Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Two Sample Test for Covariance Matrices in Ultra-High Dimension

Xiucai Ding, Yichen Hu & Zhenggang Wang

To cite this article: Xiucai Ding, Yichen Hu & Zhenggang Wang (03 Dec 2024): Two Sample Test for Covariance Matrices in Ultra-High Dimension, Journal of the American Statistical Association, DOI: 10.1080/01621459.2024.2423971

To link to this article: https://doi.org/10.1080/01621459.2024.2423971

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2024.2423971

Two Sample Test for Covariance Matrices in Ultra-High Dimension

Xiucai Ding, Yichen Hu, and Zhenggang Wang

Department of Statistics, UC Davis, Davis, CA

ABSTRACT

In this article, we propose a new test for testing the equality of two population covariance matrices in the ultra-high dimensional setting that the dimension is much larger than the sizes of both of the two samples. Our proposed methodology relies on a data splitting procedure and a comparison of a set of well selected eigenvalues of the sample covariance matrices on the split datasets. Compared to the existing methods, our methodology is adaptive in the sense that (i). it does not require specific assumption (e.g., comparable or balancing, etc.) on the sizes of two samples; (ii). it does not need quantitative or structural assumptions of the population covariance matrices; (iii). it does not need the parametric distributions or the detailed knowledge of the moments of the two populations. Theoretically, we establish the asymptotic distributions of the statistics used in our method and conduct the power analysis. We justify that our method is powerful under weak alternatives. We conduct extensive numerical simulations and show that our method significantly outperforms the existing ones both in terms of size and power. Analysis of two real datasets is also carried out to demonstrate the usefulness and superior performance of our proposed methodology. An R package `UHDtst` is developed for easy implementation of our proposed methodology. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received December 2023
Accepted October 2024

KEYWORDS
Covariance matrices; Random matrix theory; Two sample test; Ultra-high dimension

1. Introduction

Testing the equality of two population covariance matrices is of fundamental importance in statistical analysis. For two samples $\mathcal{X} := \{\mathbf{x}_i \in \mathbb{R}^p, 1 \le i \le n_1\}$ and $\mathcal{Y} := \{\mathbf{y}_j \in \mathbb{R}^p, 1 \le j \le n_2\}$ with population covariance matrices Σ_1 and Σ_2, respectively, researchers are interested in testing

$$\boldsymbol{H}_0 : \Sigma_1 = \Sigma_2. \tag{1.1}$$

Testing (1.1) is crucial in multivariate analysis and high dimensional statistics. First, the applications of many commonly used multivariate tests or techniques rely on whether the population covariance matrices are identical. For example, if two population covariance matrices are the same, the asymptotic distribution of Hotelling's T^2 statistic, which is used to test the equality of means, will be much easier to compute (Anderson 2003; Yao, Zheng, and Bai 2015). For another example, for classification, in order to correctly apply the tool of linear discriminant analysis (LDA), people need to check the equality of population covariance matrices (Anderson 2003). Second, in many gene expression data analysis, the two sample test for covariance matrices serves an important role in understanding, classifying and selecting gene associations across different phenotypes, for example, see Hu, Qiu, and Glazko (2010) and Dudoit, Fridlyand, and Speed (2002).

In this article, we will propose a novel method to test (1.1) in the ultra-high dimensional regime that

$$p \asymp n_1^{\alpha_1} \text{ and } p \asymp n_2^{\alpha_2}, \text{ for some constants } \alpha_1, \alpha_2 > 1. \tag{1.2}$$

We first summarize some related results and methodologies in Section 1.1 and then provide an overview of our approach in Section 1.2.

1.1. Some Related Results

In the literature of multivariate statistics, testing (1.1) has been well studied in the low dimensional regime when the dimension p is fixed and the sample sizes go to infinity. For example, see Anderson (2003), Sugiura and Nagao (1968), Nagao (1973), Manly and Rayner (1987), O'Brien (1992), Gupta and Tang (1984), and Perlman (1980). A common feature of these statistics is that they are based on likelihood ratios and use the eigenvalues of some sample covariance matrices.

On the other hand, the aforementioned methods may lose their validity as the dimension p diverges with the sample sizes. The main reason is that the likelihood ratio which is essentially a function of the eigenvalues of certain sample covariance matrices, is no more consistent due to the bias caused by the inconsistency of sample covariance matrices Bai et al. (2009). Motivated by this and based on results in random matrix theory, in the last decades, various modified or new statistical methodologies have been proposed in the high dimensional and comparable regime that $\alpha_1 = \alpha_2 = 1$ in (1.2). We now list but a few in the following. In Schott (2007), under the assumption that both samples are Gaussian, the author proposed a statistic based on the Frobenius norm of the difference of the sample covariance matrices of the two samples. In Bai et al. (2009), Zheng (2012),

CONTACT Xiucai Ding xcading@ucdavis.edu Department of Statistics, UC Davis, 4222 Mathematical Sciences Building, One Shields Avenue, Davis, CA 95616.
The majority of the work was carried out while ZGW was a postdoctoral researcher at UC Davis, with support from XCD.

