Statistica Sinica Preprint No: SS-2022-0337	
Title	Bandwidth Selection for Large Covariance and Precision Matrices
Manuscript ID	SS-2022-0337
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0337
Complete List of Authors	Xuehu Zhu, Jian Guo, Xu Guo, Lixing Zhu and Jiasen Zheng
Corresponding Authors	Lixing Zhu
E-mails	lzhu@hkbu.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

Statistica Sinica

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] *School of Statistics, Beijing Normal University, Beijing, China*

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

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

Abstract: For large covariance matrices and the corresponding precision matrices with banding structures, this paper develops a criterion to identify the bandwidth. The new method is based on an objective function that is discontinuous at the true bandwidth to show a "valley-cliff" pattern so that the identification of this location can be visualized and easily implemented. We offer the estimation consistency and the estimation error bound of the estimated covariance matrix and precision matrix with this estimated bandwidth. Numerical studies demonstrate the finite sample validity of the method, and a real data validity analysis is used for illustration.

*Corresponding author (L. Zhu). Email addresses: zhuxuehu@xjtu.edu.cn (X. Zhu), guojian191@mails.ucas.ac.cn (J. Guo), xustat12@bnu.edu.cn (X. Guo), lzhu@hkbu.edu.hk (L. Zhu) and jiasen_zheng@mail.tsinghua.edu.cn (J. Zheng).

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

Key words and phrases: Covariance matrix; Precision matrix; Banding; Tapering; Large p small n; Cholesky decomposition.

1. Introduction

Estimating covariance matrix and its inverse, precision matrix, is one of the fundamental problems in multivariate data analysis. Many classic statistical problems, including principal component analysis (PCA), studies of independence or conditional independence of graphical models, and confidence interval construction for parameters in linear regression, require the knowledge of covariance structure or some aspect thereof. In many cases, precision matrix can infer the conditional dependence structure of random variables. Application areas include gene expression array analysis, functional magnetic resonance imaging, text retrieval, image classification, spectroscopy, climate studies, risk management, and portfolio allocation. The sample covariance matrix is the most commonly used covariance matrix estimator, and its properties are well understood. However, it tends to be inconsistent when the dimension p is large. For more explanation about the limiting spectrum theory of large dimensional sample covariances, see Bai and Yin [1993], Johnstone [2001], Geman [1980], Wachter [1978].

Several proposals are available in the literature on covariance estima-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

tion with high-dimensional data. Among them, some methods handle the studies in which variables with a natural order or the concept of distance between variables (see, e.g., Rothman et al. [2009b]). The implicit regularization assumption is that involved variables are weakly correlated when they are distant from each other. This is equivalent to giving a covariance matrix under a distinct banding or tapering structure. Consistent estimator of covariance matrix is often constructed, for high-dimensional data, through regularization such as shrinkage: Fan et al. [2008], Maurya [2016] and Furrer and Bengtsson [2007]; banding: Bickel and Levina [2004, 2008] and Qiu and Chen [2015]; tapering: Cai et al. [2010], Xue and Zou [2014] and Qiu and Chen [2015]. Some other methods handle the studies with no notion of distance between variables, such as arrays of gene expressions. These studies require estimators that remain constant under variable permutations. Thresholding the sample covariance matrix is a solution such as Bickel and Levina [2009], Karoui [2008] and Qiu and Liyanage [2019]. Random matrix theory presented recently is another shrinkage estimation method (Zhang et al. [2013], Wang and Daniels [2014], Wang et al. [2015]).

For precision matrix, we can also assume that the variables of interest have a natural order that there is no partial correlation between two random variables when the distance between them is large enough. In this case, the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

Cholesky decomposition is often used for regularization analysis to define an estimator, see, e.g. (Pourahmadi [1999], Wu and Pourahmadi [2003], Huang et al. [2006a]). A comprehensive review of high-dimensional covariance and precision matrix estimation under different model structures can be found in Cai et al. [2016].

Suppose we observe p-dimensional independent identically distributed random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with an unknown covariance matrix $\Sigma = Var(\mathbf{X}_1) = (\sigma_{ij})_{p\times p}$ and define $\Omega = \Sigma^{-1}$. Data with natural order generally have an important parameter, the bandwidth K, which defines the number of subdiagonals that are not all zero. Take Σ as an example, i.e. $\sigma_{ij} = 0$ for all $|i - j| > K$. Moreover, banding and tapering estimators for covariance matrix or its inverse relies on good bandwidth estimators when the bandwidth K is unknown. Several methods have been proposed for estimating the bandwidth. Cross-validation (Bickel and Levina [2008]) is a way, but time-consuming. When K is relatively large, this estimation is often unstable, and estimation accuracy is an issue. Qiu and Chen [2012] proposed a non-parametric test for banding covariance matrix without assuming a parametric distribution of the high-dimensional data, and they also presented a consistent estimator of the band size. The tests in Cai and Jiang [2011] and Qiu and Chen [2012] are respectively powerful for sparse and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

dense alternatives. Another class of methods minimiz objective functions to estimate the bandwidth (Cai et al. [2010], Qiu and Chen [2015]). For example, Yi and Zou [2013] and Li and Zou [2016] treated bandwidth as a tuning parameter, gave a criterion by using Stein's unbiased risk estimation (SURE) optimal point, and offered the estimation consistency. However, these estimators are susceptible to sample effects. As pointed out by Chen et al. [2018], even if only one outlier exists in the entire data set, the statistical performance of the estimator may be completely impaired. These methods in practical use may result in underestimation. One reason behind this is that for the sum of squares of the subdiagonals of covariance matrix, the values of estimator tend to be close to each other, except for some maximum dominance, whether or not they are non-zero at the global level. Thus the global minimum (or maximum) value of a criterion at all indices is usually smaller than the true value. The hypothesis testing methods are also helpful as they can provide a practical statistical guide to whether the underlying covariance matrix is of the 'bandable' class (Cai and Jiang [2011], Qiu and Chen [2012], Shao and Zhou [2014]). But the estimation consistency and robustness against 'outlier' samples are still the issues we must handle. Qiu and Chen [2012] considered an estimator based on the difference between continuous statistics to enhance the robustness. Howev-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

er, the objective values vary from infinity to zero at the true bandwidth, which makes it challenging to choose a suitable threshold for estimation.

To address the above issues, we propose a ridge ratio thresholding method and prove the estimation consistency. We understand that almost all existing criteria in the field follow the idea of constructing continuous convex/concave objective function such that the global minimum/maximum can be used as an estimator of the bandwidth K. To achieve convexity/concavity, the objective function usually contains a penalty term. AIC and BIC are the two representatives of such methods. The approaches in this area include Qiu and Chen [2015]. However, as these criteria may be difficult to separate well from nearby values, they often product ,under- or ovproducemation at the sample level. In other words, distinguishing the value at the dedicated bandwidth from others is crucial for estimation. The current paper then proposes a general criterion motivated from Zhu et al. [2020]. To enhance the separation of the value at K from other values, instead of considering a continuous convex-concave objective function, we construct a sequence of ridge ratios as an objective function that is discontinuous at the point K. It drops significantly to zero at K, followed by a sudden rise to 1 for all indices $q > K$. That is, the key feature of our method is that the objective function exhibits a "valley-cliff" pattern at the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

true bandwidth. Therefore, at the sample level, We can quickly determine an estimator of K by using the maximum index of the values smaller than a threshold τ with $0 < \tau < 1$.

