Machine Learning
https://doi.org/10.1007/s10994-022-06138-3

High-dimensional correlation matrix estimation for general continuous data with Bagging technique

Chaojie Wang[1] · **Jin Du**[2] · **Xiaodan Fan**[2]

Received: 3 January 2020 / Revised: 24 November 2021 / Accepted: 9 January 2022

Abstract
High-dimensional covariance matrix estimation plays a central role in multivariate statistical analysis. It is well-known that the sample covariance matrix is singular when the sample size is smaller than the dimension of the variable, but the covariance estimate must be positive-definite. This motivates some modifications of the sample covariance matrix to preserve its efficient estimation of pairwise covariance. In this paper, we modify the sample correlation matrix using the Bagging technique. The proposed Bagging estimator is flexible for general continuous data. Under some mild conditions, we show theoretically that the Bagging estimator can ensure positive-definiteness with probability one in finite samples. We also prove the consistency of the bootstrap estimator of Pearson correlation and the consistency of our Bagging estimator when the dimension p is fixed. Simulation results and a real application are provided to demonstrate that our method strikes a better balance between RMSE and likelihood, and is more robust, than other existing estimators.

Keywords Bagging technique · Random matrix · Sample correlation matrix · Positive-definiteness

Editor: Pradeep Ravikumar.

✉ Xiaodan Fan
xfan@cuhk.edu.hk

Chaojie Wang
cjwang@ujs.edu.cn

Jin Du
jinyangdu@cuhk.edu.hk

[1] The Fourth Affiliated Hospital of Jiangsu University, School of Mathematical Science, Jiangsu University, Zhenjiang, China

[2] Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, SAR, China

Published online: 18 March 2022

1 Introduction

Covariance matrix estimation is a fundamental topic in multivariate statistical analyses. Traditionally, the sample covariance matrix is a convenient and efficient estimator when sample size n is much larger than dimension p. However, in recent years, more and more high-dimensional datasets with small n and large p have appeared in various applications. For instance, investors track thousands of assets in the financial market, but there are only hundreds of daily trading observations per year (Bodnar et al., 2018). For cancer diagnosis with genetic data, thousands of gene expressions can be measured using microarray techniques simultaneously, but patient cases are often rare and limited (Best et al., 2015). It is well-known that the sample covariance matrix is singular when $p > n$, but a valid covariance matrix must be positive-definite. This fatal flaw hampers the application of sample covariance matrix in high-dimensional multivariate statistical analyses, including discriminant analysis and regression models. Furthermore, Johnstone (2001) showed that the sample covariance matrix distorts the eigen-structure of the population covariance matrix and is ill-conditioned when p is large. Generally, the sample covariance matrix is an awful estimator in high-dimensional cases.

Although its performance is poor as a whole (Fan et al., 2016), each entry in the sample covariance matrix is still an efficient estimator of pairwise covariance among variables. This motivates the design of a modified version that retains efficient estimation of pairwise covariance, while avoiding the drawbacks. Ledoit and Wolf (2004) proposed a shrinkage method by taking a weighted linear combination of the sample covariance matrix and the identity matrix. The resulting matrix is positive-definite, invertible, and preserves the eigenvector structure. There is existing literature on how to choose the optimal weighted parameter to obtain better asymptotic properties (Ledoit and Wolf, 2004; Mestre and Lagunas, 2005; Mestre, 2008). However, the shrinkage operation leads to a biased estimator in finite samples. If the covariance matrix is sparse, thresholding methods may be the most intuitive idea in high-dimensional analyses. Bickel and Levina (2008) applied the hard-thresholding method to the sample covariance matrix and showed its asymptotic consistency. After that, other generalized thresholding rules were proposed and tried, such as banding (Bickel and Levina, 2008; Wu and Pourahmadi, 2009), soft-thresholding (Rothman et al., 2009), and adaptive thresholding (Cai and Liu, 2011). For further theoretical results, Cai et al. (2010) derived the optimal rate of convergence for estimating the true covariance matrix, and Cai and Zhou (2012) explored the operator norm, Frobenius norm and L_1 norm of the estimator and its inverse. The thresholding idea is an efficient method to obtain a sparse estimator, but it is hard to ensure positive-definiteness for finite samples. In fact, Guillot and Rajaratnam (2012) showed that a thresholded matrix may lose positive-definiteness quite easily. Fan et al. (2016) also demonstrated that the thresholding method sacrifices a great deal of entries and information in the sample covariance matrix to attain positive-definiteness.

From the perspective of random matrix theory, Marzetta et al. (2011) constructed a positive-definite estimator by random dimension reduction. Tucci and Wang (2019) considered a random unitary matrix with Harr measure as an alternative random operator. In this paper, inspired by the work of random matrix theory and some practical considerations, we modify the sample correlation matrix using the Bagging technique. Bagging (**B**ootstrap **Agg**regat**ing**), proposed by Breiman (1996), is an ensemble algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical inference. Surprisingly, we find that the Bagging technique can help achieve a positive-definite estimate

when $p > n$. Through a resampling procedure, the Bagging technique can "create" more linearly independent data to transform the problem into traditional cases where n/p is large. This paper contributes to the field in the following aspects: (a) we propose a new high-dimensional correlation matrix estimator for general continuous data; (b) we prove theoretically the Bagging estimator ensures positive-definiteness with probability one in finite samples, while the estimator is consistent when p is fixed; (c) we demonstrate that the Bagging estimator is competitive with existing approaches through a large number of simulation studies in various scenarios and a real application.

This paper is organized as follow: Sect. 2 proposes the Bagging estimator. Section 3 proves some relevant theoretical results. Section 4 compares our method with existing approaches through simulation studies in various scenarios and Sect. 5 provides a real application. Section 6 concludes the paper.

2 Bagging estimator

For a given training set D of size n, the Bagging technique first generates m new training sets $d_1, \cdots, d_m$, each of size n, by sampling from D uniformly with replacement. This step is called bootstrap sampling. These m bootstrap resampling sets are then fitted separately to produce estimates $h_1, \cdots, h_m$. The individual estimates $h_1, \cdots, h_m$ are then combined by averaging or voting to generate the final estimate h^{Bag}. The procedure of the Bagging algorithm is illustrated in Fig. 1.

Generally, Bagging can improve the stability and accuracy of almost every regression and classification algorithm (Breiman, 1996). In this paper, we use the Bagging technique to modify the sample correlation matrix.

Let $\mathbf{X} = (X_{ij})_{n\times p}$ be the observed dataset. X_{ij} denotes the i-th observation for the j-th variable where $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Assume row vectors $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})$ are i.i.d. for $i = 1, \cdots, n$, and follow a continuous and irreducible p-dimensional distribution with mean $\boldsymbol{\mu}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$, e.g., $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here an irreducible

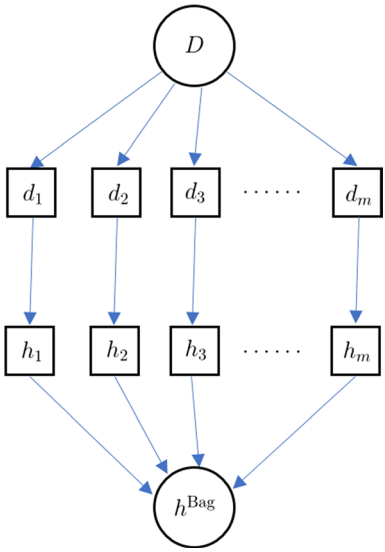


Fig. 1 The procedure of the Bagging algorithm

p-dimensional distribution denotes a p-dimensional distribution where the p components are irreducible (see Definition 5 for details). We are interested in estimating the $p \times p$ covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ for fixed p and finite sample size n when $p > n$. The sample covariance matrix is defined as

$$\mathbf{S} = \frac{1}{n-1}(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}}),$$

where $\bar{\mathbf{X}} = \mathbf{1}_{n\times 1} \cdot (\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i)$ is the matrix of sample mean vectors.

According to the variance-correlation decomposition, $\boldsymbol{\Sigma} = \mathbf{D}\boldsymbol{\Lambda}\mathbf{D}$, where $\mathbf{D}$ is the diagonal matrix of standard deviations and $\boldsymbol{\Lambda}$ is the correlation matrix with diagonal elements equal to 1. Thus, we may estimate $\mathbf{D}$ and $\boldsymbol{\Lambda}$ separately (Barnard et al., 2000). If $\mathbf{D}$ is estimated by the sample variance, i.e., $\hat{\mathbf{D}} = \text{diag}(\mathbf{S})^{1/2}$, then the problem becomes to estimate the correlation matrix $\boldsymbol{\Lambda}$. The corresponding sample version is defined as follows:

Definition 1 (Sample Correlation Matrix) Let $\mathbf{Y} = (Y_{ij})_{n\times p}$ be the matrix normalized from the original dataset $\mathbf{X}$ by columns, i.e., $Y_{ij} = (X_{ij} - \hat{\mu}_j)/\hat{\sigma}_j$ where $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ and $\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \hat{\mu}_j)^2$. Then, the sample correlation matrix $\mathbf{R}$ is defined as

$$\mathbf{R} = \frac{\mathbf{1}}{\mathbf{n-1}}\mathbf{Y}'\mathbf{Y}.$$

Note that $\text{rank}(\mathbf{R}) = n - 1$, thus $\mathbf{R}$ is still singular when $p > n$ and hence not a valid estimator of $\boldsymbol{\Lambda}$. Therefore, a modification on $\mathbf{R}$ is a must.

Definition 2 (Bagging Estimator) For a given dataset $\mathcal{L} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$, consider a simple resampling set of n observations with replacement, e.g., $\mathcal{L}^{(t)} = \{\boldsymbol{X}_1^{(t)}, \cdots, \boldsymbol{X}_n^{(t)}\}$. Using these resampled data construct the matrix $\mathbf{X}^{(t)}$, which is used to form a sample correlation matrix $\mathbf{R}^{(t)}$. Repeat this process independently for T times. Then, the Bagging estimator is defined as $\mathbf{R}^{\text{Bag}} = \frac{\mathbf{1}}{\mathbf{T}}\sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{R}^{(\mathbf{t})}$.