Zheng, Bai, and Yao (2015), and Zou et al. (2021), assuming that $p < n_1$ or $p < n_2$ so that the precision matrices exist at least for one sample, the authors proposed several tests based on the eigenvalues of the F matrices. The condition $p < n_1$ or $p < n_2$ is weakened to some extent later in Zhang, Hu, and Bai (2017). We emphasize again that the aforementioned methods are all developed in the comparable regime that $\alpha_1 = \alpha_2 = 1$ in (1.2). In fact, they usually assume that $p/n_1 \to y_1$ and $p/n_2 \to y_2$ and both y_1 and y_2 will appear in the asymptotic distributions of their proposed statistics. Moreover, to correctly implement their methodologies, people normally either need to assume the samples are Gaussian or have prior knowledge of the moments of the entries of the random vectors. More recently, Zheng et al. (2019) proposed a power-enhancement test with no distributional or sparsity assumptions, but in the comparable regime, that is, $p/n_k \to c_k \in (0, \infty)$, $k = 1, 2$.

However, much less is touched in the ultra-high dimensional regime (1.2) except for a few ones under various additional assumptions. In Srivastava and Yanagihara (2010), under the assumption that both samples are Gaussian and $n_1 \asymp n_2$ (i.e., $\alpha_1 = \alpha_2$ in (1.2)), the authors proposed a test based some normalized traces of the sample covariance matrices. In Li and Chen (2012), assuming that $n_1 \asymp n_2$ (i.e., $\alpha_1 = \alpha_2$ in (1.2)), under certain regularity conditions on Σ_1 and Σ_2, the authors proposed a test based on some U-statistics which is an unbiased estimator of the Frobenius norm of $\Sigma_1 - \Sigma_2$. Later on, with additional assumption that both Σ_1 and Σ_2 are banded, He and Chen (2018) proposed another U-statistics based test but only targeting on the super-diagonal elements of the covariance matrices. Finally, in Cai, Liu, and Xia (2013), under the assumption that $n_1 \asymp n_2$ (i.e., $\alpha_1 = \alpha_2$ in (1.2)), some moment assumptions on the random vectors and certain sparsity assumptions on Σ_1 and Σ_2, the authors introduced a test based on the maximum standardized element-wise differences between the sample covariance matrices which can be computationally very expensive. In summary, all the existing methods concerning the ultra-high dimensional setting (1.2) require that the sample sizes are comparably large $n_1 \asymp n_2$ (i.e., $\alpha_1 = \alpha_2$ in (1.2)). Moreover, they need to impose some quantitative or structural assumptions on Σ_1 and Σ_2 and distributional assumptions on the random vectors.

Motivated by the above issues, in the current article, we propose a novel methodology to test (1.1) in the ultra-high dimensional setting (1.2). Our approach does not need assumptions on the sample sizes n_1, n_2, or quantitative or structural assumptions on the population covariance matrices. Moreover, we do not require the random vectors to have specific distributions like Gaussian. An overview of our method will be given in Section 1.2.

1.2. An Overview of Our Method

In contrast to the methods developed in Srivastava and Yanagihara (2010), Li and Chen (2012), He and Chen (2018), and Cai, Liu, and Xia (2013), which all directly compare the entries of the sample covariance matrices, our proposed approach uses the eigenvalues of the sample covariance matrices. However, if we directly compare all the eigenvalues, it can result in several issues. First, since the values of the sample sizes n_1 and n_2 are different in general, a direct comparison can lead to bias especially when their orders are different. For example, if $n_1 \gg n_2$, in our regime (1.2), the sample $\mathcal{Y}$ will have much fewer nonzero eigenvalues to be considered. Second, and most importantly, as has been demonstrated in Bai et al. (2009), Ding and Wang (2023), Yao, Zheng, and Bai (2015), and Zheng, Bai, and Yao (2015), the distribution of the statistics that use all the eigenvalues usually involves more unknown quantities like the first four moments of the random samples and the detailed information of Σ_1 and Σ_2.

To address the above issues, inspired by the recent developments in random matrix theory (Li, Schnelli, and Xu 2021; Ding and Wang 2023), we only compare a subset of the eigenvalues of the two sample covariance matrices. This resolves the issue of using too many eigenvalues of one sample covariance matrix. Moreover, as will be seen in Corollary 3.1, under the null hypothesis, the asymptotic distribution of the statistic is very universal in the sense that it does not require the knowledge of any particular information of the population covariance matrices and the moments of the random vectors. In fact, as can be seen in Theorem 3.1, regardless of whether (1.1) holds, only the mean parts encode the information of the population covariance matrices. This further makes it easier to study the power of the statistics which shows that our method can reject the null hypothesis under very weak alternative.