This paper is organized as follows. Section 2 contains the criterion construction and the associated asymptotic properties. In Section 3, the method is extended to deal with the bandwidth selection of the precision matrix. Section 4 includes the selection of ridges, simulation results, and analysis of a real data set. The first part of Supplementary Materials discusses how to obtain the estimators of the covariance matrix under banding and tapering structure, the precision matrix, and the corresponding order of the matrices. The second part contains all proofs of the theoretical results.

2. Criterion Construction

2.1 Motivation and Construction

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ for $i = 1, \cdots, n$ be independent and identically distributed (i.i.d.) random variables with the mean vector μ and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$. Define

$$h(k) := \tfrac{1}{p-k} \textstyle\sum_{l=1}^{p-k} \sigma^2_{ll+k}, \text{ for } 0 \le k \le p-1.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

We presume that Σ is banded with the true bandwidth K, i.e., the following assumption:

Assumption 2.1. $\sigma_{ij} = 0$ for all $|i-j| > K$ and $h(K) > 0$.

Under this assumption, $h(0) > 0, h(1) \geq 0, \cdots, h(K-1) \geq 0$ and $h(K) > 0$, but $h(K+1) = \cdots = h(p-1) = 0$. Consider the following sequence: defining $h(p) = h(p+1) = 0$,

$$\frac{h(k+1)}{h(k)}, \text{ for } 0 \leq k \leq p. \tag{2.1}$$

We can see that this sequence has a useful pattern: when $0 \leq k < K$, $h(k+1)/h(k) \geq 0$; when $k = K$, $h(k+1)/h(k) = 0$; and when $K < k \leq p$, $h(k+1)/h(k) = 0/0 = 1$ if we temporarily define $0/0$ as 1. To avoid this undefined ratio, denote the ridge ratio sequence by adding a ridge value $c_n > 0$ to both the numerator and denominator, where c_n tends to zero at a certain rate when n, p tends to infinity. Let $s(k) = \frac{h(k+1)+c_n}{h(k)+c_n}$. It has the following property: as n and $p \to \infty$

$$s(k) = \begin{cases} \frac{h(k+1)+c_n}{h(k)+c_n} \geq 0, & \text{for } 0 \leq k < K, \\ \frac{c_n}{h(k)+c_n} \to 0, & \text{for } k = K, \\ 1, & \text{for } K+1 \leq k \leq p. \end{cases}$$

The sequence presents a good pattern to identify K: regardless of the ratio before the true K, K is the maximum index of the ratios whose values

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

are smaller than one over all k in $0 \leq k \leq p$. This looks like a valley-cliff pattern where at the location K with the value of 0 can be regarded as the valley floor and then faces a cliff with the value of one at the location $K+1$. It remains flat after the position $K+1$. At the sample level, we replace $h(k)$ with the estimators $\hat{h}(k)$ and define $\hat{h}(p) = \hat{h}(p+1) = 0$. Then the corresponding estimator of $s(k)$ is

$$\hat{s}(k) = \frac{\hat{h}(k+1) + c_n}{\hat{h}(k) + c_n}, \quad \text{for } 0 \leq k \leq p, \tag{2.2}$$

where the ridge value c_n tends to zero at a certain rate to be specified later.

To this end, we define an estimator of $h(k)$ as (see, e.g. Qiu and Chen [2015]):

$$\begin{aligned}\hat{h}(k) = \frac{1}{p-k} \sum_{l=1}^{p-k} \Bigg\{ & \frac{1}{A_n^2} \overset{*}{\sum_{i,j}} (X_{il}X_{il+k})(X_{jl}X_{jl+k}) \\ & - \frac{2}{A_n^3} \overset{*}{\sum_{i,j,m}} X_{il}X_{ml+k}(X_{jl}X_{jl+k}) \\ & + \frac{1}{A_n^4} \overset{*}{\sum_{i,j,m,q}} X_{il}X_{jl+k}X_{ml}X_{ql+k} \Bigg\},\end{aligned} \tag{2.3}$$

where $\overset{*}{\sum}$ denotes summation over all involved subscripts and $A_n^b = n!/(n-b)!$ with $0 \leq b \leq n$. Qiu and Chen [2015] has shown that it is an unbiased estimator that is a linear combination of multiple U-statistics, so we can

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

easily derive its consistency.

Once c_n is determined, we have the following result in probability,

$$\lim_{n\to\infty} \hat{s}(k) = \begin{cases} 0, & \text{for } k = K, \\ 1, & \text{for } K+1 \le k \le p-1. \end{cases}$$

Asymptotically, the sequence $\hat{s}(k)$'s has the same pattern as the sequence $s(k)$'s. Note that K is the maximum index of $s(k)$'s smaller than 1. Therefore, the bandwidth K can be estimated as: for any τ with $0 < \tau < 1$

$$\hat{K} = \arg\max_{0\le k\le p} \{k : \hat{s}(k) \le \tau\}. \tag{2.4}$$

This determination is not seriously affected even when the sequence may have multiple local minima.

2.2 Asymptotic Properties

Throughout this paper, $\|\cdot\|_{\psi_2}$ and $\|\cdot\|$ denote the Orilcz norm defined as $\|X\|_{\psi_2} = \sup_{p\ge 1} p^{-1/2} \left(\mathbb{E}|X|^p\right)^{1/p}$ and the l_2 norm of a vector, respectively. To investigate the consistency of this estimator, we state the following two assumptions.

Assumption 2.2. $\log p = o(n^{1/5})$, as $\min\{n, p\} \to \infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

Assumption 2.3. Σ is a positive definite matrix. Let $\mathbf{Z}_k = \Sigma^{-1/2}\mathbf{X}_k$. Variables $X_{il},\ 1 \le l \le p$ and $\mathbf{Z}_k$'s are sub-Gaussian vectors with $\sup_{1\le l\le p} \|X_{kl}\|_{\psi_2} < K_0$ and $E(\exp(\alpha^\top \mathbf{Z}_k)) \le \exp(K_z^2\|\alpha\|^2)$ for some constants $0 < K_0,\ K_z < \infty$.

Remark 2.1. Assumption 2.2 controls the sample size and dimensionality. As $\|\cdot\|_{\psi_2}$ in Assumption 2.3 is a sub-Gaussian norm, the class of sub-Gaussian random variables on a given probability space is the normed space. Classic examples of sub-Gaussian random variables satisfying Assumption 2.3 contain Gaussian, Bernoulli, and all bounded random variables (see, e.g., Vershynin [2010]). In particular, when $\mathbf{Z}_k$ is standard normal, $K_z = 1$, Assumption 2.3 implies that $\max_{1<j<p} \sigma_{jj} < C$ for some $C > 0$. These assumptions are similar to those in Zhao et al. [2018].

The following theorem states the convergence rate of $\hat{h}(q)$ to $h(q)$.

Theorem 2.1. *Under Assumptions 2.2 and 2.3, when $K_z \le 1$, as $\min\{n, p\}$ $\to \infty$, we have*

$$P\left(\max_{0\le k\le p} |\hat{h}(k) - h(k)| > C_0 q_n\right) = o(1),$$

where C_0 is a constant depending on K_0 and $q_n = O\left(\sqrt{\log^5(p \vee n)/n}\right)$.