The Bagging algorithm is summarized in Algorithm 1 in detail. The complete algorithm is simple, easy to implement, and requires few assumptions. Common assumptions, such as the data being Gaussian and the covariance matrix being sparse, are unnecessary in our algorithm. Compared with approaches that rely on these assumptions, our Bagging estimator is more flexible for general continuous data.

Algorithm 1 Bagging Algorithm for Correlation Matrix Estimation

1: Given dataset $\mathcal{L} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$.
2: **for** t-th iteration **do**
3: Resample n samples in $\mathcal{L}$ with replacement to construct $\mathbf{X}^{(t)}$.
4: Normalize the matrix $\mathbf{X}^{(t)}$ by columns to obtain $\mathbf{Y}^{(t)}$.
5: Calculate the sample correlation matrix $\mathbf{R}^{(t)}$.
6: **end for**
7: Average the outputs in iterations as Bagging estimator $\mathbf{R}^{\text{Bag}}$.

3 Theoretical properties

3.1 Positive-definiteness

A valid correlation matrix estimator must be positive-definite. As we shall show, our new estimator $\mathbf{R}^{\text{Bag}}$ is positive-definiteness with probability one for finite samples, although each $\mathbf{R}^{(t)}$ is still singular. It should be noted that this "magic" operation works only for the sample correlation matrix $\mathbf{R}$, rather than the sample covariance matrix $\mathbf{S}$. This may partially explain why this simple procedure has not been explored up till now.

For $\mathbf{R}^{\text{Bag}}$, we have the following decomposition,

$$\mathbf{R}^{\text{Bag}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{R}^{(t)} = \frac{1}{(n-1)T}\sum_{t=1}^{T}\mathbf{Y}^{(\mathbf{t})\prime}\mathbf{Y}^{(t)} = \frac{1}{(n-1)T}\mathbf{Z}'\mathbf{Z}, \tag{1}$$

where $\mathbf{Y}^{(t)} = (Y_{ij}^{(t)})_{n\times p}$ is the matrix normalized from the resampled dataset $\mathbf{X}^{(t)}$ by columns, i.e., $Y_{ij}^{(t)} = (X_{ij}^{(t)} - \hat{\mu}_j^{(t)})/\hat{\sigma}_j^{(t)}$, where $\hat{\mu}_j^{(t)} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}^{(t)}$ and $(\hat{\sigma}_j^{(t)})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}^{(t)} - \hat{\mu}_j^{(t)})^2$. Here

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \\ \vdots \\ \mathbf{Y}^{(T)} \end{pmatrix}_{nT\times p}$$

is a random matrix, which contains all resampled observations.

According to Equation (1), it is sufficient to show that $\Pr(\text{rank}(\mathbf{Z}) = p) = 1$ for large T. First, we clarify several definitions regarding random variables for convenience.

Definition 3 (Continuous) A random variable X is said to be continuous if $\Pr(X \in B) = 0$ for any finite or countable set B of points of the real line.

Definition 4 (Irreducible) Let W be a continuous random variable. Given random variables $U_1, \cdots, U_n$, if $W|U_1, \cdots, U_n$ is still a continuous random variable, W is said to be irreducible given $U_1, \cdots, U_n$.

Definition 5 For continuous random variables $U_1, \cdots, U_n$, if every U_i is irreducible given the remaining random variables, we say $U_1, \cdots, U_n$ are irreducible.

Corollary 1 *Let W be a continuous random variable. If W is independent of random variables $U_1, \cdots, U_n$, then W is irreducible given $U_1, \cdots, U_n$.*

Proof If W is independent of $U_1, \cdots, U_n$, then $W|U_1, \cdots, U_n$ is identically distributed with W and is a continuous random variable. □

Definition 6 (Linearly Irreducible) Let W be a continuous random variable. Given random variables $U_1, \cdots, U_n$, if

$$\Pr(W = a_1U_1 + \cdots + a_nU_n|U_1, \cdots, U_n) = 0,$$

for any $a_1, \cdots, a_n \in \mathbb{R}$, W is said to be linearly irreducible given $U_1, \cdots, U_n$.

Definition 7 For continuous random variables $U_1, \cdots, U_n$, if every U_i is linearly irreducible given the remaining random variables, we say $U_1, \cdots, U_n$ are linearly irreducible.

Corollary 2 *Let W be a continuous random variable. If W is irreducible given $U_1, \cdots, U_n$, then W is linearly irreducible given $U_1, \cdots, U_n$.*

Proof By Definition 4, $W|U_1, \cdots, U_n$ is a continuous random variable. So $\Pr(W = a|U_1, \cdots, U_n) = 0$ for any $a \in \mathbb{R}$. In particular, $\Pr(W = a_1U_1 + \cdots + a_nU_n|U_1, \cdots, U_n) = 0$ for any $a_1, \cdots, a_n \in \mathbb{R}$. □

The following lemma provides a criterion for being linearly irreducible (See Appendix A for detailed proofs of Lemmas and Theorems).

Lemma 1 *Let $U_1, \cdots, U_n$ be continuous random variables. If*

$$\Pr(a_1U_1 + \cdots + a_nU_n = 0) = 0$$

for any $a_1, \cdots, a_n \in \mathbb{R}$ which are not all zero, then $U_1, \cdots, U_n$ are linearly irreducible.

Inspired by the rank of the Gaussian ensemble in random matrix theory (Tao and Vu 2010), we show a general result for the rank of a random matrix.

Theorem 1 *For random matrix $\mathbf{M} = (M_{ij})_{q\times p}$, where M_{ij} are continuous random variables, if $\mathbf{M}$ satisfies the following conditions*: (1) *By rows, $M_{i1}, \cdots, M_{ip}$ are linearly irreducible for all i*; (2) *By columns, $M_{1j}, \cdots, M_{qj}$ are linearly irreducible for all j, then we have*

$$\Pr(\text{rank}(\mathbf{M}) = \min(q, p)) = 1.$$

Specifically, consider the rank of random matrix $\mathbf{Z}$,

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Y}^{(1)} \\ \vdots \\ \mathbf{Y}^{(T)} \end{pmatrix} = \begin{pmatrix} \frac{X_{11}^{(1)} - \hat{\mu}_1^{(1)}}{\hat{\sigma}_1^{(1)}} & \frac{X_{12}^{(1)} - \hat{\mu}_2^{(1)}}{\hat{\sigma}_2^{(1)}} & \cdots & \frac{X_{1p}^{(1)} - \hat{\mu}_p^{(1)}}{\hat{\sigma}_p^{(1)}} \\ \vdots & \vdots & & \vdots \\ \frac{X_{n1}^{(1)} - \hat{\mu}_1^{(1)}}{\hat{\sigma}_1^{(1)}} & \frac{X_{n2}^{(1)} - \hat{\mu}_2^{(1)}}{\hat{\sigma}_2^{(1)}} & \cdots & \frac{X_{np}^{(1)} - \hat{\mu}_p^{(1)}}{\hat{\sigma}_p^{(1)}} \\ \vdots & \vdots & & \vdots \\ \frac{X_{11}^{(T)} - \hat{\mu}_1^{(T)}}{\hat{\sigma}_1^{(T)}} & \frac{X_{12}^{(T)} - \hat{\mu}_2^{(T)}}{\hat{\sigma}_2^{(T)}} & \cdots & \frac{X_{1p}^{(T)} - \hat{\mu}_p^{(T)}}{\hat{\sigma}_p^{(T)}} \\ \vdots & \vdots & & \vdots \\ \frac{X_{n1}^{(T)} - \hat{\mu}_1^{(T)}}{\hat{\sigma}_1^{(T)}} & \frac{X_{n2}^{(T)} - \hat{\mu}_2^{(T)}}{\hat{\sigma}_2^{(T)}} & \cdots & \frac{X_{np}^{(T)} - \hat{\mu}_p^{(T)}}{\hat{\sigma}_p^{(T)}} \end{pmatrix}_{Tn\times p}.$$

For simplicity, delete the redundant rows in $\mathbf{Z}$, which does not change the rank of the matrix. The redundancy may come from identical resampling sets, i.e., $\mathbf{Y}^{(t_1)} \equiv \mathbf{Y}^{(t_2)}$, or may come from repetitive observations in the same resampling sets, i.e., $\boldsymbol{X}_{i_1}^{(t)} \equiv \boldsymbol{X}_{i_2}^{(t)} \equiv \boldsymbol{X}_i \in \mathcal{L}^{(t)}$. After eliminating these redundant rows, let $\tilde{T}$ be the number of distinct resampling sets in total T resampling sets, and let q_t be the number of non-repetitive observations in $\mathcal{L}^{(t)}$.