Our proposed methodology (see Algorithm 2.2) will be presented in Section 2. It consists of three important components. The first one is a data splitting procedure (see Algorithm 2.1) which divides the data in $\mathcal{X} \cup \mathcal{Y}$ into three parts, denoted as $\mathcal{X}^s, \mathcal{Y}^s$ and $\mathcal{Z}^s$ with the same size n satisfying (2.1). $\mathcal{X}^s$ and $\mathcal{Y}^s$ are the testing beds and $\mathcal{Z}^s$ is used to generate some useful quantities for us to choose a subset of the eigenvalues of the sample covariance matrices associated with $\mathcal{X}^s$ and $\mathcal{Y}^s$. The selection of the subset of the eigenvalues is done via the choice of a location parameter γ (see (2.2)) and a tuning bandwidth η_0 (see Algorithm B.1 of our supplement). Both parameters can be chosen automatically. Our statistic in (2.4) primarily considers the eigenvalues lying within the interval $[\gamma - 1.05\eta_0, \gamma + 1.05\eta_0]$. Here we choose $[\gamma - 1.05\eta_0, \gamma + 1.05\eta_0]$ instead of $[\gamma - \eta_0, \gamma + \eta_0]$ mainly for technical reasons. However, the major contribution comes from those eigenvalues lying in $[\gamma - \eta_0, \gamma + \eta_0]$. The construction of the above statistic only uses one split dataset so that some samples may be omitted. In order to use as much information as possible and stabilize our procedure, our second component of the methodology is to repeat the splitting procedure multiple times. Instead of using the statistic in (2.4) once, we generate a sequence of such statistics and construct a summary statistic called *decision ratio* (see (2.10)). Such a procedure will reduce the variability of testing (1.1) compared with only one data splitting. The last component of our methodology is to provide a critical value δ for the decision ratio to suggest whether we should accept or reject the null hypothesis. This will be done by a calibration procedure (see Algorithm B.2 of our supplement) which uses the very universal properties of our statistics under the null hypothesis (1.1).

On the theoretical side, we establish the asymptotic distributions for our statistics and decision ratio in Section 3.2. Moreover, we conduct detailed power analysis for our method-

ology in Section 3.3 which shows that our proposed method will be powerful under weak alternatives. We test our proposed methodology and compare it with the state-of-the-art methods (Srivastava and Yanagihara 2010; Li and Chen 2012; He and Chen 2018; Cai, Liu, and Xia 2013) on both simulated and two real datasets. The numerical results show that our proposed method outperforms the existing ones.

1.3. Organization of the Article

The rest of the article is organized as follows. In Section 2, we introduce our test procedure. Section 3 provides theoretical guarantees for our procedure, establishing the asymptotic distributions of the test statistics and conducting power analysis. In Section 4, we compare our proposed methodology with several existing methods via Monte Carlo simulations and two real data analysis. An online supplementary file is enclosed to provide the technical proofs in Section A, the arguments of tuning parameter selection in Section B, additional simulations in Section C, and additional discussions in Section D. An R package `UHDtst` is developed for easy implementation.

2. Methodology

In this section, we introduce our proposed methodology.

2.1. Construction of Test Statistics

Our first step is the data splitting procedure via random sampling. For some integer n satisfying that

$$n < \mathsf{N}, \text{ where } \mathsf{N} := \min\left\{\frac{\max\{n_1, n_2\}}{2}, n_1, n_2\right\}, \tag{2.1}$$

we follow Algorithm 2.1 to split the data.

Algorithm 2.1 Data splitting

Inputs: n, the datasets $\mathcal{X}$ and $\mathcal{Y}$.
Step one: Randomly sample n data points from $\mathcal{X}$ and $\mathcal{Y}$, denoted as $\mathcal{X}^s$ and $\mathcal{Y}^s$, respectively.
Step two: For the dataset with more samples, say $\mathcal{X}$ (i.e., $n_1 \geq n_2$), we randomly sample n data points from $\mathcal{X}\backslash\mathcal{X}^s$, denoted as $\mathcal{Z}^s$.
Output: The split datasets $\mathcal{X}^s$, $\mathcal{Y}^s$, and $\mathcal{Z}^s$.

Algorithm 2.1 generates three independent datasets ($\mathcal{X}^s$, $\mathcal{Y}^s$, and $\mathcal{Z}^s$) with the same sample size n. $\mathcal{X}^s$, and $\mathcal{Y}^s$ are used for testing (1.1), while $\mathcal{Z}^s$ serves as the reference dataset to generate γ, as discussed in Section 1.2. Let the sample covariance matrices associated with $\mathcal{X}^s$, $\mathcal{Y}^s$, and $\mathcal{Z}^s$ be

$$\mathcal{Q}_x = \frac{1}{\sqrt{pn}}\sum_{\mathbf{x}_i\in\mathcal{X}^s}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^\top,$$
$$\mathcal{Q}_y = \frac{1}{\sqrt{pn}}\sum_{\mathbf{y}_i\in\mathcal{Y}^s}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})^\top,$$
$$\mathcal{Q}_z = \frac{1}{\sqrt{pn}}\sum_{\mathbf{z}_i\in\mathcal{Z}^s}(\mathbf{z}_i-\bar{\mathbf{z}})(\mathbf{z}_i-\bar{\mathbf{z}})^\top,$$

where $\bar{\mathbf{x}}, \bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$ are their respective sample means. Note that the scaling $(pn)^{-1/2}$ differs from the typical $\frac{1}{n}$. As shown in Section 3, it addresses the ultra-high dimensionality (1.2).

