligned}$$

Since $k_{b,n} = o(n)$, the last term in the inequality above is the small order term of np^{-1}. Noting that $n = O(p)$ by Assumption 1, we have $P(|\hat{k}_B - \tilde{k}_B| \geq \delta\tilde{k}_B) = o(n/p) \to 0$ for any $\delta > 0$, which leads to the conclusion that $\hat{k}_B/\tilde{k}_B \to 1$, as $n \to \infty$. □

References

[1] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

[2] Bai, Z. and Saranadasa, H. (1996). Effect of High Dimension: by an Example of a Two Sample Problem. *Statistica Sinica* **6** 311-329.

[3] Bai, Z.D., Silverstein, J.W. and Yin, Y.Q. (1998), A Note on the Largest Eigenvalue of a Large-dimensional Sample Covariance Matrix. *Journal of Multivariate Analysis* **26** 166-168.

[4] Bai, Z.D. and Yin, Y.Q. (1993), Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *The Annals of Probability* **21** 1276-1294.

[5] Bickel, P. J. and Levina, E. (2008a), Regularized Estimation of Large Covariance Matrices. *The Annals of Statistics* **36** 199-227.

[6] Bickel, P. and Levina, E. (2008b), Covariance Regularization by Thresholding. *The Annals of Statistics* **36** 2577-2604.

[7] Cai, T. T., Zhang, C.H. and Zhou, H. (2010), Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics* **38** 2118-2144.

[8] Cai, T. T. and Liu, W. (2011), Adaptive Thresholding for Sparse Covariance Matrix Estimation. *Journal of the American Statistical Association* **494** 672-684.

[9] Cai, T. T., Liu, W. and Luo, X. (2011), A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association* **494** 594-607.

[10] Cai, T. T., Liu, W. and Xia, Y. (2013), Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings. *Journal of the American Statistical Association* **501** 265-277.

[11] Cai, T. and Yuan, M. (2012), Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics* **40** 2014-2042.

[12] Gorman, R. and Sejnowski, T. (1988a), Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets. *Neural Networks* **1** 75-89.

[13] Gorman, R. and Sejnowski, T. (1988b), Learned Classification of Sonar Targets Using a Massively Parallel Network . *IEEE Transactions on Acoustics, Speech and Signal Processing* **36** 1135-1140.

[14] Hall, P. and Jin, J. (2010), Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise. *The Annals of Statistics* **38** 1686-1732.

[15] Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006), Covariance Matrix Selection and Estimation via Penalised Normal Likelihood. *Biometrika* **93** 85-98.

[16] Jing, B.Y., Shao, Q.M. and Wang, Q.Y. (2003), Self-normalized Cramer Type Large Deviations for Independent Random Variables. *The Annals of Probability* **31** 2167-2215.

[17] Johnstone, I. (2001), On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics* **29** 295-327.

[18] Levina, E., Rothman, A. and Zhu, J. (2008), Sparse Estimation of Large Covariance Matrices Via a Nested Lasso Penalty. *The Annals of Applied Statistics* **2** 245-263.

[19] Qiu, Y. and Chen, S. (2012), Test for Bandedness of High-dimensional Covariance Matrices and Bandwidth Estimation. *The Annals of Statistics* **40** 1285-1314.

[20] Rothman, A. J., Levina, E. and Zhu, J. (2009), Generalized Thresholding of Large Covariance Matrices. *Journal of the American Statistical Association* **104** 177-186.

[21] Rothman, A. J., Levina, E. and Zhu, J. (2010), A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539-550.

[22] Van de Vaart, A.W. (2000), *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[23] Wu, W. B. and Pourahmadi, M. (2003), Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data. *Biometrika* **93** 831-844.

[24] Xue, L. and Zou, H. (2012), Regularized Rank-based Estimation of High-dimensional Nonparanormal Graphical Models. *The Annals of Statistics* **40** 2541-2571.

[25] Yi, F. and Zou, H. (2013), SURE-tuned Tapering Estimation of Large Covariance Matrices. *Computational Statistics and Data Analysis* **58** 339-351.