Note that in each resampling set, there exists a perfect linear relationship among non-repetitive rows due to the sample mean $\hat{\mu}_j^{(t)}$, which decreases the degrees of freedom of observations by one. Thus, there are only $q_t - 1$ free observations in each resampling

set. Without loss of generality, assume the first $q_t - 1$ rows in each resampling set are non-repetitive. We have submatrix $\mathbf{G}$ of $\mathbf{Z}$,

$$\mathbf{G} = \begin{pmatrix} \frac{X_{11}^{(1)} - \hat{\mu}_1^{(1)}}{\hat{\sigma}_1^{(1)}} & \frac{X_{12}^{(1)} - \hat{\mu}_2^{(1)}}{\hat{\sigma}_2^{(1)}} & \cdots & \frac{X_{1p}^{(1)} - \hat{\mu}_p^{(1)}}{\hat{\sigma}_p^{(1)}} \\ \vdots & \vdots & & \vdots \\ \frac{X_{q_1-1,1}^{(1)} - \hat{\mu}_1^{(1)}}{\hat{\sigma}_1^{(1)}} & \frac{X_{q_1-1,2}^{(1)} - \hat{\mu}_2^{(1)}}{\hat{\sigma}_2^{(1)}} & \cdots & \frac{X_{q_1-1,p}^{(1)} - \hat{\mu}_p^{(1)}}{\hat{\sigma}_p^{(1)}} \\ \vdots & \vdots & & \vdots \\ \frac{X_{11}^{(\tilde{T})} - \hat{\mu}_1^{(\tilde{T})}}{\hat{\sigma}_1^{(\tilde{T})}} & \frac{X_{12}^{(\tilde{T})} - \hat{\mu}_2^{(\tilde{T})}}{\hat{\sigma}_2^{(\tilde{T})}} & \cdots & \frac{X_{1p}^{(\tilde{T})} - \hat{\mu}_p^{(\tilde{T})}}{\hat{\sigma}_p^{(\tilde{T})}} \\ \vdots & \vdots & & \vdots \\ \frac{X_{q_{\tilde{T}}-1,1}^{(\tilde{T})} - \hat{\mu}_1^{(\tilde{T})}}{\hat{\sigma}_1^{(\tilde{T})}} & \frac{X_{q_{\tilde{T}}-1,2}^{(\tilde{T})} - \hat{\mu}_2^{(\tilde{T})}}{\hat{\sigma}_2^{(\tilde{T})}} & \cdots & \frac{X_{q_{\tilde{T}}-1,p}^{(\tilde{T})} - \hat{\mu}_p^{(\tilde{T})}}{\hat{\sigma}_p^{(\tilde{T})}} \end{pmatrix}_{\sum_{t=1}^{\tilde{T}}(q_t-1)\times p}.$$

Here submatrix $\mathbf{G} = (G_{ij}^{(t)})$ has the same rank as $\mathbf{Z}$, where $G_{ij}^{(t)} = \frac{X_{ij}^{(t)} - \hat{\mu}_j^{(t)}}{\hat{\sigma}_j^{(t)}}$, $i = 1, \cdots, q_t - 1$, $j = 1, \cdots, p, t = 1, \cdots, \tilde{T}$.

Lemma 2 *$G_{ij}^{(t)}$ is a continuous random variable.*

According to Theorem 1 and Lemma 2, we show $\Pr(\text{rank}(\mathbf{G}) = \min(\sum_{t=1}^{\tilde{T}}(q_t - 1), p)) = 1$.

Theorem 2 *For random matrix $\mathbf{G}$, we have*

$$\Pr(\text{rank}(\mathbf{G}) = \min(\sum_{t=1}^{\tilde{T}}(q_t - 1), p)) = 1.$$

The total number of distinct sets is $\binom{n+k-1}{k}$ if we draw k samples from n different elements with replacement (Pishro-Nik 2016). Here we have $k = n$ in our Bagging algorithm. Thus, the number of distinct resampling sets $\tilde{T}$ goes to $\binom{2n-1}{n}$ with probability 1 as $T \to \infty$.

Since there are $q_t - 1$ free observations in each resampling set and $q_t - 1 \geq 1$ holds except for the n sets in which the elements are all the same, we have $\sum_{t=1}^{\tilde{T}}(q_t - 1) \geq \tilde{T} - n$. Thus, $\sum_{t=1}^{\tilde{T}}(q_t - 1) \geq \binom{2n-1}{n} - n$ as $T \to \infty$. Even if n is small, $\binom{2n-1}{n}$ can be quite large. For example, when $n = 30$, $\binom{2n-1}{n} \approx 5.9 \times 10^{16}$. Thus, even in the cases where $p \gg n$, we still have $\Pr(\text{rank}(\mathbf{Z}) = p) = 1$ as long as $\binom{2n-1}{n} - n > p$.

In practice, it does not need too many resampling times T to ensure the full rank. Let $\tau = \binom{2n-1}{n}$ and consider resampling p times i.e., $T = p$. Note that the number of resampling sets with rank at least 1 is $\tau - n$. The probability of obtaining p distinct resampling sets with rank at least 1 is

$$\frac{\tau-n}{\tau}\cdot\frac{\tau-n-1}{\tau}\cdots\frac{\tau-n-p+1}{\tau}=\prod_{i=0}^{p-1}(1-\frac{n+i}{\tau})$$
$$\geq (1-\frac{n+p-1}{\tau})^{p-1}=1-\frac{(n+p-1)(p-1)}{\tau}+o(\frac{n+p-1}{\tau}),$$

where $o(\frac{n+p-1}{\tau})$ denotes a higher order term of $\frac{n+p-1}{\tau}$. Since $\tau \gg n$ and $\tau \gg p$ (e.g., for $n=30$, $\tau \approx 5.9\times 10^{16}$), then $\frac{n+p-1}{\tau}$ is close to 0. Thus, the probability is quite close to 1. It illustrates that we could obtain a full rank matrix with only p resampling times with high probability. Since $\text{rank}(\mathbf{Z})=\text{rank}(\mathbf{R}^{\text{Bag}})$, we have $\Pr(\text{rank}(\mathbf{R})=p)=1$ and thus our $\mathbf{R}^{\text{Bag}}$ is not singular.

It is worth mentioning that if we estimate the covariance matrix directly rather than the correlation matrix, i.e., without the standardization step, the Bagging estimator is not positive-definite. Similarly to the decomposition in Equation (1), we have

$$\mathbf{S}^{\text{Bag}}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{S}^{(t)}=\frac{1}{(n-1)T}\sum_{t=1}^{T}(\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)})'(\mathbf{X}^{(t)}-\bar{\mathbf{X}}^{(t)}).$$

The corresponding random matrix $\tilde{\mathbf{Z}}$ is

$$\tilde{\mathbf{Z}}=\begin{pmatrix} X_{11}^{(1)}-\hat{\mu}_1^{(1)} & X_{12}^{(1)}-\hat{\mu}_2^{(1)} & \cdots & X_{1p}^{(1)}-\hat{\mu}_p^{(1)} \\ \vdots & \vdots & & \vdots \\ X_{n1}^{(1)}-\hat{\mu}_1^{(1)} & X_{n2}^{(1)}-\hat{\mu}_2^{(1)} & \cdots & X_{np}^{(1)}-\hat{\mu}_p^{(1)} \\ \vdots & \vdots & & \vdots \\ X_{11}^{(T)}-\hat{\mu}_1^{(T)} & X_{12}^{(T)}-\hat{\mu}_2^{(T)} & \cdots & X_{1p}^{(T)}-\hat{\mu}_p^{(T)} \\ \vdots & \vdots & & \vdots \\ X_{n1}^{(T)}-\hat{\mu}_1^{(T)} & X_{n2}^{(T)}-\hat{\mu}_2^{(T)} & \cdots & X_{np}^{(T)}-\hat{\mu}_p^{(T)} \end{pmatrix}_{Tn\times p}=\mathbf{AX},$$

where $\mathbf{A}$ is a $Tn\times n$ constant matrix. This means $\tilde{\mathbf{Z}}$ is only a linear transformation of $\mathbf{X}$. We have

$$\text{Rank}(\tilde{\mathbf{Z}})\leq\text{Rank}(\mathbf{X})=n.$$

Thus, the Bagging sample covariance matrix is still singular.

3.2 Mean squared error

In addition to the guarantee of positive-definiteness, our Bagging estimator $\mathbf{R}^{Bag}$ performs well in terms of mean squared error (MSE). The MSE of a matrix estimator is defined by the Frobenius norm, i.e.,

$$\text{MSE}(\hat{\boldsymbol{\Lambda}})=||\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}||_F^2=\sum_{i,j}(\hat{\lambda}_{ij}-\lambda_{ij})^2,$$

where $||\cdot||_F$ is the Frobenius norm of a matrix, $\hat{\boldsymbol{\Lambda}}=(\hat{\lambda}_{ij})_{p\times p}$ and $\boldsymbol{\Lambda}=(\lambda_{ij})_{p\times p}$ are the estimated and true correlation matrix respectively.

For the sample correlation matrix $\mathbf{R}=(r_{ij})_{p\times p}$, the MSE of $\mathbf{R}$ is

$$\text{MSE}(\mathbf{R}) = E||\mathbf{R} - \boldsymbol{\Lambda}||_F^2 = E\sum_{i,j}(r_{ij} - \lambda_{ij})^2 = \sum_{i,j}\text{MSE}(r_{ij}).$$

Although the performance of the sample correlation matrix is poor as a whole when $p > n$ due to being singular, each entry of it is still an efficient estimator of pairwise covariance among variables. We next show that our Bagging estimator is consistent when p is fixed.

Theorem 3 *The mean squared error of r_{ij}^{Bag} is no more than the average of mean-squared error of $r_{ij}^{(t)}$, i.e.,*

$$\text{MSE}(r_{ij}^{\text{Bag}}) \leq \frac{1}{T}\sum_{t=1}^{T}\text{MSE}(r_{ij}^{(t)}),$$

where $r_{ij}^{(t)}$ denotes the i-th row and j-th column entry of $\mathbf{R}^{(\mathbf{t})}$.

Since each resampling set $\mathcal{L}^{(t)}$ has the identical distribution, Theorem 3 leads to $\text{MSE}(r_{ij}^{Bag}) \leq \text{MSE}(r_{ij}^{(t)})$ directly. Thus, it is sufficient to show that $r_{ij}^{(t)}$ is a consistent estimator, which further leads to $\text{MSE}(r_{ij}^{(t)}) \to 0$ as n goes into infinity.

For a general bivariate distribution (X, Y) with finite forth moments, Lehmann (1999) showed that the limit distribution of $\sqrt{n}(r_{XY} - \rho)$ is asymptotically normal with mean 0 and constant variance, where r_{XY} is the sample correlation coefficient and ρ is the true value of correlation coefficient. It also implies that r_{XY} is a consistent estimator of ρ. Here we proposed its bootstrap version to show that $r_{XY}^{(t)}$ is asymptotically consistent.