Remark 2.2. Here the value of C_0 is unknown, and therefore the result is mainly used for the theoretical investigation. In Section 2.3, to estimate

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

the bandwidth of the covariance matrix, we suggest using the ridge value c_n without involving the unknown constant C_0. Moreover, under the same conditions of Lemma A.1 in Qiu and Chen [2015], the conclusion in Theorem 2.1 can be improved to be

$$P\left(\max_{0\le k\le M}|\hat{h}(k)-h(k)|>C_0q_n\right)=o(1),$$

where C_0 is some constant and $q_n=O\left(\sqrt{K\log(p\vee n)/(np)}\right)$. It is worth mentioning that Lemma A.1 in Qiu and Chen [2015] requires that the components of $\mathbf{Z}_k$ are independent with identical first four moments. These conditions are different from Assumption 2.3 in the current paper.

The following theorem states the consistency of the estimator $\hat{K}$ determined by the criterion in (2.4).

Theorem 2.2. *Under Assumptions (2.1),(2.2) and (2.3), if c_n satisfies $c_n\to 0$, $c_n/h(K)\to 0$ and $q_n/c_n=o(1)$ with q_n defined in Theorem 2.1, then we have $P(\hat{K}=K)\to 1$ as $n,p\to\infty$.*

2.3 Tuning Parameter Selection

This subsection presents some discussions and suggestions for selecting the tuning parameters c_n and τ and an estimation algorithm.

For c_n the selection range is quite wide in theory. As we do not have a full data-driven algorithm to select it, it is often the case to recommend a

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

value based on the rule of thumb, like any correlation method with penalties (e.g., the BIC criterion). But if some prior information on the upper bound of the true value K is available, we propose the following semi-data-driven algorithm. From Theorem 2.1, we can see that.

$$\max_{0\le k\le p} |\hat{h}(k) - h(k)| \log(p \cdot n) = O_p(q_n \log(p \cdot n)).$$

Note that if for two large integers $M_1 < M_2$ such that $K < M_1$, M_2 has the same order as p, we then have $\max_{M_1\le k\le M_2} |\hat{h}(k) - h(k)| = \max_{M_1\le k\le M_2} |\hat{h}(k)|$, which has the same rate of convergence as $\max_{0\le k\le p} |\hat{h}(k) - h(k)|$. Therefore, we can define a ridge c_n as

$$c_n = \delta \log(p \cdot n) \max_{M_1\le k\le M_2} |\hat{h}(k)| = O_p(q_n \log(p \cdot n)), \tag{2.5}$$

where $\delta \in (0, \infty)$ is an adjustment parameter. Hence c_n satisfies all assumptions in Theorem 2.2 and is adaptive to the data.

Thus, to use this data-driven ridge, we need prior information on the upper bound of the true bandwidth K. Assume that the true bandwidth K may diverge to infinity at a rate slower than p and n. Then we can use $M_1 = \min\{[n/2], [p/4]\}$ such that $K/M_1 \to 0$. To balance between computational complexity and approximation, we in numerical studies use $M_2 = \min\{[\lambda M_1], p\}$ for a $\lambda > 2$ and to ensure M_1 large enough in finite sample scenarios, we use $M_1 = \max\{\min\{[n/2], [p/4]\}, 20\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

Note that δ is used to adjust the size of c_n. In practice, when p and n are not large, $\max_{M_1 \le k \le M_2} |\hat{h}(k)|$ will not be close to zero, so c_n will be large and the ratio will quickly reach 1, leading to an underestimation. In the numerical studies in this paper, we recommend a value of δ as

$$\delta = \begin{cases} \frac{1}{5}, & \text{if } n \le 50, \quad p \le 50, \\ 1, & \text{otherwise.} \end{cases} \tag{2.6}$$

The next issue is about the selection of the threshold τ. This issue is relatively less important because of the fairly wide range of choices in the interval $(0, 1)$. As a compromise, the median value $\tau = 0.5$ could be recommended to handle the overestimation and underestimation. However, we note that the term $\log(p \cdot n)$, when p and n are large, could result in a relatively large c_n such that the curve of the sequence would become flatter than that with smaller c_n. In this case, choosing $\tau = 0.5$ would more likely cause underestimation. Additionally, an underestimated bandwidth value would cause the covariance matrix estimator to be less accurate. Therefore, for the problem studied in this paper, we recommend a relatively large threshold value $\tau = 0.75$. The details can be found in Supplementary Material showing that this value produces stable results.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

3. Application to the Precision Matrix

When the variables of interest have a natural order, it is usually assumed that partial correlation between two random variables is zero when their distance is large enough. Specifically,

Assumption 3.1. $\omega_{ij} = 0$ for all $|i-j| > K$ and $\frac{1}{p-K}\sum_{l=1}^{p-K}|\omega_{ll+K}| > 0$.

The bandwidth of the precision matrix $\Omega = \Sigma^{-1} = (\omega_{ij})_{p\times p}$ is K. Similarly to the covariance matrix case, let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ for $i = 1, \cdots, n$ be the observations collected from the ith subject. Here, $\mathbf{X}_i$ is independent and normally distributed with mean zero and covariance matrix Σ. The Cholesky decomposition of Σ is

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

where L is a lower triangular matrix whose diagonal elements are all equal to 1 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

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

Let $\varepsilon_i = T\mathbf{X}_i$. An et al. [2014] showed that if the bandwidth of Ω is K, then

$$X_{ij} = \begin{cases} \varepsilon_{ij}, & \text{for } j = 1, \\ \sum\limits_{q=(j-K)_1}^{j-1}(-t_{jq}X_{iq} + \varepsilon_{ij}), & \text{for } j > 1, \end{cases} \tag{3.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

where $(j-K)_1 = \max\{1, j-K\}$, the elements of ε_i are independent and normally distributed with mean zero, and the covariance matrix of ε_i is D. When the precision matrix Ω has a band structure, Rothman et al. [2009a] showed that the Cholesky factor T has the same bandwidth K as Ω. We can then turn to estimate the bandwidth of T.

Let M be an upper bound of K, $t_j^{(k)} = (t_{j,(j-k)_1}, \ldots, t_{j,j-1})^\top$, $\chi = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$, and χ_j be the jth column of χ, $\chi_j^{(k)} = (\chi_{(j-k)_1}, \ldots, \chi_{j-1})$. By fitting the regression equation (3.7), we can define an estimator $\hat{t}_j^{(M)}$ of $t_j^{(M)}$ as:

$$\hat{t}_j^{(M)} = -({\chi_j^{(M)}}^\top \chi_j^{(M)})^{-1} {\chi_j^{(M)}}^\top \chi_j. \tag{3.8}$$

Let $l(k) = \frac{1}{p-k}\sum_{l=1}^{p-k} |t_{l+k,l}|$. Then an estimator of $l(k)$ is defined as

$$\hat{l}(k) = \frac{1}{p-k}\sum_{j=k+1}^{p} |\hat{t}_{j,j-k}^{(M)}|, \ k = 0, \ldots, M, \tag{3.9}$$

where $\hat{t}_{j,j-k}^{(M)}$ stands for the $(j-k-(j-M)_1+1)$th element of $\hat{t}_j^{(M)}$.

Remark 3.1. Without the band structure of Ω, the regression equations are $X_{i1} = \varepsilon_{i1}$ and $X_{ij} = \sum_{q=1}^{j-1}(-t_{jq}X_{iq} + \varepsilon_{ij})$ for $j > 1$. In the case of large p and small n, the estimates of T obtained by fitting these regression equations may not work well, some regularization of T is often imposed (Levina et al. [2008], Huang et al. [2006b]). However, if Assumption 3.1 holds and $M < n$, a good estimator of T can be constructed in the large p

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

and small n setting.

The following two theorems state the estimation consistency of relevant statistics.