We denote the nonzero eigenvalue sequences of the sample covariance matrices associated with $\mathcal{X}^s$, $\mathcal{Y}^s$, and $\mathcal{Z}^s$ as $\{\lambda_j\}_{j=1}^n$, $\{\mu_j\}_{j=1}^n$, and $\{\gamma_j\}_{j=1}^n$, respectively, assuming they are in decreasing order. Let

$$\gamma := \text{Median}\{\gamma_j\}. \tag{2.2}$$

Given the mollifier function

$$\mathcal{K}(x) := \begin{cases} 0 & |x| \geq 1.05 \\ 1 & |x| \leq 1 \\ \exp\left(\frac{1}{(0.05)^2} - \frac{1}{(0.05)^2-(|x|-1)^2}\right) & 1 < |x| < 1.05 \end{cases} \tag{2.3}$$

and $\eta_0 \equiv \eta_0(n) \ll 1$ chosen from Algorithm B.1 of our supplement, we will use the statistic

$$\mathbb{T} := \mathbb{T}_x - \mathbb{T}_y, \tag{2.4}$$

where

$$\mathbb{T}_x := \sum_{j=1}^n \left(\frac{\lambda_j-\gamma}{\eta_0}\right)\mathcal{K}\left(\frac{\lambda_j-\gamma}{\eta_0}\right),$$
$$\mathbb{T}_y := \sum_{j=1}^n \left(\frac{\mu_j-\gamma}{\eta_0}\right)\mathcal{K}\left(\frac{\mu_j-\gamma}{\eta_0}\right). \tag{2.5}$$

We provide a few remarks. First, $\mathcal{K}(x)$ is a smooth version of the indicator function $\mathcal{I}(x) = \mathbf{1}_{|x|\leq 1}$, mainly for technical reasons related to the Helffer-Sjöstrand formula; see the discussion around (A.23) of our supplement. Moreover, similar results also hold for general test functions other than (2.5), see Remark A.2 for details. Second, as shown in Figure 1, according to the definition in (2.3), instead of using all the eigenvalues, the statistics in (2.5) mostly sum up the properly scaled and shifted

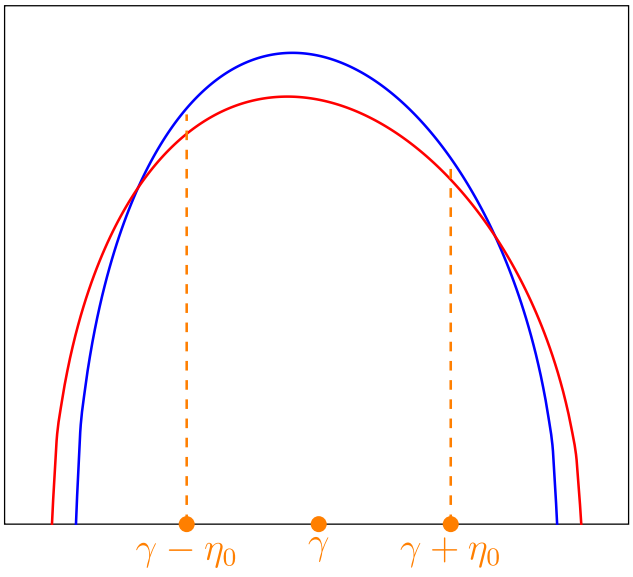

Figure 1. Illustration of the statistics $\mathbb{T}_x$ and $\mathbb{T}_y$. Our test statistics $\mathbb{T}_x$ and $\mathbb{T}_y$ mostly focus on the eigenvalues within the interval $[\gamma-\eta_0, \gamma+\eta_0]$ and therefore are affected by difference between the spectral densities (blue: ϱ_1, red: ϱ_2) within this interval.

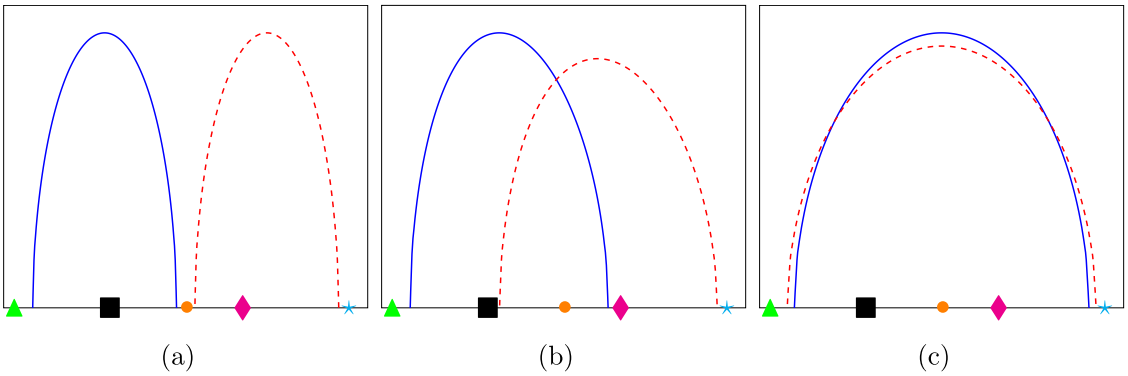

Figure 2. Possible examples of the LSDs of split datasets $(\mathcal{X}^s, \mathcal{Y}^s)$ and locations of γ. This figure illustrates how the densities separate and emphasizes that for the test to be effective, γ computed from $\mathcal{Z}^s$ should lie in the overlapping region of the supports of both densities.