Table 1: *Average and standard deviation in parentheses of the proposed band width estimators and Bickel and Levina's CV estimators (BL) for the banding estimation under the covariance Design (A) with* $\theta^{-1} = 0.7$ *and* $\theta^{-1} = 0.9$ *in (7.2) for the standard normal and standardized* t_5 *innovations.*

		Covariance (A) with $\theta^{-1} = 0.7$					
		Normal			t-distribution		
n	p	True	Proposed	BL	True	Proposed	BL
40	40	5	4.65(1.059)	4.40(1.616)	5	4.76(1.049)	4.76(2.010)
40	200	5	4.71(0.528)	5.06(2.206)	5	4.69(0.538)	5.28(2.716)
40	400	5	4.73(0.442)	5.43(2.516)	5	4.78(0.436)	5.97(3.409)
40	1000	5	4.87(0.332)	5.98(3.510)	5	4.86(0.361)	6.75(3.959)
60	40	5	5.35(1.169)	5.35(1.799)	5	5.34(1.187)	5.70(2.947)
60	200	5	5.26(0.483)	5.65(2.149)	5	5.23(0.465)	6.38(2.677)
60	400	5	5.16(0.372)	6.28(2.865)	5	5.17(0.379)	6.36(3.250)
60	1000	5	5.11(0.308)	6.93(3.695)	5	5.09(0.291)	7.52(3.869)
		Covariance (A) with $\theta^{-1} = 0.9$					
40	40	17	17.45(6.329)	17.99(8.154)	17	17.32(6.808)	24.06(8.613)
40	200	17	17.23(2.614)	16.37(4.970)	17	17.01(2.335)	16.49(5.506)
40	400	17	17.12(1.738)	16.38(5.316)	17	16.96(1.685)	16.14(4.985)
40	1000	17	17.02(1.084)	17.84(7.044)	17	16.99(1.005)	17.93(7.565)
60	40	19	19.58(6.894)	22.24(11.38)	19	19.16(6.228)	28.23(10.153)
60	200	19	19.01(2.750)	17.69(4.537)	19	19.08(2.633)	19.23(5.446)
60	400	19	19.05(1.766)	18.89(5.294)	19	19.04(1.876)	19.36(6.092)
60	1000	19	19.00(1.063)	19.93(6.696)	19	18.99(1.156)	20.56(6.901)

Table 2: *Average and standard deviation in parentheses of the proposed band width estimators and Bickel and Levina's CV estimators (BL) for the banding estimation under the covariance Design (B) and (C) with $\xi = 0.5$ and $\beta = 1.5$ in (7.2) for the standard normal and standardized t_5 innovations.*

		Covariance (B)					
		Normal			t-distribution		
n	p	True	Proposed	BL	True	Proposed	BL
40	40	2	1.70(0.585)	1.81(0.948)	2	1.79(0.618)	2.44(1.739)
40	200	2	1.88(0.330)	2.38(1.497)	2	1.89(0.318)	2.81(2.171)
40	400	2	1.96(0.200)	2.81(1.914)	2	1.96(0.196)	3.11(2.352)
40	1000	2	2.00(0.063)	3.33(2.611)	2	1.98(0.128)	4.68(3.956)
60	40	2	2.08(0.511)	2.16(1.079)	2	2.10(0.523)	2.49(1.826)
60	200	2	2.01(0.099)	2.59(1.406)	2	2.01(0.141)	3.15(2.147)
60	400	2	2.00(0.045)	2.98(1.949)	2	2.00(0)	4.21(3.042)
60	1000	2	2.00(0)	3.81(2.769)	2	2.00(0)	4.77(3.783)
		Covariance (C)					
40	40	2	1.74(0.842)	1.90(1.022)	2	1.77(0.873)	2.44(1.882)
40	200	2	1.76(0.462)	2.38(1.470)	2	1.75(0.473)	3.03(2.322)
40	400	2	1.85(0.369)	2.88(2.060)	2	1.88(0.337)	3.31(2.492)
40	1000	2	1.95(0.214)	3.23(2.360)	2	1.94(0.237)	4.15(3.444)
60	40	2	2.17(0.869)	2.23(1.185)	2	2.17(0.879)	2.85(2.025)
60	200	2	2.05(0.219)	2.74(1.570)	2	2.08(0.323)	3.34(2.673)
60	400	2	2.02(0.147)	3.02(1.874)	2	2.03(0.159)	3.77(2.752)
60	1000	2	2.00 (0)	3.79(2.666)	2	2.00(0.017)	3.92(2.912)