Theorem 3.1. *Suppose that* $\mathbf{X}_i$*, for* $i = 1, \ldots, n$*, are independent identically normally distributed. Under Assumption 3.1, if* $K \leq M < n$*, then*

$$P\left(\max_{0\leq k\leq M} |\hat{l}(k) - l(k)| > C_1\gamma_n\right) = o_p(1),$$

as $\min\{n, p\} \to \infty$*, where* C_1 *is a constant and* $\gamma_n = O(\sqrt{\log p/n})$.

Based on Theorem 3.1, we can similarly define an objective function as that in (2.2):

$$\hat{r}(k) = \frac{\hat{l}(k+1) + \tilde{c}_n}{\hat{l}(k) + \tilde{c}_n}, \text{ for } 0 \leq k \leq M-1, \tag{3.10}$$

where the choice of $\tilde{c}_n$ is discussed in the following theorem. Thus, the bandwidth K of the precision matrix can be estimated as: for $0 < \tau < 1$,

$$\hat{K} = \underset{0\leq k\leq M-1}{\arg\max} \{k : \hat{r}(k) \leq \tau\}. \tag{3.11}$$

Like that in Subsection 2.3, we also recommend the thresholding value $\tau = 0.75$, and the bandwidth upper bound $M_1 = \max\{\min\{[p/4], [n/2]\}, 20\}$. The ridge $\tilde{c}_n$ is similarly defined as:

$$\tilde{c}_n = \delta \log(n) \max_{M_1\leq k\leq M_2} |\hat{l}(k)|, \tag{3.12}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

where δ is the same value defined in (2.6) and $M_2 = \min\{[\lambda M_1], M-1\}$ with $\lambda \in (2,3)$. The following theorem states the estimation consistency.

Theorem 3.2. *Under the normality assumption of* $\mathbf{X}_i$ *and Assumption (3.1), when* $\tilde{c}_n \to 0$, $\tilde{c}_n / l(K) \to 0$ *and* $\tilde{c}_n \sqrt{n/\log p} \to \infty$, *then* $P(\hat{K} = K) \to 1$, *as* $n, p \to \infty$.

We have obtained bandwidth estimators of the covariance matrix and precision matrix with band structure using the proposed ridge ratio thresholding method. We also discuss how to apply the estimated bandwidth to estimating the covariance matrix and precision matrix and give the properties of the corresponding estimators in Supplementary Materials.

4. Numerical Studies

In this section, we will utilize several numerical studies first to select the appropriate value of λ and then assess the finite sample performance of the proposed method and compare it with some state-of-the-art approaches.

4.1 Selection of λ

Consider two covariance structures similarly to the examples in Qiu and Chen [2015]. The data generation process used in this experiment is as follows:

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

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

where Z_{ij} are generated i.i.d. from $N(0,1)$ and $\Sigma = (\sigma_{ij})_{1\le i,j\le p}$ is the covariance matrix. Consider the two designs as:

1. $\sigma_{ij} = 3^{-|i-j|/2} I(|i-j| \le K)$, for $K = 4$,
2. $\sigma_{ij} = I(i=j) + 0.2I(0 < |i-j| \le K)$, for $K = 8$.

In this example, we consider the true bandwidth to be $K = 4, 8$ in two scenarios: $n = 50,\ p = 300$; and $n = 200,\ p = 100$. We search for a value of λ by maximizing the correct rate of the determined bandwidth in the interval $[1, 3]$ with the grid points $1 : 0.2 : 3$. For each λ, we performed 50 replications to obtain the mean and correct rate. Figures 1 and 2 plot the mean values and the correct rates of the determined bandwidth for different λ.

Obviously, from these two figures, the proposed method is not very sensitive to the choice of λ when it is within the interval $[2, 3]$, and its correct rate well keeps equal to 100%. The numerical studies not reported in this paper indicate that when $\lambda > 3$, the correct ratio also keeps equal to 100%. Therefore it is sensible to recommend the median value of 2.5 as a suitable value of λ.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

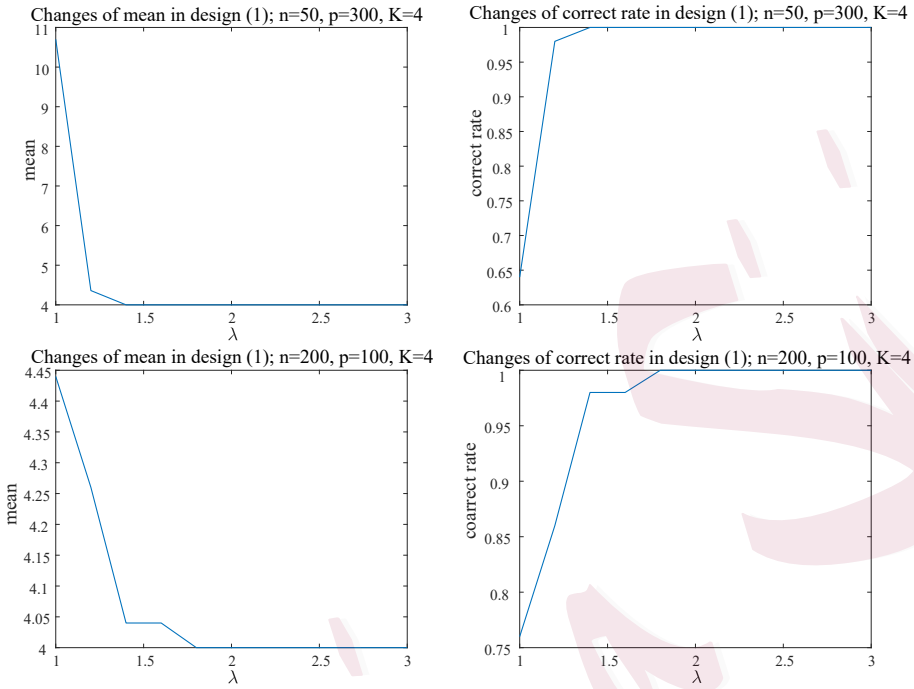

Figure 1: The results of λ and the estimated bandwidth mean under the covariance structure (1).

4.2 Simulation Study

In this subsection, we consider two sets of numerical experiments below. The first set, including **Examples 1–3**, is used to compare our method with Qiu and Chen's estimator in Qiu and Chen [2015] and Bickel and Levina's estimator in Bickel and Levina [2008].

Write our method and their methods as VCC, QC, and BL, respectively, and use "True" to indicate the true bandwidth K. To make a fair comparison between QC and VCC, we adopt the same upper bound

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

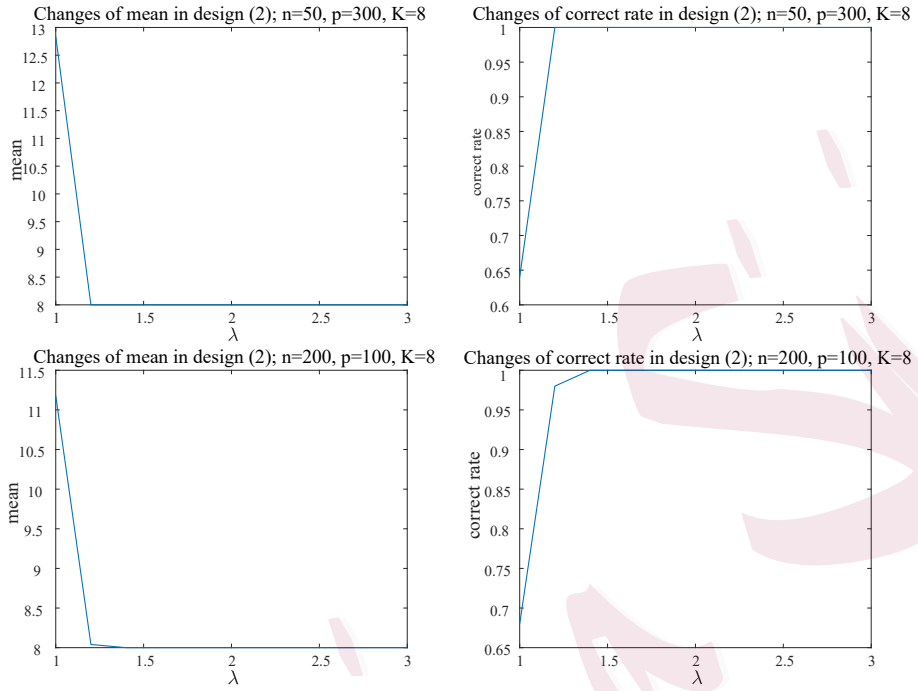

Figure 2: The results of λ and the estimated bandwidth mean under the covariance structure (2).