eigenvalues within the small neighborhood $[\gamma - \eta_0, \gamma + \eta_0]$. For the choice of γ, we refer to Remark A.4 of the supplement for detailed discussions. Third, since the statistics in (2.4) use only one split dataset, to use all available information and stabilize the procedure, we can repeat the data splitting Algorithm 2.1 and construct the statistics (2.4) multiple times, as described in the next subsection.

Denote

$$\mathsf{v} = \frac{1}{2\pi^2} \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{(\mathcal{K}(x_1) - \mathcal{K}(x_2))^2}{(x_1 - x_2)^2} \, \mathrm{d}x_1 \mathrm{d}x_2. \tag{2.6}$$

We will see later from Section 3.2 that under the null hypothesis (1.1), $\mathbb{T}$ will be asymptotically $\mathcal{N}(0, 2\mathsf{v})$. Therefore, under the nominal level α, we should reject the null hypothesis if $|\mathbb{T}| > \sqrt{2\mathsf{v}} z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)\%$ quantile of the $\mathcal{N}(0, 1)$ random variable.

Remark 2.1. Note that the sample covariance matrices $\mathcal{Q}_x$ and $\mathcal{Q}_y$ have at most rank n and therefore n nonzero eigenvalues. Even though the full data dimension is p, our test statistics $\mathbb{T}_x$ and $\mathbb{T}_y$ defined in (2.5) only sum over the n nonzero eigenvalues, as the additional $p - n$ zero eigenvalues do not contribute to the randomness of the sum.

2.2. Test Procedure

To fully use the data and stabilize our testing procedure, we repeat the data splitting (Algorithm 2.1) and test statistic construction (2.4) multiple times. We generate a sequence of test statistics and compute a summary statistic called the *decision ratio* (2.10), which provides a more robust assessment of the null hypothesis (1.1), see Remark A.3 of the supplement for the theoretical motivation behind Algorithms 2.1 and 2.2.

Before stating our inferential procedure, we define the *efficient data splitting* scheme.

Definition 2.1. We call $\mathcal{X}^s, \mathcal{Y}^s, \mathcal{Z}^s$ from Algorithm 2.1 is an ϵ-efficient data splitting if

$$\max\{|\gamma - \mu_1|, |\gamma - \mu_n|\} \le \mathrm{Range}(\mathcal{Y}^s) - \epsilon \text{ and}$$
$$\max\{|\gamma - \lambda_1|, |\gamma - \lambda_n|\} \le \mathrm{Range}(\mathcal{X}^s) - \epsilon,$$

where $\mathrm{Range}(\cdot)$ is the range of the eigenvalues of the associated sample covariance matrices.

Remark 2.2. Definition 2.1 essentially states that the eigenvalue ranges of $\mathcal{Q}_x$ and $\mathcal{Q}_y$ must overlap, and γ computed from $\mathcal{Q}_z$ should lie in the overlapped interval; otherwise, we should quickly reject the null hypothesis, boosting the power of our test. Figure 2 illustrates this with three subfigures showing various cases of split datasets. The blue and red curves represent the possible limiting distributions of the eigenvalues of $\mathcal{X}^s$ and $\mathcal{Y}^s$, while ▲ (leftmost), ■, ●, ♦, and ⋆ (rightmost) represent five possible locations of γ derived from $\mathcal{Z}^s$.

First, in Figure 2(a), regardless of the locations of γ, it does not satisfy Definition 2.1. In this setting, we should reject the null hypothesis in (1.1) without testing, as concluded from Lemma A.2. However, if γ is in the orange bullet ● position and we use our statistic (2.4), we may fail to reject (1.1) since a small neighborhood of γ contains no eigenvalues of $\mathcal{Q}_x$ and $\mathcal{Q}_y$. To address this and boost power, Algorithm 2.2 (see (2.7)) directly rejects $\boldsymbol{H}_0$ if a data splitting like Figure 2(a) happens. Second, the data splitting is efficient if γ is in the ● spot in Figure 2(b) or ■, ●, or ♦ spot in Figure 2(c). To boost power, Algorithm 2.2 only considers efficient splitting for cases in Figure 2(b) and (c) (see (2.8)).

We now propose our two sample test procedure in Algorithm 2.2. The algorithm can be implemented automatically using our R package `UHDtst`.