Theorem 4 *Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be i.i.d. according to some bivariate distribution (X, Y), which has finite forth moments, with means $E(X) = \xi$, $E(Y) = \eta$, variances $\text{Var}(X) = \sigma^2$, $\text{Var}(Y) = \tau^2$, and correlation coefficient ρ. Let $(X_1^{(t)}, Y_1^{(t)}), \cdots, (X_n^{(t)}, Y_n^{(t)})$ be the t-th bootstrap resampling set. The bootstrap sample correlation is defined as*

$$r_{XY}^{(t)} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}^{(t)})(Y_i^{(t)} - \bar{Y}^{(t)})}{S_X^{(t)}S_Y^{(t)}},$$

where

$$\bar{X}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}X_i^{(t)}, \quad \bar{Y}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}Y_i^{(t)},$$

$$(S_X^{(t)})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}^{(t)})^2, \quad (S_Y^{(t)})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i^{(t)} - \bar{Y}^{(t)})^2.$$

Then, as n goes to infinity, the bootstrap sample correlation $r_{XY}^{(t)}$ is a consistent estimator of ρ.

By Theorems 3 & 4, we have the following corollary.

Corollary 3 *Under the mild condition that the p-dimensional distribution has finite forth moments, MSE of the Bagging estimator converges to zero, i.e.,*

$$\text{MSE}(\mathbf{R}^{\text{Bag}}) \leq \text{MSE}(\mathbf{R}^{(t)}) \to 0,$$

as $n \to \infty$ for fixed p. It implies that the Bagging estimator $\mathbf{R}^{\text{Bag}}$ is consistent.

4 Simulations

In this section, simulation studies are presented to compare the performance of the Bagging estimator with other classic approaches, including graphical lasso (glasso, Friedman et al., 2008), the hard-threshold method (H-threshold, Bickel and Levina, 2008), the shrinkage estimator (Ledoit and Wolf, 2004 and the traditional sample correlation matrix. Two criteria are used to evaluate the performance of estimators: comparable log-likelihood ℓ and root-mean-square error (RMSE). Log-likelihood measures the fitness of observed data, which depends on the assumed distribution. Here comparable log-likelihood ℓ is the core of the log-likelihood function with common constant terms omitted. RMSE measures the difference between the true values and estimators. The RMSE of an estimator is defined as follows:

$$\text{RMSE}(\hat{\boldsymbol{\Lambda}}) = \frac{1}{p}||\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}||_F = \frac{1}{p}\sqrt{\sum_{i,j}(\hat{\lambda}_{ij} - \lambda_{ij})^2},$$

where $||\cdot||_F$ is the Frobenius norm of a matrix, $\hat{\boldsymbol{\Lambda}} = (\hat{\lambda}_{ij})_{p\times p}$ and $\boldsymbol{\Lambda} = (\lambda_{ij})_{p\times p}$ are the estimated and true correlation matrix respectively.

In the following simulation studies, we synthesize data from assumed distributions with known correlation matrix. The true correlation matrix is generated as follows:

$$\boldsymbol{\Sigma} = A'A \quad \text{and} \quad \boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Sigma})^{-1/2}\boldsymbol{\Sigma}\text{diag}(\boldsymbol{\Sigma})^{-1/2} \tag{2}$$

where $A = (a_{ij})_{p\times p}$, $a_{ij} \sim \text{Unif}(-1, 1)$ are i.i.d for $i, j = 1, \cdots, p$. The randomly generated correlation matrices are positive-definite and symmetric. They are general correlation matrices without any special structures.

Then, we obtain the estimated covariance matrix using generated data sets. Considering the uncertainty of Monte Carlo simulations, we repeat the experiments, including generation of random covariance matrices and data synthesis, 100 times independently in each setting. The means and standard errors of ℓ and RMSE are reported for comparison. See the supplementary materials for the detailed R codes.

4.1 Case 1: multivariate Gaussian data

In this case, the data sets are generated from a multivariate Gaussian distribution with mean zero and a general correlation matrix. Here the true correlation matrix is generated randomly according to Equation (2). Table 1 presents the means and standard errors of ℓ_N and RMSE in the case of $p = 50, n = 20$ and $p = 200, n = 100$ respectively.

The only required parameter in the Bagging estimator is the resampling times T. In practice, increasing the resampling times may improve the accuracy of estimation. Figure 2, which is from one of following simulation studies, demonstrates the relationship between T and RMSE. At the beginning, the RMSE of the estimator decays with the increase of T and then converges to a stable level. In the following simulation studies, T is set as 100 to balance accuracy of estimation and computation cost.

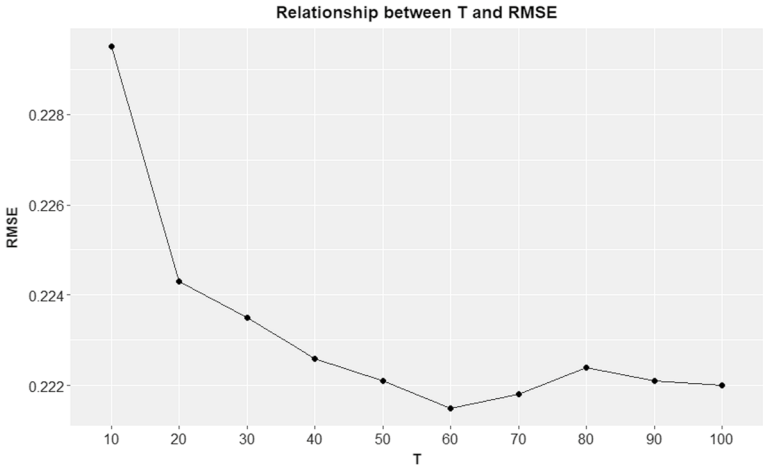


Fig. 2 The RMSE of Bagging estimator decays with the increase of T at the beginning and then seemingly converges to a stable level

Table 1 The means and standard errors of two criteria across 100 independent experiments for multivariate Gaussian data

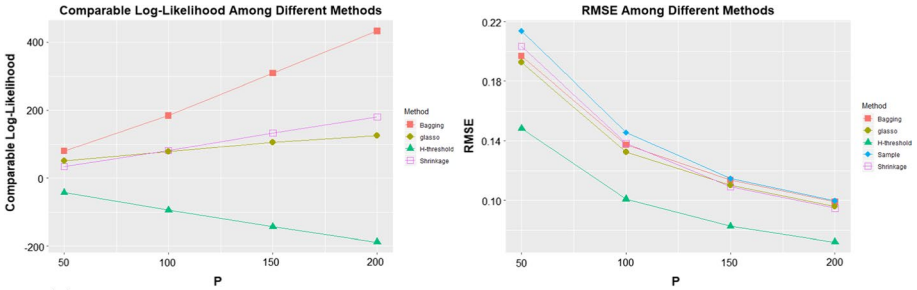
	$p = 50, n = 20$		$p = 200, n = 100$	
	ℓ_N	RMSE	ℓ_N	RMSE
Bagging	94.46(2.31)	0.2192(0.0075)	435.19(2.66)	0.0998(0.0009)
H-threshold	−44.74(1.19)	0.1422(0.0028)	−189.64(1.40)	0.0718(0.0004)
Shrinkage	46.61(3.18)	0.2264(0.0152)	180.71(1.48)	0.0965(0.0016)
glasso	71.61(0.96)	0.2186(0.0078)	126.21(1.28)	0.0964(0.0010)
Sample	–	0.2384(0.0163)	–	0.1016(0.0018)

From Table 1, we find that the hard-threshold method sacrifices much information of the covariance matrix to attain positive-definiteness. The comparable log-likelihood of the thresholded estimator is quite low, though its RMSE performs well. Our Bagging estimator has significant advantages over compared approaches on comparable log-likelihood ℓ_N. This demonstrates that the Bagging estimator fits the observed data better. Note that ℓ_N of the sample correlation estimator would be infinite when $p > n$ due to the estimator being singular, making the estimator invalid. For RMSE, the performance of Bagging and glasso are close, and better than the shrinkage estimator and the sample correlation estimator; but not as good as the H-threshold estimator.

The results of more scenarios under different settings are shown in Fig. 3. Here the sample size n is set as $n = p/2$ varying with the number of variables p. In summary, the Bagging estimator strikes a better balance between RMSE and likelihood.

4.2 Case 2: multivariate *t*-distribution data

Besides traditional multivariate Gaussian data, the Bagging estimator also works on general continuous distributions, such as multivariate t-distributions. In the following



(a) Comparable log-likelihood **(b)** RMSE

Fig. 3 **a** For comparable log-likelihood ℓ_N, our Bagging estimator beats others significantly across all values of p. **b** For RMSE, the Bagging estimator is second only to the hard-threshold method, which has the worst performance from the perspective of ℓ_N

simulation studies, data are generated from the multivariate t-distribution with mean zero and a general correlation matrix, which is still randomly generated from Equation (2). The multivariate t-distribution is a generalization to random vectors of Student's t-distribution (Genz and Bretz, 2009). The density function is defined as

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Lambda},\upsilon)=\frac{\Gamma[(\upsilon+p)/2]}{\Gamma(\upsilon/2)\upsilon^{p/2}\pi^{p/2}|\boldsymbol{\Lambda}|^{1/2}}\left[1+\frac{1}{\upsilon}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})\right]^{-(\upsilon+p)/2}.$$

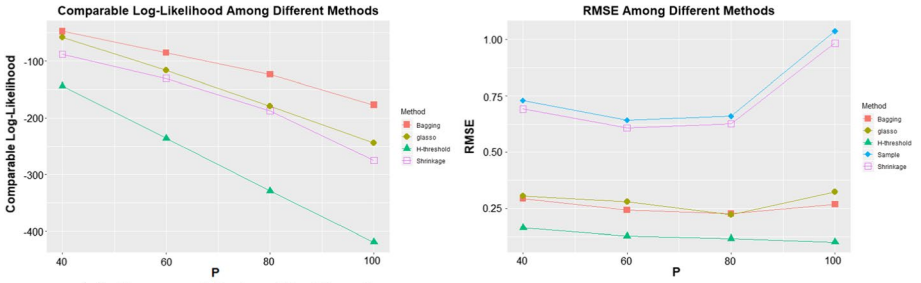
where $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are the mean vector parameter and correlation matrix parameter respectively. Here υ denotes the degrees of freedom of the distribution. As $\upsilon\to\infty$, the multivariate t-distribution converges to the multivariate Gaussian distribution asymptotically. So the degrees of freedom υ is set to 3 to distinguish from the Gaussian cases. The resampling times T is still set as 100, the same as in Sect. 4.1. Table 2 presents the means and standard errors of ℓ_t and RMSE in the case of $p=50, n=20$ and $p=100, n=50$.