$M_1 = \max\{\min\{[n/2], [p/4]\}, 20\}$ of K. We search for the minimum value point for BL and QC method in $0, \ldots, M_1$. The second set, including Example 4, forces on precision matrix and compares VCC with the hypothesis testing procedure (Backward-Forward procedure) in An et al. [2014]. Each experiment is repeated 100 times for QC and VCC throughout this subsection. Compared with the Backward-Forward procedure, the replication time is 1000, so the Type I error can be well controlled.

The data are generated from

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

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

where Z_{ij} are i.i.d. respectively from $N(0,1)$ and t_5 that denotes the standardized t-distribution with degrees 5 of freedom.

Example 1. Consider the following example similarly to that in Qiu and Chen [2015] but with truncated covariance matrix $\Sigma = (\sigma_{ij})_{1\le i,j\le p}$ as:

A. $\sigma_{ij} = \theta^{-|i-j|} I(|i-j| \le K)$, with $K = 5$ and $\theta = 0.7^{-1}$;

B. $\sigma_{ij} = I(i=j) + \xi |i-j|^{-\beta} I(0 < |i-j| \le K)$, with $K = 2$, $\xi = 0.5$ and $\beta = 1.5$.

We design the same samples sizes and dimensions as those in Qiu and Chen [2015], which are $n = 40,\ 60$ and $p = 40,\ 200,\ 400,\ 1000$, respectively.

Tables 1 and 2 report the mean and standard deviations of the estimated bandwidth by these three methods. The results show that VCC has the best performance with less deviation among the three contenders, and QC has a better performance than BL.

Table 1: Mean and standard deviation of the estimated bandwidth by VCC, QC, and BL under the covariance structure A in **Example 1**.

		Covariance (A) with $\theta^{-1} = 0.7$							
		Normal				t-distribution			
n	p	True	VCC	QC	BL	True	VCC	QC	BL
40	40	5	4.72(1.301)	6.34(1.387)	5.56(1.833)	5	4.80(0.876)	6.32(1.377)	5.56(2.392)
40	200	5	4.75(0.956)	6.56(1.343)	8.70(4.446)	5	5.12(0.782)	6.76(1.386)	8.76(5.142)
40	400	5	4.86(0.492)	6.39(1.392)	9.94(5.199)	5	4.80(0.568)	6.46(1.396)	9.90(5.390)
40	1000	5	5(0)	6.21(1.402)	10.64(5.524)	5	5(0)	6.36(1.375)	10.74(5.677)
60	40	5	4.97(1.086)	6.35(1.359)	5.34(1.683)	5	4.94(0.565)	6.39(1.355)	5.94(2.182)
60	200	5	4.77(0.583)	6.28(1.386)	10.85(6.428)	5	4.80(0.538)	6.24(1.319)	11.89(8.128)
60	400	5	5(0)	6.68(1.385)	11.64(6.844)	5	5(0)	6.44(1.366)	16.54(7.830)
60	1000	5	5(0)	6.62(1.316)	14.14(8.600)	5	5(0)	6.38(1.376)	16.96(8.856)

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

Table 2: Mean and standard deviation of the estimated bandwidth by VCC, QC, and BL under the covariance structure B in **Example 1.**

		Covariance (B) with $\xi = 0.5,\ \beta = 1.5$							
		Normal				t-distribution			
n	p	True	VCC	QC	BL	True	VCC	QC	BL
40	40	2	2.35(1.225)	3.51(1.322)	3.46(2.346)	2	2.35(1.086)	3.38(1.316)	4.10(2.576)
40	200	2	1.91(0.795)	3.34(1.307)	6.67(4.803)	2	1.83(0.377)	3.39(1.286)	9.48(5.926)
40	400	2	2(0)	3.79(1.258)	8.20(4.872)	2	1.98(0.140)	3.47(1.283)	10.79(6.256)
40	1000	2	2(0)	3.29(1.274)	9.20(5.737)	2	2(0)	3.61(1.263)	10.42(5.919)
60	40	2	2.29(0.795)	3.37(1.308)	3.36(1.967)	2	4.63(0.847)	3.20(1.310)	2.49(2.977)
60	200	2	2(0)	3.40(1.223)	7.45(6.195)	2	1.97(0.171)	3.46(1.329)	10.86(7.702)
60	400	2	2(0)	3.22(1.307)	9.41(7.354)	2	2(0)	3.52(1.306)	14.93(9.305)
60	1000	2	2(0)	3.23(1.302)	11.03(9.220)	2	2(0)	3.30(1.291)	15.13(10.051)

As the dimension p increases, the deviation and standard deviation decrease, while QC and BL do not. When $p = 1000$, VCC's deviation and standard deviation are equal to 0. This means that VCC always makes the correct decision in this simulation.

Example 2. This model with normal data is similar to the example in Bickel and Levina [2008]: $\Sigma = (\sigma_{ij})_{1\le i,j\le p}$ with

$$\sigma_{ij} = \frac{1}{2}\left[||i-j|+1|^{2H} - 2|i-j|^{2H} + ||i-j|-1|^{2H}\right] I(|i-j| \le K).$$

Table 3: Mean and standard deviation of the estimated bandwidth by VCC and BL for **Example 2.**

			Mean(Std)		
p	H	True(K)	VCC	BL	QC
10	0.5	0	0.0(0.0)	0.0(0.0)	0.0(0.0)
10	0.7	5	5(0.0)	5.0(2.8)	2.6(0.7)
100	0.5	0	0.0(0.0)	0.0(0.0)	0.0(0.0)
100	0.7	4	4(0.0)	4.9(3.2)	17.0(9.8)
200	0.5	0	0.0(0.0)	0.0(0.0)	0.0(0.0)
200	0.7	3	3(0.0)	4.5(2.7)	24.6(16.1)

The sample size and the dimension are $n = 100$ and $p = 10, 100, 200$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

respectively. The results are summarized in Table 3. The results clearly show the superiority of VCC to BL and QC.

Example 3. To further check the performance of VCC under banding structures, consider the following covariance structure:

$$\sigma_{ij} = I(i=j) + \sum_{l=1}^{K} \xi l^{-\beta/2} I(|i-j|=l), \text{ with } \xi = 0.5 \text{ and } \beta = 0.9$$

with larger bandwidths $K = 4, 13, 19$. The sample sizes and dimension are $n = 50,\ 100$ and $p = 50,\ 500,\ 1000$, respectively. Table 4 reports the averages, standard deviations, and frequencies of the bandwidth estimators by QC and VCC. Some findings from Table 4 are as follows.