Table 3: *Average and standard deviation in parentheses of the proposed band width estimators for the tapering estimation under the covariance Design (A) with* $\theta^{-1} = 0.7$ *and* $\theta^{-1} = 0.9$*, (B) and (C) with* $\xi = 0.5$ *and* $\beta = 1.5$ *in (7.2) and (7.3) for the standard normal and standardized* t_5 *innovations.*

		Normal		t-distribution		Normal		t-distribution	
		Covariance (A) with $\theta^{-1} = 0.7$				Covariance (A) with $\theta^{-1} = 0.9$			
n	p	True	Proposed	True	Proposed	True	Proposed	True	Proposed
40	40	3	3.42(0.741)	3	3.49(0.791)	11	10.97(3.251)	11	10.82(3.364)
40	200	3	3.40(0.490)	3	3.34(0.476)	11	11.47(1.691)	11	11.33(1.492)
40	400	3	3.36(0.479)	3	3.33(0.471)	11	11.47(1.205)	11	11.32(1.136)
40	1000	3	3.27(0.442)	3	3.26(0.439)	11	11.37(0.744)	11	11.35(0.679)
60	40	4	3.83(0.789)	4	3.84(0.819)	12	12.26(3.595)	12	12.04(3.101)
60	200	4	3.90(0.342)	4	3.93(0.299)	13	12.61(1.757)	13	12.51(1.387)
60	400	4	3.97(0.176)	4	3.97(0.180)	13	12.66(1.209)	13	12.63(1.143)
60	1000	4	4.00(0)	4	4.00(0)	13	12.70(0.738)	13	12.68(0.765)
		Covariance (B)				Covariance (C)			
n	p	True	Proposed	True	Proposed	True	Proposed	True	Proposed
40	40	2	1.60(0.549)	2	1.68(0.516)	2	1.64(0.677)	2	1.64(0.659)
40	200	2	1.81(0.391)	2	1.82(0.385)	2	1.69(0.463)	2	1.71(0.456)
40	400	2	1.91(0.281)	2	1.90(0.301)	2	1.80(0.400)	2	1.82(0.383)
40	1000	2	1.98(0.140)	2	1.97(0.171)	2	1.92(0.272)	2	1.91(0.292)
60	40	2	1.94(0.403)	2	1.94(0.401)	2	1.93(0.618)	2	1.96(0.653)
60	200	2	2.00(0)	2	2.00(0.056)	2	1.99(0.155)	2	1.99(0.161)
60	400	2	2.00(0)	2	2.00(0)	2	2.00(0.045)	2	2.00(0)
60	1000	2	2.00(0)	2	2.00(0)	2	2.00(0)	2	2.00(0)

Table 4: *Empirical average and standard deviation in parentheses of the proposed threshold estimators and Bickel and Levina (BL)'s under the covariance Design (A) in (7.2) and (D) in (7.4) for the normal distributed data.*