Remark 2.3. Several remarks are in order on Algorithm 2.2. First, as discussed in Remark 2.2, Steps one and two are mainly employed to increase the power under the alternative. In fact, under the null hypothesis when (1.1) holds, as can be seen from the proof of Corollary 3.2, with high probability, (3.7) holds. In other words, (2.7) and (2.8) will be skipped and all the split datasets will be used. Second, as will be seen in Corollary 3.2, when the null hypothesis (1.1) holds, conditional on the datasets, $\{\mathsf{c}_i\}$ in (2.9) can be asymptotically regarded as a sequence of iid Bernoulli random variables with probability $\mathsf{p} = \alpha$. Consequently, when n is sufficiently large, asymptotically, it suffices to check $\boldsymbol{H}_0 : \mathsf{p} = \alpha$ Vs $\boldsymbol{H}_a : \mathsf{p} > \alpha$ which can be done using the Binomial test or its Gaussian approximation when K is large. That is, under the nominal level α, we need to reject the null hypothesis if

$$\mathsf{DR} > \frac{1}{\mathsf{K}} \mathcal{B}_{\mathsf{K},\alpha}(1-\alpha), \text{ or } \mathsf{DR} > \alpha + z_{1-\alpha/2} \frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{\mathsf{K}}}, \tag{2.11}$$

Algorithm 2.2 Two sample test procedure

Inputs: n, Type I error α, and the datasets $\mathcal{X}$ and $\mathcal{Y}$.
Step one: Run Algorithm 2.1 K times (say $\mathsf{K} = 1000$) and record the split datasets as $(\mathcal{X}_i^s, \mathcal{Y}_i^s, \mathcal{Z}_i^s)$, whose associated eigenvalues as $(\{\lambda_j^i\}, \{\mu_j^i\}, \{\gamma_j^i\})$, $1 \le j \le n, 1 \le i \le \mathsf{K}$. For $1 \le i \le \mathsf{K}$, let the ranges of $\{\lambda_j^i\}, \{\mu_j^i\}$ be as $\mathsf{R}_i(x), \mathsf{R}_i(y)$, respectively. For $1 \le i \le \mathsf{K}$ and a given small positive value $\tilde{\epsilon}$ (say $\tilde{\epsilon} = 0.05$), if

$$\max\{|\lambda_1^i - \mu_n^i|, |\mu_1^i - \lambda_n^i|\} > \mathsf{R}_i(x) + \mathsf{R}_i(y) + \tilde{\epsilon}, \tag{2.7}$$

we record that $\mathsf{c}_i = 1$ and denote $\mathcal{S}_0 := \{1 \le i \le \mathsf{K} |$ (2.7) is satisfied$\}$.
Step two: Compute the median values for $\{\gamma_j^i\}$ as in (2.2) and denote them as $\{\gamma^i\}$. For some small value ϵ (say $\epsilon = 0.05$), denote the set

$$\mathcal{S}_1 := \{\{1, 2, \ldots, \mathsf{K}\} \backslash \mathcal{S}_0 | \text{Definition 2.1 is satisfied}\}. \tag{2.8}$$

If $\mathcal{S}_0 \cup \mathcal{S}_1 = \emptyset$, redo Steps one and two until $\mathcal{S}_0 \cup \mathcal{S}_1 \neq \emptyset$.
Step three: For $i \in \mathcal{S}_1$, together with $(\{\lambda_j^i\}, \{\mu_j^i\})$, run Algorithm B.1 from our supplement to choose a sequence of tuning parameters $\{\eta_0^i\}$. Using the above quantities, construct a sequence of statistics $\mathbb{T}_i$ following (2.4) and (2.5).
Step four: For $i \in \mathcal{S}_1$ and the given type one error α, we record

$$\mathsf{c}_i = \mathbf{1}(|\mathbb{T}_i| \ge z_{1-\alpha/2}\sqrt{2\mathsf{v}}), \tag{2.9}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\%$ quantile of the $\mathcal{N}(0,1)$ random variable.
Step five: Calculate the decision ratio (DR)

$$\mathsf{DR} = \frac{1}{|\mathcal{S}_0 \cup \mathcal{S}_1|} \sum_{i=1}^{|\mathcal{S}_0 \cup \mathcal{S}_1|} \mathsf{c}_i. \tag{2.10}$$

Output: Reject the null hypothesis (1.1) if $\mathsf{DR} > \delta$. Here $\delta > \alpha$ is some control threshold which can be tuned using Algorithm B.2 from our supplement.

where $\mathcal{B}_{\mathsf{K},\alpha}(1-\alpha)$ is the $(1-\alpha)\%$ quantile of a binomial distribution with parameters K and α; see Corollary 3.2 for more discussions.

Remark 2.4. To implement Algorithm 2.2, several parameters need to be chosen. The first one is the split sample size n, which should satisfy (2.1). Generally, increasing n leads to more informative statistics. However, it is important to strike a balance between choosing a reasonably large value for n and ensuring that N and n are separated to allow for multiple splitting procedures. Section C.2 of the supplement provides a detailed comparison of different choices of $n \equiv n(\mathsf{N})$, suggesting that when N is large enough, smaller n values can be used to reduce computational burden while maintaining robust performance. In our R package `UHDtst`, for N satisfying (2.1), we set the default value of $n = \mathsf{N} - 5$, which achieves a balance for both larger and smaller N values.

The second parameter is the window size in Step three (i.e., η_0 in (2.5)). η_0 can be regarded as a bandwidth that controls the number of eigenvalues used in the test. Inspired by this, we provide a smoothing-based approach to choose η_0 in Algorithm B.1 of our supplement. More details can be found in Section B of the supplement.