More scenarios under different settings are explored in Fig. 4. Also, the sample size n is set as $n=p/2$.

Table 2 and Fig. 4 draw similar conclusions to those in Table 1 and Fig. 3. They demonstrate that our Bagging estimator is not only suitable for Gaussian data, but also can be applied to non-Gaussian data.

Table 2 The means and standard errors of two criteria across 100 independent experiments for Multivariate t-distribution data ($\upsilon=3$)

	$p=50, n=20$		$p=100, n=50$	
	ℓ_t	RMSE	ℓ_t	RMSE
Bagging	−48.15(4.13)	0.2921(0.0496)	−169.11(5.47)	0.2171(0.0398)
H-threshold	−185.61(8.15)	0.1432(0.0034)	−424.73(15.78)	0.1005(0.0012)
Shrinkage	−108.91(9.37)	1.2597(1.7466)	−261.19(10.43)	0.7906(0.5331)
glasso	−61.47(4.92)	0.3218(0.0751)	−247.09(10.72)	0.2378(0.0583)
Sample	–	1.3275(1.8385)	–	0.8333(0.5611)



(a) Comparable log-likelihood **(b)** RMSE

Fig. 4 **a** For comparable log-likelihood ℓ_t, our Bagging estimator beats others significantly across all values of p. **b** For RMSE, the Bagging estimator is second only to the hard-threshold method, which has the worst performance from the perspective of ℓ_t

5 Application

This section presents a real application to demonstrate the performance of our estimator. The original dataset, contributed by Bhattacharjee et al. (2001), is a famous gene expression dataset on lung cancer patients. It contains 203 specimens, including 139 adenocarcinomas resected from the lung ("AD" samples) and 64 other samples, and 12,600 transcript sequences. Here we focus on the 139 "AD" samples ($n = 139$) and assume they are independent, identically distributed, and follow a Gaussian distribution. For simplicity, we use a standard deviation threshold of 500 expression units to select the 186 most variable transcript sequences ($p = 186$). Then, a subset of 70 "AD" samples are sampled randomly without replacement to form a covariance matrix estimator. We repeat the experiments and the sampling procedure for 100 times independently. The comparable log-likelihood and RMSE for different covariance matrix estimations are summarized in Table 3, where RMSE is calculated using the sample covariance matrix of the full 139 samples instead of the unknown "true" covariance matrix. It shows our Bagging estimator has significant advantages over other estimators in terms of likelihood, and is competitive in terms of RMSE.

Figure 5 presents the sample correlation matrix of the full 139 samples and the Bagging estimator with a subset of 70 samples in one of experiments. It demonstrates that our Bagging estimator is quite close to the "true" value.

Table 3 Means and standard errors of two criteria across 100 independent experiments

	ℓ_N	RMSE
Bagging	518.27(7.50)	0.0875(0.0056)
H-threshold	−133.16(33.14)	0.1930(0.0018)
Shrinkage	274.45(1.87)	0.0841(0.0055)
glasso	262.45(3.04)	0.0864(0.0056)
Sample	–	0.0879(0.0056)

6 Summary

In this paper, we propose a novel approach to estimate high-dimensional correlation matrices when $p > n$ with finite samples. Through the procedure of Bootstrap resampling, we show that the Bagging estimator ensures positive-definiteness with probability one in finite samples. Furthermore, our estimator is flexible for general continuous data under some mild conditions. The common assumptions in analogous problems, such as sparse structure and having a Gaussian distribution, are unnecessary in our framework. Through simulation studies and a real application, our method is demonstrated to strike a better balance between RMSE and likelihood. The selected four approaches for comparison represent different but classical ideas to solve the high-dimensional covariance matrix problem; so the results are representative.

It should be noted that our Bagging estimator is devoted to solving problems with little prior knowledge. If one has the prior information on the structure of the covariance matrix, e.g., block or banding, specific approaches are certainly better than our general method. The choice of estimation method still depends on specific scenarios and applications. Some theoretical aspects can be explored further in future research, e.g., the convergence rate of the Bagging estimator when both p and n go to infinity.

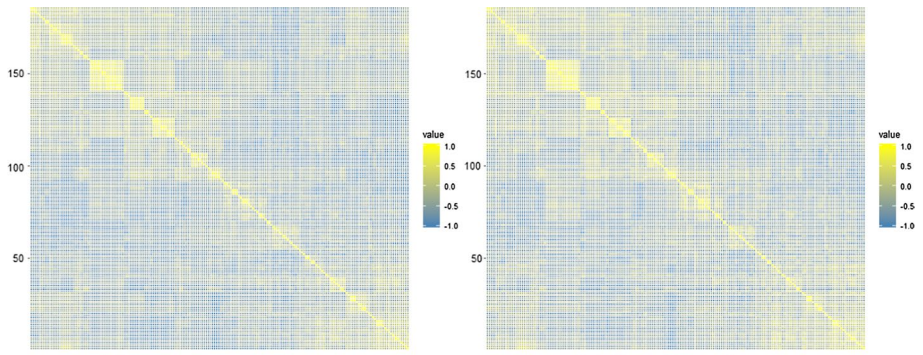
Appendix A: proofs of Lemmas & Theorems

Proof of Lemma 1 Without loss of generality, assume $a_1 \neq 0$. So

$$\Pr(U_1 = -\sum_{i=2}^{n} \frac{a_i}{a_1} U_i) = 0.$$

This implies that

$$\Pr(U_1 = \sum_{i=2}^{n} b_i U_i | U_2, \cdots, U_n) = 0,$$



(a) "True" correlation matrix **(b)** Bagging estimator

Fig. 5 **a** Heat map of the sample correlation matrix of the full 139 samples, which is viewed as "true" correlation matrix for comparison. **b** Heat map of the Bagging estimator on a subset of 70 samples

where $b_i = -a_i/a_1$. Here b_i can be any value in $\mathbb{R}$ since the equation holds for any $a_i \in \mathbb{R}$. Since U_1 is a continuous random variable, by Definition 4, U_1 is linearly irreducible given remaining random variables.

Similarly, we have every U_i is linearly irreducible given remaining random variables. Thus, $U_1, \cdots, U_n$ are linearly irreducible. □

Proof of Theorem 1 If $q \le p$, we need to show that $\Pr(\text{rank}(\mathbf{M}) = q) = 1$. Construct a square submatrix $\mathbf{G}_{q\times q}$ using the first q columns of $\mathbf{M}$. Since $\mathbf{G}$ is a square matrix, $\Pr(\text{rank}(\mathbf{G}) = q) = 1$ means that $\mathbf{G}$ is singular with probability 0, i.e., $\Pr(\det(\mathbf{G}) = 0) = 0$. $\mathbf{G}$ is singular if and only if $\boldsymbol{G}_i$ lies in the span of $\boldsymbol{G}_1, \cdots, \boldsymbol{G}_{i-1}$ for some i, where $\boldsymbol{G}_i$ may be a row or column vector of $\mathbf{G}$. Since this theorem is symmetric by rows or columns, we assume $\boldsymbol{G}_i$ are row vectors of $\mathbf{G}$ without loss of generality. Thus,

$$\Pr(\det(\mathbf{G}) = 0) \le \sum_{i=1}^{q} \Pr(\boldsymbol{G}_i \in V_i),$$

where $V_i := \text{span}(\boldsymbol{G}_1, \cdots, \boldsymbol{G}_{i-1})$. Here we define V_1 the null space. Obviously, we have $\Pr(\boldsymbol{G}_1 \in V_1) = 0$. Then we show $\Pr(\boldsymbol{G}_i \in V_i) = 0$ for any $1 < i \le q$.

According to condition (2), $G_{1j}, \cdots, G_{qj}$ are linearly irreducible for all j. So, for any $1 < i \le q$, $G_{1j}, \cdots, G_{ij}$ are linearly irreducible for all j. This means G_{ij} is linearly irreducible given $G_{1j}, \cdots, G_{i-1,j}$ for all j. By Definition 4, we have

$$\Pr(G_{ij} = a_1 G_{1j} + \cdots + a_{i-1} G_{i-1,j} | G_{1j}, \cdots, G_{i-1,j}) = 0$$

holds for all j and for any $a_1, \cdots, a_{i-1}$. Thus, $\Pr(\boldsymbol{G}_i \in V_i | \boldsymbol{G}_1, \cdots, \boldsymbol{G}_{i-1}) = 0$ holds for any $1 < i \le q$.

By integrating $\boldsymbol{G}_1, \cdots, \boldsymbol{G}_{i-1}$ out, we get

$$\Pr(\boldsymbol{G}_i \in V_i) = 0$$

for any i. Thus,

$$\Pr(\det(\mathbf{G}) = 0) \le \sum_{i=1}^{q} \Pr(\boldsymbol{G}_i \in V_i) = 0$$

as desired. So, we have $\Pr(\text{rank}(\mathbf{G}) = q) = 1$. And since $\text{rank}(\mathbf{G}) \le \text{rank}(\mathbf{M}) \le q$, then $\Pr(\text{rank}(\mathbf{M}) = q) = 1$.

If $q > p$, we take the first p rows of $\mathbf{M}$ to construct a square submatrix $\mathbf{G}_{p\times p}$. Similarly we have $\Pr(\text{rank}(\mathbf{G}) = p) = 1$. And since $\text{rank}(\mathbf{G}) \le \text{rank}(\mathbf{M}) \le p$, then $\Pr(\text{rank}(\mathbf{M}) = p) = 1$.

Thus, generally, we have $\Pr(\text{rank}(\mathbf{M}) = \min(q, p)) = 1$. □

Proof of Lemma 2 Note that

$$G_{ij}^{(t)} = \frac{X_{ij}^{(t)} - \hat{\mu}_j^{(t)}}{\hat{\sigma}_j^{(t)}},$$

is a function of $X_{1j}^{(t)}, \cdots, X_{q_t,j}^{(t)}$, where $X_{1j}^{(t)}, \cdots, X_{q_t,j}^{(t)}$ are independent continuous random variables.