		Covariance (A) with $\theta = 0.7^{-1}$				
n	p	True	BL	1st iteration	2nd iteration	5th iteration
40	40	0.64	0.92(0.177)	0.76(0.104)	0.70(0.090)	0.66(0.095)
40	200	0.86	1.28(0.078)	1.05(0.053)	0.98(0.049)	0.95(0.046)
40	400	0.89	1.38(0.061)	1.13(0.042)	1.06(0.039)	1.03(0.038)
40	1000	0.92	1.48(0.047)	1.21(0.033)	1.15(0.031)	1.13(0.030)
60	40	0.64	0.85(0.171)	0.72(0.120)	0.67(0.116)	0.64(0.121)
60	200	0.85	1.20(0.063)	1.00(0.041)	0.94(0.037)	0.92(0.036)
60	400	0.88	1.25(0.042)	1.05(0.031)	1.00(0.029)	0.98(0.028)
60	1000	0.91	1.31(0.031)	1.11(0.021)	1.06(0.019)	1.04(0.019)
		Covariance (A) with $\theta = 0.9^{-1}$				
40	40	0	0.09(0.133)	0.06(0.099)	0.04(0.082)	0.02(0.063)
40	200	0.56	0.79(0.098)	0.65(0.070)	0.60(0.064)	0.57(0.064)
40	400	0.65	0.95(0.074)	0.78(0.052)	0.72(0.046)	0.70(0.044)
40	1000	0.72	1.09(0.052)	0.90(0.040)	0.84(0.037)	0.82(0.036)
60	40	0	0.08(0.153)	0.05(0.113)	0.04(0.098)	0.03(0.084)
60	200	0.56	0.76(0.075)	0.64(0.053)	0.60(0.051)	0.58(0.052)
60	400	0.65	0.89(0.059)	0.74(0.040)	0.70(0.035)	0.68(0.034)
60	1000	0.72	1.00(0.037)	0.84(0.028)	0.80(0.026)	0.78(0.025)
		Covariance (D)				
40	40	0.58	0.76(0.241)	0.62(0.164)	0.58(0.144)	0.54(0.149)
40	200	0.73	1.07(0.082)	0.87(0.054)	0.81(0.050)	0.78(0.051)
40	400	0.78	1.17(0.063)	0.95(0.041)	0.90(0.038)	0.87(0.037)
40	1000	0.83	1.28(0.042)	1.05(0.030)	0.99(0.028)	0.97(0.027)
60	40	0.61	0.79(0.196)	0.67(0.125)	0.63(0.111)	0.61(0.108)
60	200	0.73	1.00(0.079)	0.84(0.049)	0.79(0.046)	0.77(0.047)
60	400	0.78	1.08(0.052)	0.91(0.034)	0.87(0.032)	0.85(0.031)
60	1000	0.82	1.17(0.034)	0.98(0.025)	0.94(0.023)	0.92(0.023)