The third parameter is the ϵ-efficiency parameter. Current choice of ϵ appears to be robust and achieve a constantly high ratio of ϵ-efficiency under various settings; we refer to Section C.1 for a comprehensive discussion and supporting evidence.

The last parameter is the threshold δ. As discussed in Remark 2.3, according to (2.11), our theory has provided some theoretically justified values for δ when n is sufficiently large. For finite n, in order to improve the accuracy and power, in Algorithm B.2 of our supplement, we provide a calibration procedure to choose δ. The motivation is inspired by the results in Corollary 3.1 that when the null hypothesis (1.1) holds, the distribution of (2.4) only relies on (2.6) which is irrelevant of the matrix Σ that $\Sigma_1 = \Sigma_2 = \Sigma$. Therefore, we can calibrate a δ for any combinations of $(n_1, n_2, p, \mathsf{K})$ simply using iid multivariate Gaussian samples for both $\mathcal{X}$ and $\mathcal{Y}$; see Section B of our supplement for more details.

3. Theoretical Guarantees

This section analyzes the proposed Algorithm 2.2, focusing on the asymptotic distributions of our proposed statistics (2.4) and (2.5), as well as the power of statistic (2.4). For a $k \times k$ symmetric matrix H, its empirical spectral distribution (ESD) is defined as $\mu_H := \frac{1}{k}\sum_{i=1}^k \delta_{\lambda_i(H)}$, where δ is the Dirac's delta function and $\{\lambda_i(H)\}$ are the eigenvalues of H. For any probability measure ν defined on $\mathbb{R}$, its Stieltjes transform is defined as

$$m_\nu(z) = \int \frac{1}{x - z} \mathrm{d}\nu(x), \tag{3.1}$$

where $z \in \mathbb{C}_+ := \{E + \mathrm{i}\eta : E \in \mathbb{R}, \eta > 0\}$. For two sequences of deterministic positive values a_n and b_n, we write $a_n = \mathrm{O}(b_n)$ if $a_n \le Cb_n$ for some constant $C > 0$, and $a_n \asymp b_n$ if both $a_n = \mathrm{O}(b_n)$ and $b_n = \mathrm{O}(a_n)$. Moreover, we write $a_n = \mathrm{o}(b_n)$ if $a_n \le c_n b_n$ for some positive sequence $c_n \downarrow 0$. Moreover, for a sequence of random variables $\{x_n\}$ and positive real values $\{a_n\}$, we use $x_n = \mathrm{O}_{\mathbb{P}}(a_n)$ to state that x_n/a_n is stochastically bounded. Similarly, we use $x_n = \mathrm{o}_{\mathbb{P}}(a_n)$ to say that x_n/a_n converges to zero in probability.

3.1. Some Background in Random Matrix Theory

In this section, we provide some background and preliminary results. Throughout this section, we will need the following mild assumptions.

Assumption 3.1. We assume that the following conditions are satisfied:

1. For dimensionality, we assume (1.2) holds.
2. We assume that the two samples are iid generated according to $\boldsymbol{x}_i = \Sigma_1^{1/2}\boldsymbol{x}_i, 1 \le i \le n_1$, and $\boldsymbol{y}_j = \Sigma_2^{1/2}\boldsymbol{y}_j \in \mathbb{R}^p, 1 \le j \le n_2$, where the entries of $\boldsymbol{x}_i = (x_{is})$ and $\boldsymbol{y}_j = (y_{jt}), 1 \le i \le n_1,\ 1 \le j \le n_2$, are independent and satisfy that for some positive sequence $(C_k)_{k\in\mathbb{N}}$ that for all $k \in \mathbb{N}$
$$\mathbb{E}x_{is} = 0, \quad \mathbb{E}x_{is}^2 = 1, \quad \mathbb{E}\,|x_{is}|^k \le C_k, \tag{3.2}$$
$$\mathbb{E}y_{jt} = 0, \quad \mathbb{E}y_{jt}^2 = 1, \quad \mathbb{E}\,|y_{jt}|^k \le C_k. \tag{3.3}$$
3. For Σ_1 and Σ_2, we assume that all of their eigenvalues are bounded from above and below away from zero.

Remark 3.1. Several remarks on Assumption 3.1 are in order. First, condition 1 specifies the ultra-high dimensional regime. Second, condition 2 provides a commonly used data generating model in high-dimensional data analysis (e.g., Chen, Zhang, and Zhong 2010; He and Chen 2018; Dobriban and Owen 2019; Ke, Ma, and Lin 2023). Moreover, we assume centered data for ease of statements, but our results can be easily generalized to the nonzero mean setting (see Theorem 2.23 of Bloemendal et al. (2016)). The moment assumptions in (3.2) and (3.3) can be weakened with additional technical efforts (see Ding and Yang 2018; Yang 2019). Finally, condition 3 imposes a mild assumption on the population covariance matrices.