Let $X_{1j}^{(t)} = X$. Given $X_{2j}^{(t)}, \cdots, X_{q_t,j}^{(t)}$, for any $b \in \mathbb{R}$, we have

$$
\begin{aligned}
&\Pr(G_{ij}^{(t)} = b | X_{2j}^{(t)}, \cdots, X_{q_t j}^{(t)}) \\
=& \Pr(\frac{X_{ij}^{(t)} - \frac{1}{n}\sum_{i=1}^{n} X_{ij}^{(t)}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}^{(t)} - \frac{1}{n}\sum_{k=1}^{n} X_{kj}^{(t)})^2}} = b | X_{2j}^{(t)}, \cdots, X_{q_t j}^{(t)}) \\
=& \Pr(\frac{AX+B}{\sqrt{CX^2+DX+E}} = b | X_{2j}^{(t)}, \cdots, X_{q_t j}^{(t)}),
\end{aligned}
$$

where $CX^2 + DX + E > 0$ for all X, $A \neq 0$, and A, B, C, D, E are constants. Consider that

$$
\begin{aligned}
&\frac{AX+B}{\sqrt{CX^2+DX+E}} = b \\
\Rightarrow& AX + B = b\sqrt{CX^2+DX+E} \\
\Rightarrow& (A^2 - b^2C)X^2 + (2AB - b^2D)X + (B^2 - b^2E) = 0,
\end{aligned}
$$

there are at most two zero points in solution space $\boldsymbol{\Omega}$. Since $X = X_{1j}^{(t)}$ is independent of $X_{2j}^{(t)}, \cdots, X_{q_t j}^{(t)}$, and X is a continuous random variable, by Corollary 1, $X | X_{2j}^{(t)}, \cdots, X_{q_t j}^{(t)}$ is a continuous random variable. For any finite set $\boldsymbol{\Omega}$, $\Pr(X \in \boldsymbol{\Omega}) = 0$. So, we have

$$
\Pr(G_{ij}^{(t)} = b | X_{2j}^{(t)}, \cdots, X_{q_t j}^{(t)}) = 0.
$$

By integrating $X_{2j}^{(t)}, \cdots, X_{q_t j}^{(t)}$ out, we have $\Pr(G_{ij}^{(t)} = b) = 0$ for any $b \in \mathbb{R}$. Then for any finite or countable set B of points of the real line, we have $\Pr(G_{ij}^{(t)} \in B) = 0$. □

Proof of Theorem 2 According to Theorem 1 and Lemma 1, we only need to check that $\mathbf{G}$ satisfies the two conditions: (1) By rows, $G_{i1}^{(t)}, \cdots, G_{ip}^{(t)}$ are linearly irreducible for all i, t; (2) By columns, $G_{1j}^{(1)}, \cdots, G_{q_1-1,j}^{(1)}, \cdots, G_{1j}^{(\bar{T})}, \cdots, G_{q_{\bar{T}}-1,j}^{(\bar{T})}$ are linearly irreducible for all j.

(1): Note that if $\boldsymbol{X}_i^{(t)} = (X_{i1}^{(t)}, \cdots, X_{ip}^{(t)})$ follows a continuous and irreducible p-dimensional distribution, then $\{X_{i1}^{(t)}, \cdots, X_{ip}^{(t)}\}$ are irreducible. Since $\boldsymbol{X}_i^{(t)}, i = 1, \cdots, q_t$, are independent random vectors, according to Corollary 1, we have

$$
\mathcal{X}^{(t)} = \{X_{11}^{(t)}, \cdots, X_{1p}^{(t)}, \cdots, X_{q_t,1}^{(t)}, \cdots, X_{q_t p}^{(t)}\}
$$

are irreducible. For any $a_1, \cdots, a_p \in \mathbb{R}$, not all zero, we explore the probability of the equation

$$
a_1 G_{i1}^{(t)} + \cdots + a_p G_{ip}^{(t)} = 0.
$$

Without loss of generality, assume $a_1 \neq 0$. Let $X_{11}^{(t)} = X$. Given $\mathcal{X}^{(t)} \backslash X_{11}^{(t)}$, we have

$$
\begin{aligned}
&\Pr(a_1 G_{i1}^{(t)} + \cdots + a_p G_{ip}^{(t)} = 0 | \mathcal{X}^{(t)} \backslash X_{11}^{(t)}) \\
=& \Pr(G_{i1}^{(t)} = F | \mathcal{X}^{(t)} \backslash X_{11}^{(t)}) \\
=& \Pr(\frac{AX+B}{\sqrt{CX^2+DX+E}} = F | \mathcal{X}^{(t)} \backslash X_{11}^{(t)}),
\end{aligned}
$$

where $CX^2 + DX + E > 0$ for all X, $A \neq 0$ and A, B, C, D, E, F are constants. Since X is irreducible given $\mathcal{X}^{(t)} \backslash X_{11}^{(t)}$, similarly with the proof of Lemma 2, we have

$$\Pr(a_1 G_{i1}^{(t)} + \cdots + a_p G_{ip}^{(t)} = 0 | \mathcal{X}^{(t)} \backslash X_{11}^{(t)}) = 0.$$

By integrating $\mathcal{X}^{(t)} \backslash X_{11}^{(t)}$ out, we have $\Pr(a_1 G_{i1}^{(t)} + \cdots + a_p G_{ip}^{(t)} = 0) = 0$. According to Lemma 2, $G_{ij}^{(t)}$ are continuous random variable. Then, by Lemma 1, we have $G_{i1}^{(t)}, \cdots, G_{ip}^{(t)}$ are linearly irreducible for all i, t. The random matrix $\mathbf{G}$ satisfies condition (1).

(2) Within the t-th resampling set, there are q_t different independent samples $\boldsymbol{X}_1^{(t)}, \cdots, \boldsymbol{X}_{q_t}^{(t)}$. For any t and column j, $G_{1j}^{(t)}, \cdots, G_{q_t j}^{(t)}$ come from independent samples $X_{1j}^{(t)}, \cdots, X_{q_t j}^{(t)}$ with a linear constraint that $\sum_{i=1}^{q_t} n_{ij}^{(t)} G_{ij}^{(t)} = 0$, where $n_{ij}^{(t)} \geq 1$ denotes the number of repeated observations $G_{ij}^{(t)}$ and $\sum_{i=1}^{q_t} n_{ij}^{(t)} = n$. So, without loss of generality, we can show that the first $q_t - 1$ elements that $G_{1j}^{(t)}, \cdots, G_{q_t - 1, j}^{(t)}$ are linearly irreducible. For any $a_1^{(t)}, \cdots, a_{q_t - 1}^{(t)} \in \mathbb{R}$, since $X_{1j}^{(t)}, \cdots, X_{q_t j}^{(t)}$ are independent and irreducible,

$$\Pr\left(\sum_{i=1}^{q_t - 1} a_i^{(t)} G_{ij}^{(t)} = 0\right) = \Pr\left(\sum_{i=1}^{q_t - 1} a_i^{(t)} \frac{X_{ij}^{(t)} - \hat{\mu}_j^{(t)}}{\hat{\sigma}_j^{(t)}} = 0\right) = \Pr\left(\sum_{i=1}^{q_t - 1} a_i^{(t)} (X_{ij}^{(t)} - \hat{\mu}_j^{(t)}\right) = 0$$

$$= \Pr\left(\sum_{i=1}^{q_t - 1} (a_i^{(t)} - \frac{n_{ij}^{(t)} \sum_{s=1}^{q_t - 1} a_s^{(t)}}{n}\right)\left(X_{ij}^{(t)} - \frac{n_{q_t j}^{(t)} \sum_{s=1}^{q_t - 1} a_s^{(t)}}{n} X_{q_t j}^{(t)} = 0\right) = 0$$

if the coefficients $a_i^{(t)} - \frac{n_{ij}^{(t)} \sum_{s=1}^{q_t - 1} a_s^{(t)}}{n}$, $i = 1, \cdots, q_t - 1$, and $\sum_{s=1}^{q_t - 1} a_s^{(t)}$ are not all zero. Note that the coefficients are all zero if and only if $a_1^{(t)} = \cdots = a_{q_t - 1}^{(t)} = 0$. Thus, for any $a_1^{(t)}, \cdots, a_{q_t - 1}^{(t)}$ not all zero, $\Pr\left(\sum_{i=1}^{q_t - 1} a_i^{(t)} G_{ij}^{(t)} = 0\right) = 0$. According to Lemma 1, we have that $G_{1j}^{(t)}, \cdots, G_{q_t - 1, j}^{(t)}$ are linearly irreducible for any t and j.

Between different resampling sets, e.g., the t_1-th set and t_2-th set, let $X = X_{1j}^{(t_1)}$. If $X \notin \mathcal{X}_{\cdot j}^{(t_2)}$, then $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1} - 1, j}^{(t_1)}$ are independent with $G_{1j}^{(t_2)}, \cdots, G_{q_{t_2} - 1, j}^{(t_2)}$. Since $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1} - 1, j}^{(t_1)}$ are linearly irreducible, then $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1} - 1, j}^{(t_1)}$ are linear irreducible given $G_{1j}^{(t_2)}, \cdots, G_{q_{t_2} - 1, j}^{(t_2)}$. Also, we have that $G_{1j}^{(t_2)}, \cdots, G_{q_{t_2} - 1, j}^{(t_2)}$ are linear irreducible given $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1} - 1, j}^{(t_1)}$. Thus, $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1} - 1, j}^{(t_1)}, G_{1j}^{(t_2)}, \cdots, G_{q_{t_2} - 1, j}^{(t_2)}$ are linearly irreducible.