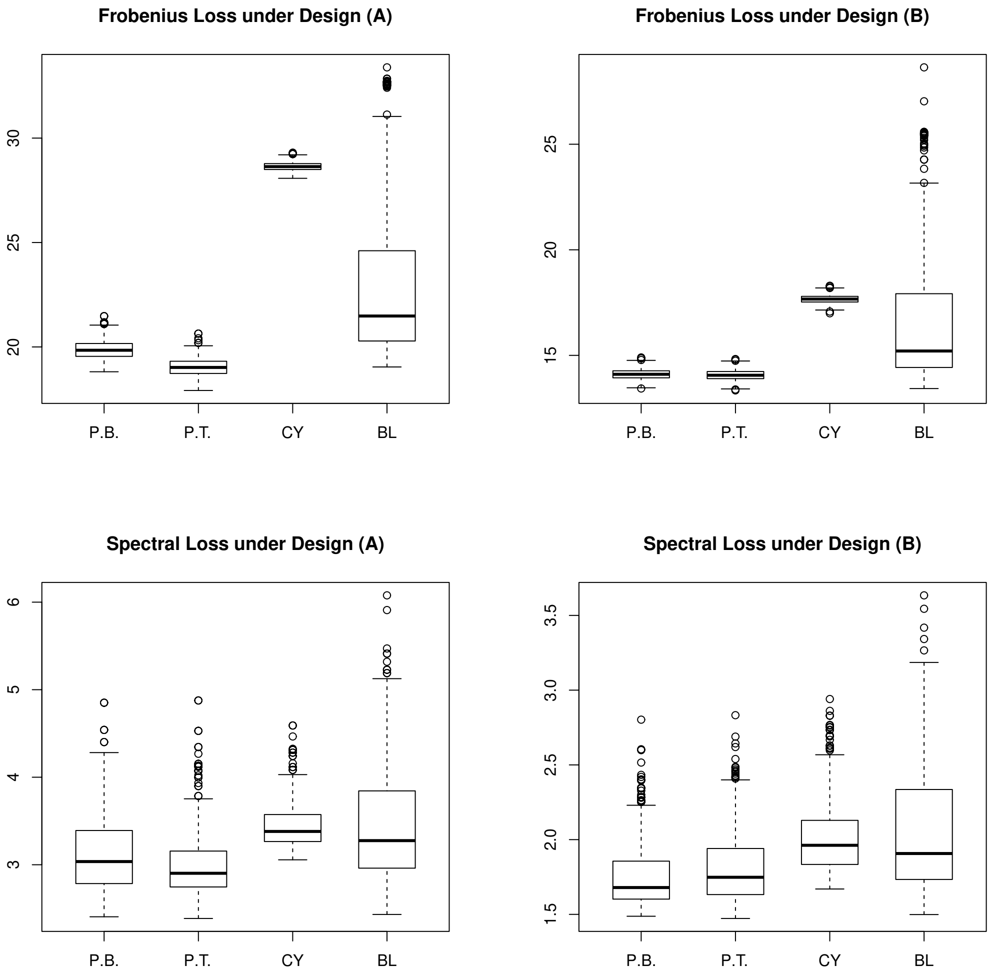

Figure 1: *Box-plots of the Frobenius and Spectral loss of the banding estimator with the proposed band width selector (PB) and Bickel and Levina's selector (BL), the tapering estimator with the proposed band width selector (PT) and Cai and Yuan's adaptive blocking estimator (CY) for covariance Deign (A) with* $\theta = 0.7^{-1}$ *and Design (B) with* $\xi = 0.5$ *and* $\beta = 1.5$*,* $n = 40$*,* $p = 1000$ *and Gaussian data.*