First, when $K = 4$, VCC has stable results and a high frequency of correct decisions, while QC tends to mate the bandwidth grossly. Moreover, except for the cases of $p = 50$ and $n = 50$, in more detail, QC has a lower proportion of correct decisions, less than 35%. Except for the cases of $n = 50$ and $p = 50$, VCC can have more than 75% of correct decisions, and when $K = 4$, the proportion of correct decisions of VCC is 100%. The performance of VCC is significantly better than QC. Secondly, as the value of K increases, the standard deviation of VCC increases reasonably, and the proportion of correct decisions decreases.

Let $M_n(k) = \frac{1}{p}\sum_{|l_1-l_2|>k} \sigma_{l_1 l_2}^2 + \frac{1}{np}\sum_{|l_1-l_2|\le k} \sigma_{l_1 l_1}\sigma_{l_2 l_2}$ and $\hat{M}_n(k)$ denote the estimator of $M_n(k)$ defined in Qiu and Chen [2015]. Figures 3 and 4

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

plot the curves of the objective functions of QC and VCC at the population level and their box plots at the sample level, respectively. The box plots in Figures 3 and 4 show the advantage of discontinuity of the objective function we defined and the results of QC. We can observe that for $k > K$, almost all values of $\hat{s}(k)$ are above the threshold 0.75 and $\hat{s}(K)$ is much smaller than 0.75. Further, when $p = 100$, $p = 1000$ and $K = 19$, $\hat{h}(1)$ attains the global minimum. But the discontinuity at the true bandwidth K greatly separates $\hat{s}(K)$ from all consecutive ratios. This explains why VCC performs better than QC and BL.

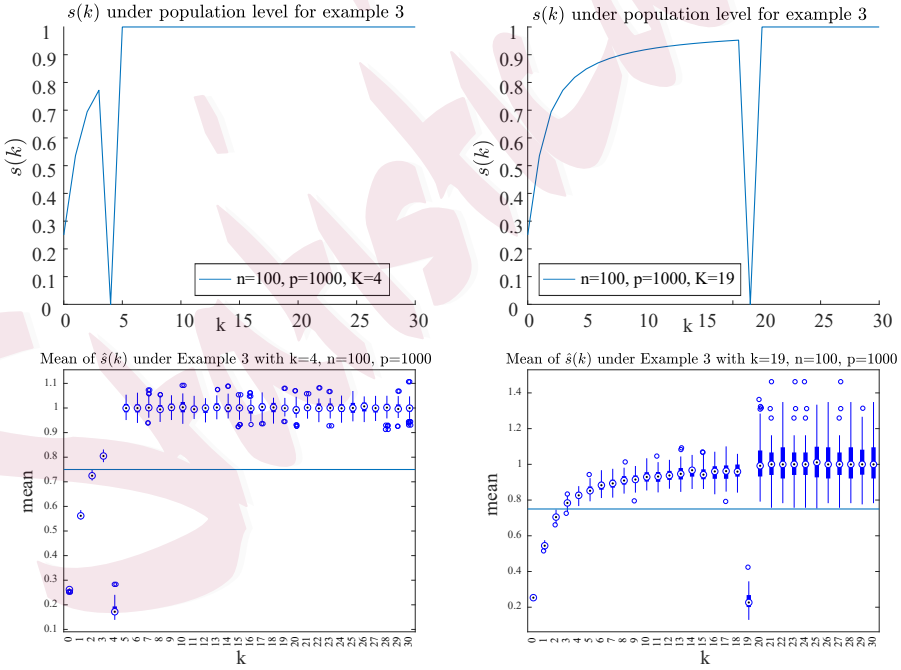

Figure 3: The true curves of $s(k)$ and boxplots of $\hat{s}(k)$ for **Example 3** with $K = 4, 19$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

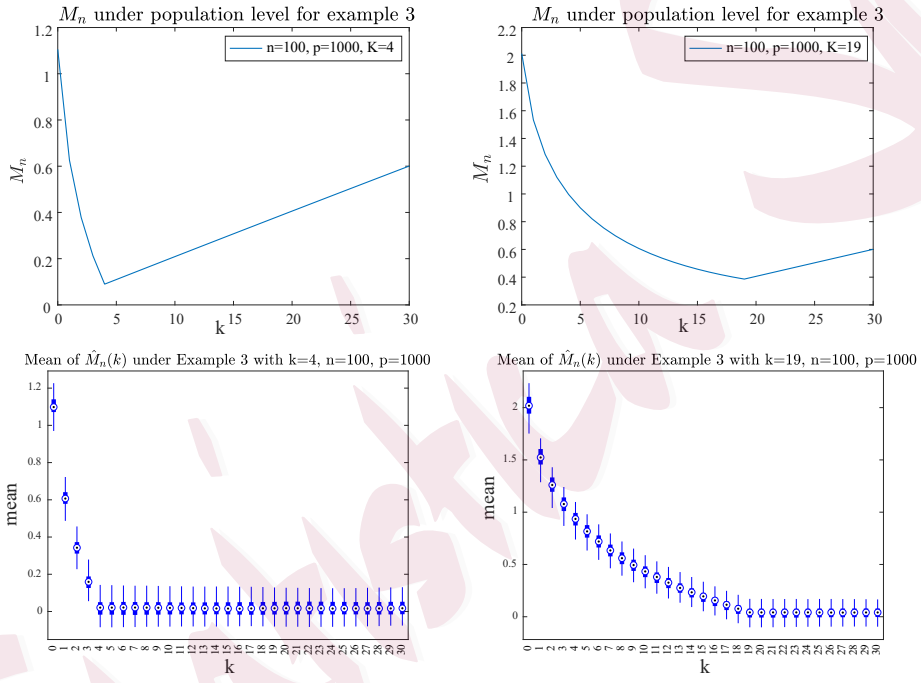

Figure 4: The true curves of M_n and boxplots of $\hat{M}_n$ for **Example 3** with $K = 4, 19$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

In summary, VCC works better than QC and BL; in some cases, the advantage is very significant.

Table 4: Mean, standard deviation, and frequencies of the estimated bandwidth by VCC and QC for **Example 3.**

			Example 3							
			Mean(Std)		frequencies under VCC			frequencies under QC		
n	p	true(K)	VCC	QC	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$	$\hat{K} < K$	$\hat{K} = K$	$\hat{K} > K$
50	50	4	4.35(1.120)	11.36(5.943)	2	92	6	0	24	76
50	500	4	4(0)	13.73(7.802)	0	100	0	0	22	78
50	1000	4	4(0)	14.13(8.553)	0	100	0	0	24	76
100	50	4	4(0)	12.37(6.350)	0	100	0	0	22	78
100	500	4	4(0)	24.08(17.176)	0	100	0	0	14	86
100	1000	4	4(0)	27.33(17.044)	0	100	0	0	16	84
50	50	13	11.43(6.627)	15.91(2.878)	28	61	11	0	36	64
50	500	13	13.95(2.396)	18.98(4.948)	7	85	8	0	26	74
50	1000	13	13.31(1.361)	18.74(4.898)	0	94	6	0	26	74
100	50	13	13.11(4.364)	16.26(2.953)	0	100	0	0	31	69
100	500	13	13.34(1.430)	31.53(14.239)	0	94	6	0	14	86
100	1000	13	13(0)	29.06(14.083)	0	100	0	0	19	81
50	50	19	12.89(1.100)	19.41(0.9331)	58	34	8	0	39	61
50	500	19	19.28(0.792)	22.36(2.5605)	0	86	14	0	26	74
50	1000	19	19.21(0.795)	21.65(2.532)	0	93	7	0	35	65
100	50	19	11.72(3.662)	19.41(0.494)	17	77	6	0	59	41
100	500	19	19.44(1.929)	33.78(12.387)	0	94	6	0	22	78
100	1000	19	19(0)	33.79(12.194)	0	100	0	0	21	79

Now we examine the finite sample performance of VCC for precision matrix and compare it with the hypothesis testing procedures (Backward and Forward procedure) developed in An et al. [2014]. We write them as BackE and ForE. Because of an inverse matrix involved in their computing process, An et al. [2014] considered the upper bound $M = 2K$ in their estimating algorithm. Again, we adopt the model used in An et al. [2014] for a fair comparison.