If $X \in \mathcal{X}_{\cdot j}^{(t_2)}$, then $X \in \mathcal{X}_{\cdot j}^{(t_1)} \cap \mathcal{X}_{\cdot j}^{(t_2)}$. For any $a_1^{(t_1)}, \cdots, a_{q_{t_1} - 1}^{(t_1)}, a_1^{(t_2)}, \cdots, a_{q_{t_2} - 1}^{(t_2)} \in \mathbb{R}$, which are not all zeros, there are two cases. If $a_1^{(t_1)}, \cdots, a_{q_{t_1} - 1}^{(t_1)}$ are all zero (or $a_1^{(t_2)}, \cdots, a_{q_{t_2} - 1}^{(t_2)}$ are all zero), given $\mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X$, we have

$$\Pr\left(\sum_{i=1}^{q_{t_1} - 1} a_i^{(t_1)} G_{ij}^{(t_1)} + \sum_{i=1}^{q_{t_2} - 1} a_i^{(t_2)} G_{ij}^{(t_2)} = 0 | \mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X\right)$$

$$= \Pr\left(\sum_{i=1}^{q_{t_2} - 1} a_i^{(t_2)} G_{ij}^{(t_2)} = 0 | \mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X\right) = 0,$$

since $G_{1j}^{(t_2)}, \cdots, G_{q_{t_2} - 1, j}^{(t_2)}$ are linearly irreducible (or $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1} - 1, j}^{(t_1)}$ are linearly irreducible). If $a_1^{(t_1)}, \cdots, a_{q_{t_1} - 1}^{(t_1)}$ are not all zero and $a_1^{(t_2)}, \cdots, a_{q_{t_2} - 1}^{(t_2)}$ are not all zero, given $\mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X$, we have

$$\Pr\Big(\sum_{i=1}^{q_{t_1}-1} a_i^{(t_1)} G_{ij}^{(t_1)} + \sum_{i=1}^{q_{t_2}-1} a_i^{(t_2)} G_{ij}^{(t_2)} = 0 | \mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X\Big)$$
$$= \Pr\Big(\frac{A_1 X + B_1}{\sqrt{C_1 X^2 + D_1 X + E_1}} + \frac{A_2 X + B_2}{\sqrt{C_2 X^2 + D_2 X + E_2}} = 0 | \mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X\Big)$$

where A_1, B_1 are not both zero and A_2, B_2 are not both zero, since $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1}-1,j}^{(t_1)}$ are linearly irreducible and $G_{1j}^{(t_2)}, \cdots, G_{q_{t_2}-1,j}^{(t_2)}$ are linearly irreducible. Here $C_1 X^2 + D_1 X + E_1 > 0$ and $C_2 X^2 + D_2 X + E_2 > 0$ for all X, and $A_1, B_1, C_1, D_1, E_1, A_2, B_2, C_2, D_2, E_2$ are constants. Consider that

$$\frac{A_1 X + B_1}{\sqrt{C_1 X^2 + D_1 X + E_1}} + \frac{A_2 X + B_2}{\sqrt{C_2 X^2 + D_2 X + E_2}} = 0$$
$$\Rightarrow (A_1 X + B_1)^2 (C_2 X^2 + D_2 X + E_2) = (A_2 X + B_2)(C_1 X^2 + D_1 X + E_1),$$

there are at most 4 zero points in solution space $\boldsymbol{\Omega}$. Since $X = X_{1j}^{(t_1)}$ is independent of $\mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X$, and X is a continuous random variable, by Corollary 1, $X | X_{2j}^{(t)}, \cdots, X_{q_t,j}^{(t)}$ is a continuous random variable. For any finite set $\boldsymbol{\Omega}$, the probability $\Pr(X \in \boldsymbol{\Omega}) = 0$. So, we have

$$\Pr\left(\sum_{i=1}^{q_{t_1}-1} a_i^{(t_1)} G_{ij}^{(t_1)} + \sum_{i=1}^{q_{t_2}-1} a_i^{(t_2)} G_{ij}^{(t_2)} = 0 \middle| \mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X\right) = 0.$$

By integrating $\mathcal{X}_{\cdot j}^{(t_1)} \cup \mathcal{X}_{\cdot j}^{(t_2)} \backslash X_{ij}^{(t_1)}$ out, we have

$$\Pr\left(\sum_{i=1}^{q_{t_1}-1} a_i^{(t_1)} G_{ij}^{(t_1)} + \sum_{i=1}^{q_{t_2}-1} a_i^{(t_2)} G_{ij}^{(t_2)} = 0\right) = 0.$$

According to Lemma 2, $G_{ij}^{(t)}$ is a continuous random variable. Then, by Lemma 1, $G_{1j}^{(t_1)}, \cdots, G_{q_{t_1}-1,j}^{(t_1)}, G_{1j}^{(t_2)}, \cdots, G_{q_{t_2}-1,j}^{(t_2)}$ are linearly irreducible for all j. Similarly, we can generalize the results that $G_{1j}^{(1)}, \cdots, G_{q_1-1,j}^{(1)}, \cdots, G_{1j}^{(\tilde{T})}, \cdots, G_{q_{\tilde{T}}-1,j}^{(\tilde{T})}$ are linearly irreducible for all j. The random matrix $\mathbf{G}$ satisfies condition (2).

By Theorem 1, $\Pr(\text{rank}(\mathbf{G}) = \min(\sum_{t=1}^{\tilde{T}} (q_t - 1), p)) = 1$ as required. □

Proof of Theorem 3 Note that

$$\frac{1}{T}\sum_{t=1}^{T}\left(r_{ij}^{(t)} - \lambda_{ij}\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\left(r_{ij}^{(t)}\right)^2 - \frac{2\lambda_{ij}}{T}\sum_{t=1}^{T} r_{ij}^{(t)} + \lambda_{ij}^2.$$

Applying the Jensen's inequality to the first term,

$$\frac{1}{T}\sum_{t=1}^{T}\left(r_{ij}^{(t)}-\lambda_{ij}\right)^2 \geq \left(\frac{1}{T}\sum_{t=1}^{T}r_{ij}^{(t)}\right)^2 - \frac{2\lambda_{ij}}{T}\sum_{t=1}^{T}r_{ij}^{(t)}+\lambda_{ij}^2 = \left(\frac{1}{T}\sum_{t=1}^{T}r_{ij}^{(t)}-\lambda_{ij}\right)^2 = \left(r_{ij}^{Bag}-\lambda_{ij}\right)^2.$$

Integrating both sides of the inequality over the distribution of $\mathbf{X}$, by definition, we have

$$\mathrm{MSE}(r_{ij}^{Bag}) \leq \frac{1}{T}\sum_{t=1}^{T}\mathrm{MSE}(r_{ij}^{(t)}).$$

□

Theorem 4 *Note that*

$$\begin{aligned} r_{XY}^{(t)} &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(t)}-\bar{X}^{(t)})(Y_i^{(t)}-\bar{Y}^{(t)})}{S_X^{(t)}S_Y^{(t)}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\bar{X}^{(t)})(Y_i^{(t)}-\bar{Y}^{(t)})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\bar{X}^{(t)})^2\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(t)}-\bar{Y}^{(t)})^2}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\xi)(Y_i^{(t)}-\eta)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\bar{X}^{(t)})^2\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(t)}-\bar{Y}^{(t)})^2}} \\ &\quad - \frac{(\bar{X}^{(t)}-\xi)(\bar{Y}^{(t)}-\eta)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\bar{X}^{(t)})^2\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(t)}-\bar{Y}^{(t)})^2}}. \end{aligned} \tag{3}$$

To characterize the property of bootstrap sample mean and variance, we introduce $Z^{(t)} = (z_1, \cdots, z_n) \sim Multinomial(n; \frac{1}{n}, \cdots, \frac{1}{n})$ *to denote the number of occurrences for* $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ *in the t-th resampling set, where* $\sum_{i=1}^{n} z_i = n$. *For each* z_i, *we have* $E(z_i) = n \cdot \frac{1}{n} = 1$ *and* $\mathrm{Var}(z_i) = n \cdot \frac{1}{n}(1-\frac{1}{n}) = 1 - \frac{1}{n}$.

Then, the bootstrap sample mean $\bar{X}^{(t)}$ *can be written as*

$$\bar{X}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}X_i^{(t)} = \frac{1}{n}\sum_{i=1}^{n}z_iX_i.$$

Since $Z^{(t)}$ *is independent with* $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, *then we have the expectation*

$$E[\bar{X}^{(t)}] = E[\frac{1}{n}\sum_{i=1}^{n}z_iX_i] = \frac{1}{n}\sum_{i=1}^{n}E[z_i]\cdot E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\xi = \xi.$$

The variance can be calculated by the law of total variance as follows:

$$\begin{aligned}
&\mathrm{Var}[\bar{X}^{(t)}] = \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} z_i X_i] \\
&= \mathrm{Var}[E[\frac{1}{n}\sum_{i=1}^{n} z_i X_i | z_1, \cdots, z_n]] + E[\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} z_i X_i | z_1, \cdots, z_n]] \\
&= \mathrm{Var}[\frac{\xi}{n}\sum_{i=1}^{n} z_i] + E[\frac{\sigma^2}{n^2}\sum_{i=1}^{n} z_i^2] \\
&= \mathrm{Var}[\xi] + \frac{\sigma^2}{n}(2 - \frac{1}{n}) = \frac{\sigma^2}{n}(2 - \frac{1}{n}) \to 0 \quad \text{as } n \to \infty,
\end{aligned}$$

since $\sum_{i=1}^{n} z_i = n$ *and* $E[z_i^2] = \mathrm{Var}(z_i) + [E(z_i)]^2 = 2 - \frac{1}{n}$. *Thus, the bootstrap sample mean* $\bar{X}^{(t)}$ *is a consistent estimator of* ξ. *Similarly,* $\bar{Y}^{(t)}$ *is a consistent estimator of* η.