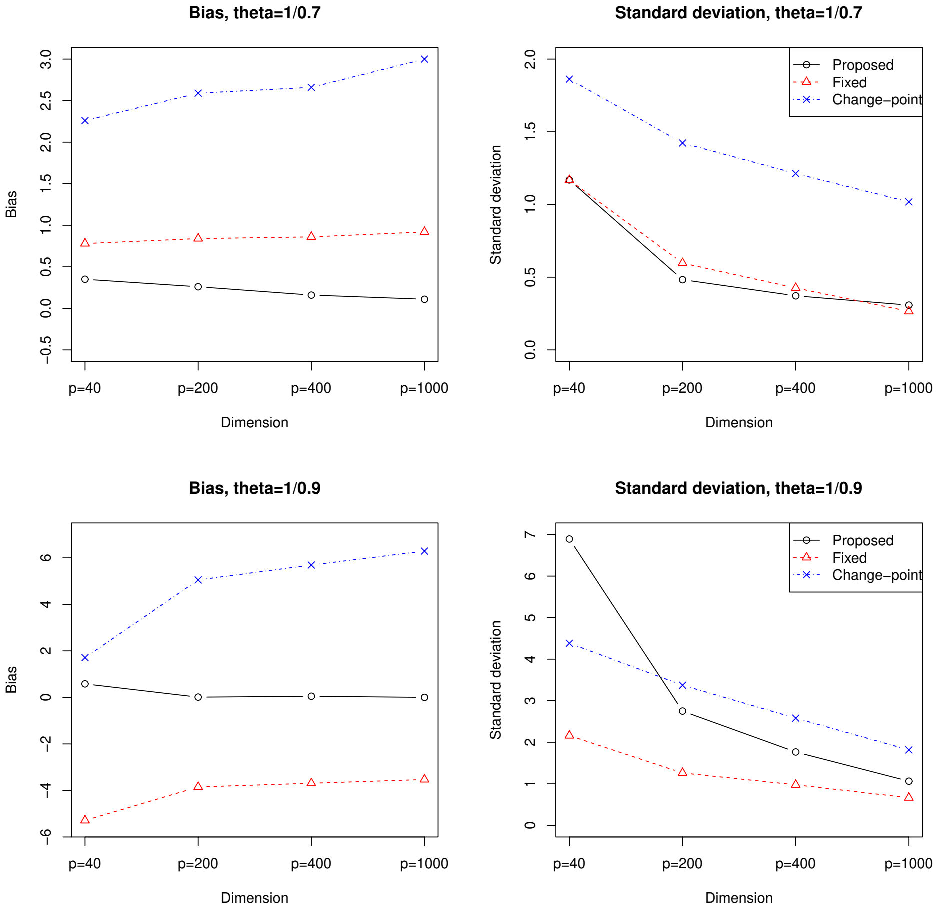

Figure 2: *Empirical bias and standard deviation of the proposed method (4.3), the Fixed and Change-point estimator of Qiu and Chen (2012) for covariance (A) with* $\theta = 0.7^{-1}, 0.9^{-1}$ *and* $n = 60$ *under standard normal innovation.*

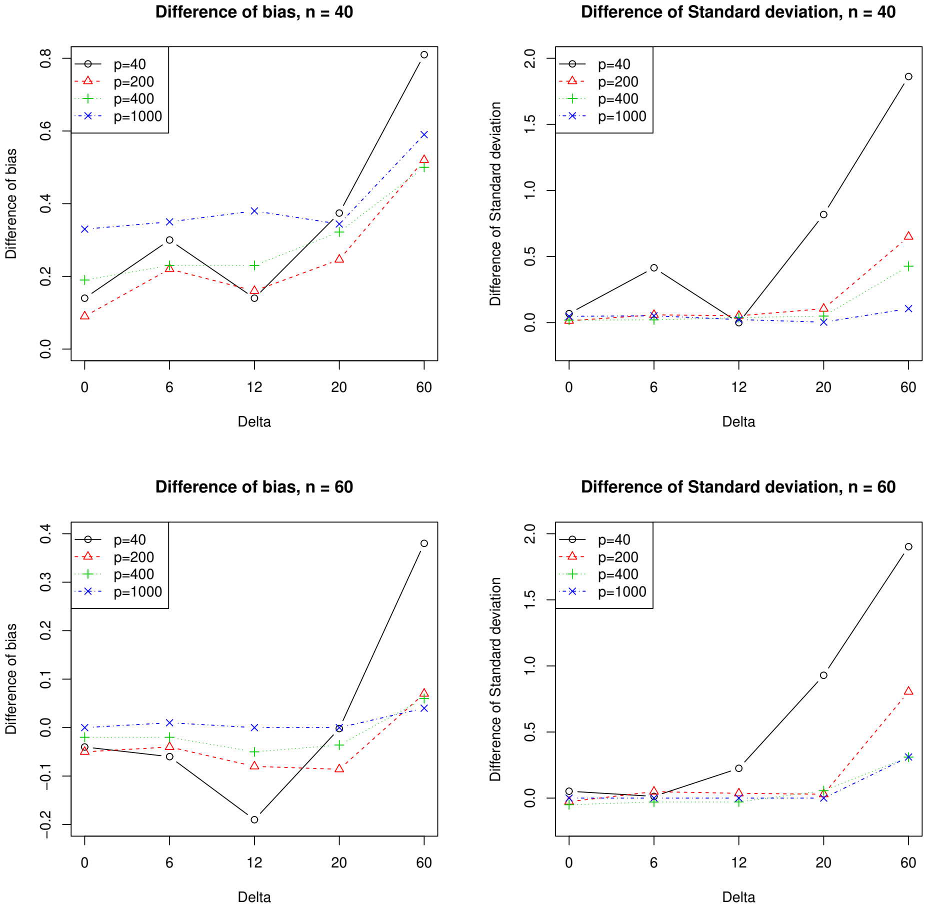

Figure 3: *Differences in the absolute bias and standard deviation of SURE and the proposed band width estimator (SURE minus Proposed) for the tapering estimation under covariance (A) with* $\theta = 0.7^{-1}$ *and* $N(0,1)$ $(\Delta = 0)$*, standardized Gamma innovation with* $\Delta = 6, 12, 20, 60$.