Example 4: Consider the following precision matrix $\Omega = (\omega_{ij})$ with

$$\Omega_{ij} = I(i = j) + \sum_{l=1}^{K} 3^{-l/2} I(|i - j| = l),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

where $K = 2, 4, 6, 8$. The results are reported in Table 5.

The results of the three methods in Table 5 clearly show that when $K \leq 6$, BackE performs well, and VCC works similarly to BackE. ForE is not as good as VCC and BackE. When $K = 8$, the performance of BackE is much worse than VCC.

Table 5: Percentages (%) of correct identifications of K by our proposed method(VCC) and Backward and Forward estimators (AGL) for the normally distributed data in **Example 4**.

	p														
	30			100			200			500			1000		
n	VCC	Backward	Forward	VCC	Backward	Forward	VCC	Backward	Forward	VCC	Backward	Forward	VCC	Backward	Forward
							$K = 2$								
50	100	97.5	99.8	97.3	98.9	99.5	99.2	98.6	99.7	99	99	99.8	100	98.8	99.7
200	99.1	97.4	99.9	100	98.6	99.8	100	99.2	99.7	100	98.8	99.8	100	99.1	100
400	100	97.3	99.6	100	98.4	99.8	100	99	99.8	100	98.8	100	100	99.3	100
							$K = 4$								
50	92.4	73.6	7.5	100	99	52	99.5	99.2	83.8	99.7	99.5	99.8	99.8	99.7	99.9
200	98.3	98.8	98	99	99.2	99.7	100	99.5	99.8	100	98.9	99.8	100	99.7	99.7
400	100	98.6	99.6	100	99.4	99.9	100	99.1	99.8	100	99.7	99.7	100	99.4	99.9
							$K = 6$								
50	17.7	1.5	0	39.3	2.4	0	75.1	4.9	0	100	16.2	0.4	100	47.9	0.2
200	59.3	18.7	0.4	94.1	72.7	5.7	100	96.4	14.5	99.2	99.8	47.6	100	100	84.7
400	83.5	67.1	7.6	100	99.8	38.7	99	99.6	70.8	100	99.5	99.3	100	99.8	99.6
							$K = 8$								
50	11.1	0	0	3	0	0	6	0	0	31.1	0	0	59.4	0	0
200	5	0	0	11.3	0	0	35	0	0	76.5	0	0	98	0	0
400	18.3	0	0	23.2	0	0	64.3	0	0	96	10	20	100	4	2

4.3 Two Real Data Examples

In this subsection, we illustrate the application of VCC to the sonar data and the ionospheric data. Both datasets are available in the UCI database.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

4.3.1 Sonar Dataset

This data set was analysed in Yi and Zou [2013] and Qiu and Chen [2015]. There are 218 observations, 60 input variables, and one output variable. The output target is mine or rock, of which 97 are from rock and 111 from mine. They were considered two data sets, and two corresponding matrices were estimated. Yi and Zou [2013] found that the values on the diagonal of the sample covariance matrix decayed significantly along the direction away from the diagonal. This finding shows the banding structure that combines the sample covariance matrix with the estimated bandwidth yields better results. Figure 5 plots the curves of the function $\hat{h}(k)$ on the covariance matrix defined in (2.3) and the function $\hat{l}(k)$ on the precision matrix defined in (3.9). It can be found that the covariance matrix has a clear hierarchical nature and the accuracy matrix has a large variation in the subdiagonal. Thus, assuming that the covariance matrix has a potentially bundleable structure is reasonable.

Different methods yielded different estimated bandwidths for the covariance matrix. Qiu and Chen [2015] and Bickel and Levina [2008] derived bandwidth estimators of 26 and 37 (QC) and 35 and 44 (BL) for the rock and mine classes, respectively. The proposed VCC gives values of 3 and 27 for the rock and metal groups, respectively. The estimated $\hat{s}(k)$ are shown

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

in Figure 6.

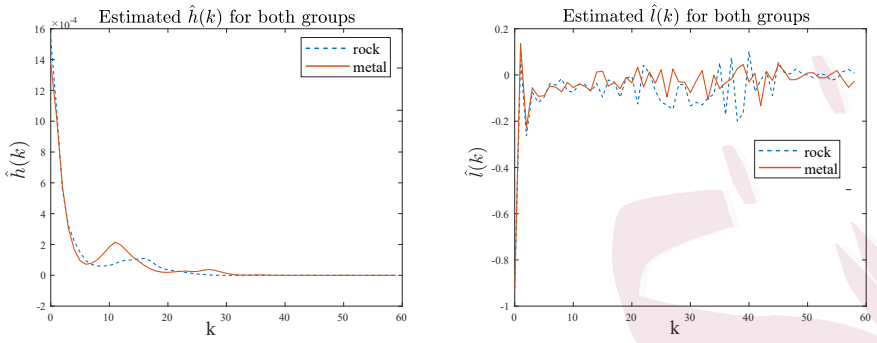

Figure 5: The value of the estimated $\hat{h}(k)$ and $\hat{l}(k)$ under two types of Sonar data

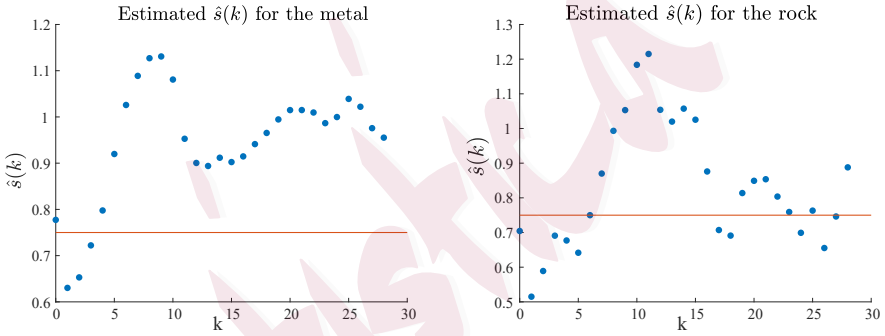

Figure 6: The value of the estimated $\hat{s}(k)$ under two types of Sonar data

To examine the estimation efficiency of these three methods, we used linear discriminant analysis for data classification. Here, the sample covariance matrix used in the linear discriminant analysis is replaced with a banding sample covariance matrix that combines the estimated bandwidths obtained by the above methods. The output correct rates were 0.6394 (VCC), 0.5769 (QC), and 0.5337 (BL), respectively. The performance of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

the three classifiers demonstrates the superiority of VCC concerning QC and BL.