Since $\bar{X}^{(t)} \to \xi$ *in probability,* $\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}^{(t)})^2$ *can be written as*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}^{(t)})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)})^2 - (\bar{X}^{(t)})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i \cdot X_i^2 - (\bar{X}^{(t)})^2 \to \frac{1}{n}\sum_{i=1}^{n} z_i \cdot X_i^2 - \xi^2.$$

Then we have the expectation

$$\begin{aligned}
&E[\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}^{(t)})^2] \to E[\frac{1}{n}\sum_{i=1}^{n} z_i \cdot X_i^2] - \xi^2 = \frac{1}{n}\sum_{i=1}^{n} E[z_i] \cdot E[X_i^2] - \xi^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\xi^2 + \sigma^2) - \xi^2 = \sigma^2.
\end{aligned}$$

Since the forth moment is finite, the variance can be calculated by the law of total variance as follows:

$$\begin{aligned}
&\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}^{(t)})^2] \to \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} z_i \cdot X_i^2] \\
&= \mathrm{Var}[E[\frac{1}{n}\sum_{i=1}^{n} z_i X_i^2 | z_1, \cdots, z_n]] + E[\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} z_i X_i^2 | z_1, \cdots, z_n]] \\
&= \mathrm{Var}[\frac{\xi^2 + \sigma^2}{n}\sum_{i=1}^{n} z_i] + E[\frac{\mathrm{Var}(X_1^2)}{n^2}\sum_{i=1}^{n} z_i^2] \\
&= \mathrm{Var}[\xi^2 + \sigma^2] + \frac{\mathrm{Var}(X_1^2)}{n}(2 - \frac{1}{n}) = \frac{\mathrm{Var}(X_1^2)}{n}(2 - \frac{1}{n}) \to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Thus, $\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}^{(t)})^2$ *is a consistent estimator of* σ^2. *Similarly,* $\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(t)} - \bar{Y}^{(t)})^2$ *is a consistent estimator of* τ^2.

Back to Equation 3, since $\bar{X}^{(t)}$ *and* $\bar{Y}^{(t)}$ *are consistent estimators of* ξ *and* η, *and the denominators are consistent estimators of* σ^2 *and* τ^2, *then we have the second term in Equation 3 tends to* 0 *in probability as* $n \to \infty$, *i.e.*,

$$\frac{(\bar{X}^{(t)}-\xi)(\bar{Y}^{(t)}-\eta)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\bar{X}^{(t)})^2\cdot\frac{1}{n}\sum_{i=1}^{n}(Y_i^{(t)}-\bar{Y}^{(t)})^2}}\to 0. \tag{4}$$

For the first term in Equation 3, the denominators are consistent with σ^2 and τ^2. So, we focus on the numerator $\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\xi)(Y_i^{(t)}-\eta)$. By introducing Z, we have

$$\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\xi)(Y_i^{(t)}-\eta)=\frac{1}{n}\sum_{i=1}^{n}z_i(X_i-\xi)(Y_i-\eta).$$

Then we have the expectation

$$E(\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\xi)(Y_i^{(t)}-\eta))=E(\frac{1}{n}\sum_{i=1}^{n}z_i(X_i-\xi)(Y_i-\eta))$$
$$=\frac{1}{n}\sum_{i=1}^{n}E(z_i)E[(X_i-\xi)(Y_i-\eta)]=\frac{1}{n}\sum_{i=1}^{n}\sigma\tau\rho=\sigma\tau\rho.$$

The variance can be calculated by the law of total variance as follows:

$$\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\xi)(Y_i^{(t)}-\eta))=\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}z_i(X_i-\xi)(Y_i-\eta))$$
$$=\mathrm{Var}[E(\frac{1}{n}\sum_{i=1}^{n}z_i(X_i-\xi)(Y_i-\eta)|z_1,\cdots,z_n)]+E[\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}z_i(X_i-\xi)(Y_i-\eta)|z_1,\cdots,z_n)]$$
$$=\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n}z_i\cdot E((X_i-\xi)(Y_i-\eta))]+E[\frac{1}{n^2}\sum_{i=1}^{n}z_i^2\cdot\mathrm{Var}((X_i-\xi)(Y_i-\eta))]$$
$$=\mathrm{Var}[\sigma\tau\rho\cdot\frac{1}{n}\sum_{i=1}^{n}z_i]+\frac{1}{n}(2-\frac{1}{n})\mathrm{Var}((X_1-\xi)(Y_1-\eta)).$$
$$=\mathrm{Var}[\sigma\tau\rho]+\frac{1}{n}(2-\frac{1}{n})\mathrm{Var}((X_1-\xi)(Y_1-\eta))\to 0,\quad \text{as } n\to\infty.$$

Thus, we have the numerator $\frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)}-\xi)(Y_i^{(t)}-\eta)$ converges to $\sigma\tau\rho$ in probability. To sum up, the bootstrap correlation coefficient $r_{XY}^{(t)}$ is a consistent estimator of ρ. □

Corollary 3 *By Theorem 4, the bootstrap correlation coefficient $r_{ij}^{(t)}$ is a consistent estimator. Suppose that there exists some $\epsilon>0$ satisfying $\liminf_n E(r_{ij}^{n,(t)}-\lambda_{ij})^2\geq\epsilon>0$, where $r_{ij}^{n,(t)}$ denotes $r_{ij}^{(t)}$ with n samples. It follows that there exists N such that for $n>N$ we have $E(r_{ij}^{n,(t)}-\lambda_{ij})^2\geq\epsilon/2$.*

By Paley-Zygmund's inequality, we have

$$\Pr\left[(r_{ij}^{n,(t)}-\lambda_{ij})^2\geq\frac{\epsilon}{4}\right]\geq\Pr\left[(r_{ij}^{n,(t)}-\lambda_{ij})^2\geq\frac{E(r_{ij}^{n,(t)}-\lambda_{ij})^2}{2}\right]\geq\frac{(E(r_{ij}^{n,(t)}-\lambda_{ij})^2)^2}{4E(r_{ij}^{n,(t)}-\lambda_{ij})^4}\geq\frac{\epsilon^2/4}{4\times 16},$$

where we used $E(r_{ij}^{n,(t)}-\lambda_{ij})^4\leq 16$ since both $r_{ij}^{n,(t)}$ and λ_{ij} are bounded in $[1,-1]$. This is a contradiction with the fact that $r_{ij}^{n,(t)}$ is a consistent for λ_{ij}. Thus, we have

$$\mathrm{MSE}(r_{ij}^{(t)}) = E(r_{ij}^{(t)} - \lambda_{ij})^2 \to 0,$$

as $n \to \infty$. *According to Theorem* 3, *we have*

$$\mathrm{MSE}(r_{ij}^{Bag}) \le \mathrm{MSE}(r_{ij}^{(t)}) \to 0,$$

as $n \to \infty$. *By the definitions that* $\mathrm{MSE}(\mathbf{R}^{\mathrm{Bag}}) = \sum_{i,j} \mathrm{MSE}(r_{ij}^{Bag})$ *and* $\mathrm{MSE}(\mathbf{R}^{(t)}) = \sum_{i,j} \mathrm{MSE}(r_{ij}^{(t)})$, *we have*

$$\mathrm{MSE}(\mathbf{R}^{\mathrm{Bag}}) \le \mathrm{MSE}(\mathbf{R}^{(t)}) \to 0,$$

as $n \to \infty$ *for fixed* p. *It implies that the Bagging estimator* $\mathbf{R}^{\mathrm{Bag}}$ *is consistent.* □

Acknowledgements We thank anonymous reviewers for correcting the errors in our previous proof and encouraging us to find the current new approach for proving the consistency. Especially, a suggestion to use the Paley-Zygmund inequality made our proof more rigorous.

Author Contributions Conceptualization, CW and XF; methodology and formal analysis, CW; writing-original draft, CW; writing—review and editing, JD and XF; funding acquisition, XF and CW. All authors have read and agreed to the published version of the manuscript.

Funding The authors gratefully acknowledge two grants from the Research Grants Council of the Hong Kong SAR, China (CUHK14173817, CUHK14303819), one CUHK direct grant (4053357) and one UJS direct grant (5501190012).

Data availability Data on results is available upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability Code on results is available upon request.

References

Barnard, J., McCulloch, R., & Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica, 10*(4), 1281–1311.

Best, M. G., Sol, N., Kooi, I., Tannous, J., Westerman, B. A., Rustenburg, F., et al. (2015). RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell, 28*(5), 666–676.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling

reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences, 98*(24), 13790–13795.
Bickel, P. J., & Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics, 36*(6), 2577–2604.
Bickel, P. J., Levina, E., et al. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics, 36*(1), 199–227.
Bodnar, T., Parolya, N., & Schmid, W. (2018). Estimation of the global minimum variance portfolio in high dimensions. *European Journal of Operational Research, 266*(1), 371–390.
Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
Cai, T., & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association, 106*(494), 672–684.
Cai, T. T., & Zhou, H. H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. Statistica Sinica (pp. 1319–1349).
Cai, T. T., Zhang, C. H., Zhou, H. H., et al. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics, 38*(4), 2118–2144.
Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal, 19*(1), C1–C32.
Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432–441.
Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. New York: Springer.
Guillot, D., & Rajaratnam, B. (2012). Retaining positive definiteness in thresholded matrices. *Linear Algebra and its Applications, 436*(11), 4143–4160.
Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics, 29*(2), 295–327.
Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis, 88*(2), 365–411.
Lehmann, E. L. (1999). Elements of large-sample theory. Springer, 1999.
Marzetta, T. L., Tucci, G. H., & Simon, S. H. (2011). A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Transactions on Information Theory, 57*(9), 6256–6271.
Mestre, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Transactions on Information Theory, 54*(11), 5113–5129.
Mestre, X., & Lagunas, M. A. (2005). Diagonal loading for finite sample size beamforming: An asymptotic approach. Robust Adaptive Beamforming (pp. 200–266).
Neudecker, H., & Wesselman, A. M. (1990). The asymptotic variance matrix of the sample correlation matrix. *Linear Algebra and its Applications, 127,* 589–599.
Pishro-Nik, H. (2016). Introduction to probability, statistics, and random processes.
Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association, 104*(485), 177–186.
Tao, T., & Vu, V. (2010). Random matrices: Universality of local eigenvalue statistics up to the edge. *Communications in Mathematical Physics, 298*(2), 549–572.
Tucci, G. H., & Wang, K. (2019). New methods for handling singular sample covariance matrices. *IEEE Transactions on Information Theory, 65*(2), 770–786.
Wu, W. B., & Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica, 19*(4), 1755–1768.