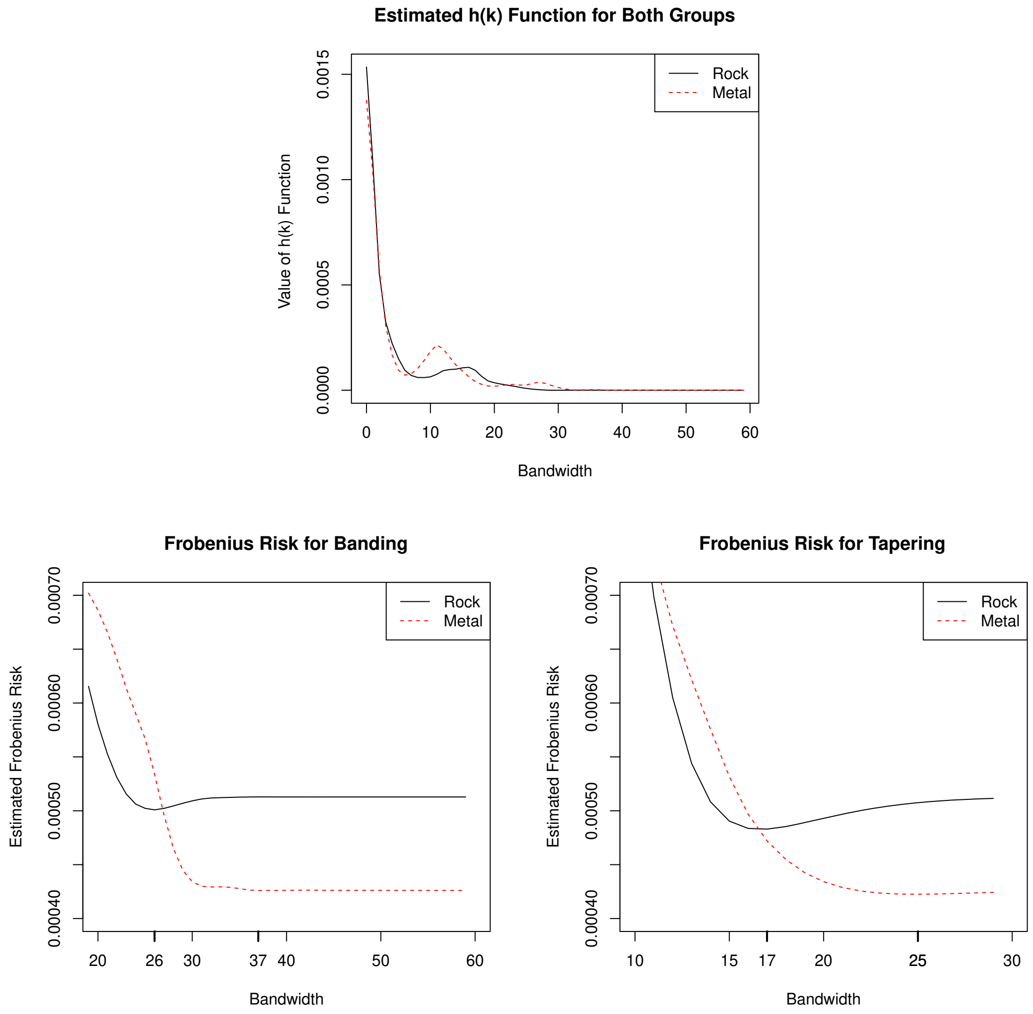

Figure 4: *Estimated $h(k)$ and estimated Frobenius loss of the banding and the tapering estimators for the metal and the rock groups of the sonar spectrum data.*