4.3.2 Ionospheric dataset

Ionospheric data are mainly used to predict atmospheric structure based on radar echoes of free electrons in a given ionosphere. This is a binary classification problem. The data set consists of 351 observations, 34 input variables, and ,one output variable, including two types of labels, "g" and "b" for "good" and "bad," respectively. Similarly, Figure 8 plots the line graphs of the function $\hat{h}(k)$ defined in (2.3) and the function $\hat{l}(k)$ defined in (3.9). It is clear that as k increases, $\hat{h}(k)$ gradually approaches zero, but $\hat{l}(k)$ does not. Therefore, it is reasonable to consider the frequency banding assumption on the covariance matrix. Then we estimate the bandwidth to obtain an effective classifier. The estimator based on VCC is 26. The classifier for the sonar data is obtained using linear discriminant analysis. The corresponding accuracy is 0.8666. When applying QC and BL, the estimated bandwidths are 29 and 20, and the accuracies of the corresponding classifier is 0.8547 and 0.8575, respectively. The estimated $\hat{s}(k)$ is shown in Figure 8

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

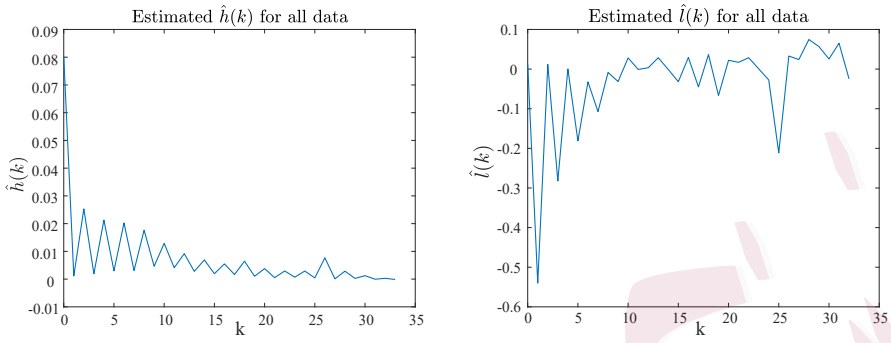

Figure 7: The value of the estimated $\hat{h}(k)$ and $\hat{l}(k)$ for all ionosphere data

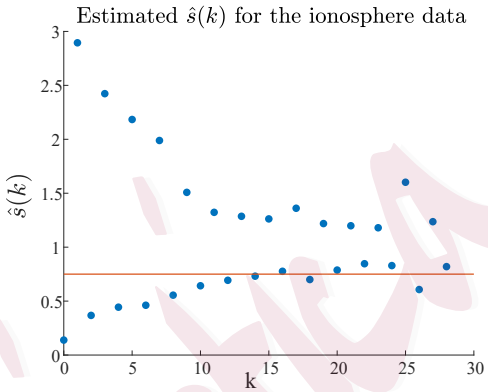

Figure 8: The value of the estimated $\hat{s}(k)$ under ionosphere data

5. Conclusion

This paper proposes a novel approach called "valley-cliff criterion" (VC-C) to determine the band sizes of the large-dimensional covariance matrix. It can also apply to the bandwidth selection problem of the precision matrix. The new approach is computationally efficient, and the resulting estimation is consistent. Unlike the traditional methods that construct a convex/concave objective function to search for a minimizer/maximizer as

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

an estimator, the key feature of the new criterion is its discontinuity of the objective function at the true bandwidth such that the corresponding value of the objective function can be significantly stood out for identification. Our method can be nested in a class of regularized estimators of covariance and precision matrices. This methodology should have the potential to be applied to other order determination problems with large-dimensional covariance matrices. The research is ongoing.

6. Supplementary Material

In the online supplementary material, we discusses how bandwidth estimation can be applied to the estimation of covariance matrices and precision matrices. This supplementary material also contains the part of numerical studies and all proofs of the theoretical results.

7. Acknowledgement

The authors are grateful to Drs. Songxi Chen, Yumou Qiu, and Baiguo An for providing partial codes. Xuehu Zhu's research was supported by the National Key R&D Program of China (2022YFA1003803), the National Social Science Foundation of China (21BTJ048), and the Zhongying Young Scholar Program. Xu Guo's research was supported by a grant from the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

Natural Science Foundation of China (NSFC12071038). The research of Lixing Zhu was supported by a grant from the University Grants Council of Hong Kong and a grant from the Natural Science Foundation of China (NSFC12131006). The thanks go to the Editor, Associate Editor, and two referees for their constructive suggestions that significantly improved an early manuscript.

References

B. G. An, J. H. Guo, and Y. F. Liu. Hypothesis testing for band size detection of high-dimensional banded precision matrices. *Biometrika*, 101(2):477–483, 2014.

Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

P. J. Bickel and E. Levina. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

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

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2009.

T. T. Cai and T. F. Jiang. Limiting laws of coherence of random matrices with applications to

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, 39(3):1496–1525, 2011.

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.

T. T. Cai, Z. Ren, and H. H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *The Annals of Statistics*, 10 (1):1–59, 2016.

M. J. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under hubers contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

J. Q. Fan, Y. Y. Fan, and J. C. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

S. Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2): 252–261, 1980.

J. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006a.

J. H. Huang, N. P. Liu, M. Pourahmadi, and L. X. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006b.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

I. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

N. E. Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, 36(6):2717–2756, 2008.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, 2(1):245–263, 2008.

D. N. Li and H. Zou. Sure information criteria for large covariance matrix estimation and their asymptotic properties. *IEEE Transactions on Information Theory*, 62(4):2153–2169, 2016.

A. Maurya. A well-conditioned and sparse estimation of covariance and inverse covariance matrices using a joint penalty. *Journal of Machine Learning Research*, 17(1):4457–4484, 2016.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

Y. M. Qiu and S. X. Chen. Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *The Annals of Statistics*, 40(3):1285–1314, 2012.

Y. M. Qiu and S. X. Chen. Bandwidth selection for high-dimensional covariance matrix estimation. *Journal of the American Statistical Association*, 110(511):1160–1174, 2015.

Y. M. Qiu and J.S.S Liyanage. Threshold selection for covariance estimation. *Biometrics*, 75 (3):895–905, 2019.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

A. J. Rothman, E. Levina, and J. Zhu. A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97(3):539–550, 2009a.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009b.

Q. M. Shao and W. X. Zhou. Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *The Annals of Probability*, 42(2): 623–648, 2014.

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

K. W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, 6(1):1–18, 1978.

C. Wang, G.M. Pan, T.J. Tong, and L.X. Zhu. Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statistica Sinica*, 25(3):993–1008, 2015.

Y. Wang and M.J. Daniels. Computationally efficient banding of large covariance matrices for ordered data and connections to banding the inverse cholesky factor. *Journal of Multivariate Analysis*, 130:21–26, 2014.

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

L. Z. Xue and H. Zou. Rank-based tapering estimation of bandable correlation matrices. *Sta-*

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0337

tistica Sinica, 24(1):83–100, 2014.

F. Yi and H. Zou. Sure-tuned tapering estimation of large covariance matrices. *Computational Statistics and Data Analysis*, 58(1):339–351, 2013.

M.Y. Zhang, F. Rubio, and D.P. Palomar. Improved calibration of high-dimensional precision matrices. *IEEE Transactions on Signal Processing*, 61(6):1509–1519, 2013.

J. L. Zhao, H. Y. Zhao, and L. X. Zhu. Pivotal variable detection of the covariance matrix and its application to high-dimensional factor models. *Statistics and Computing*, 28(4): 775–793, 2018.

X. H. Zhu, X. Guo, T. Wang, and L. X. Zhu. Dimensionality determination: a thresholding double ridge ratio criterion. *Computational Statistics and Data Analysis*, 146:106910, 2020.