For the split datasets $\mathcal{X}^s$ and $\mathcal{Y}^s$ from our Algorithm 2.1, under Assumption 3.1, we can write the sample covariance matrix as follows
$$\mathcal{Q}_x = \frac{1}{\sqrt{pn}}\Sigma_1^{1/2}XX^\top\Sigma_1^{1/2},\ \mathcal{Q}_y = \frac{1}{\sqrt{pn}}\Sigma_2^{1/2}YY^\top\Sigma_2^{1/2}, \tag{3.4}$$
where X contains the samples from $\mathcal{X}^s$ and Y contains that from $\mathcal{Y}^s$. Since the matrices
$$Q_x = \frac{1}{\sqrt{pn}}X^\top\Sigma_1 X,\ Q_y = \frac{1}{\sqrt{pn}}Y^\top\Sigma_2 Y,$$
have the same nonzero eigenvalues with $\mathcal{Q}_x$ and $\mathcal{Q}_y$, it is sufficient to work with Q_x and Q_y. It is well-known that the limiting spectral distributions (LSD) of Q_x and Q_y can be best described by their Stieltjes transforms, denoted as $m_1(z)$ and $m_2(z)$. The following lemma characterizes $m_1(z)$ and $m_2(z)$.

Lemma 3.1. Suppose $\{\sigma_j^{(i)}\}_{j=1}^p$ is the sequence of the eigenvalues of Σ_i and let $\phi := \frac{p}{n}$. Then for each $z \in \mathbb{C}_+$, there exists a unique $m_i \equiv m_i(z) \in \mathbb{C}_+,\ i = 1, 2$, satisfying
$$\frac{1}{m_i} = -z + \frac{1}{p}\sum_{j=1}^p \frac{\phi}{\phi^{1/2}(\sigma_j^{(i)})^{-1} + m_i}.$$

Proof. See Lemma 2.3 of Ding and Wang (2023). □

It is well-known that given $m_i(z)$, people can obtain its associated density function in the sense of (3.1) using the inverse formula (Bai and Silverstein 2010) (also see (A.8)). Let ϱ_i be the asymptotic density associated with m_i in Lemma 3.1, $i = 1, 2$. In (Ding and Wang 2023, Lemma 2.5), it has been also proved that $\varrho_i, i = 1, 2$, are both supported on some single intervals that for some constants $\gamma_\pm^i$
$$\operatorname{supp}\varrho_i \cap (0,\infty) = [\gamma_-^i, \gamma_+^i], \tag{3.5}$$
where $\gamma_+^i - \gamma_-^i = \mathrm{O}(1)$ and $\gamma_-^i, \gamma_+^i \asymp (p/n)^{1/2}$ for $i = 1, 2$.

3.2. Asymptotic Distributions of the Statistics

In this section, we establish the CLTs for the statistics (2.4) an (2.5). For γ in (2.2), we define
$$\mathsf{M}_x := n\int_{\mathbb{R}} \frac{t-\gamma}{\eta_0}\mathcal{K}\left(\frac{t-\gamma}{\eta_0}\right)\mathrm{d}\varrho_1(t),$$
$$\mathsf{M}_y := n\int_{\mathbb{R}} \frac{t-\gamma}{\eta_0}\mathcal{K}\left(\frac{t-\gamma}{\eta_0}\right)\mathrm{d}\varrho_2(t). \tag{3.6}$$

Theorem 3.1. Suppose Assumption 3.1 holds and $\max\{\gamma_-^1, \gamma_-^2\} < \gamma < \min\{\gamma_+^1, \gamma_+^2\}$. For some small constants $\tau_1, \tau_2 > 0$ that $n^{-1+\tau_2} < \eta_0 \le n^{-\tau_1}$, we have that for $\mathbb{T}_x$ and $\mathbb{T}_y$ in (2.5) and v in (2.6)
$$\mathbb{T}_x - \mathsf{M}_x \Rightarrow \mathcal{N}(0, \mathsf{v}),\ \mathbb{T}_y - \mathsf{M}_y \Rightarrow \mathcal{N}(0, \mathsf{v}).$$

Proof. See Section A.3. □

Remark 3.2. Theorem 3.1 establishes the asymptotic normality for the statistics $\mathbb{T}_x$ and $\mathbb{T}_y$ with asymptotically identical variances, regardless of whether $\boldsymbol{H}_0$ in (1.1) holds. The differences lie in the mean parts M_x and M_y in (3.6). Under (1.1), $\varrho_1 \equiv \varrho_2$, we have that $\mathsf{M}_x = \mathsf{M}_y$ and $\mathbb{T}$ has zero mean. When (1.1) fails, $\varrho_i, i = 1, 2$, which encode the information of Σ_i via Lemma 3.1, will be different, causing $\mathsf{M}_x \ne \mathsf{M}_y$. Regarding the typical (n, p) sizes needed for the asymptotics to work well in practice, we have included more detailed discussions in Section C.5 of the supplement.

For the distribution of $\mathbb{T}$ under the null hypothesis (1.1), due to the independent splitting in Algorithm 2.1, Theorem 3.1 immediately yields the following result.

Corollary 3.1. Suppose Assumption 3.1 holds. Then under the null hypothesis $\boldsymbol{H}_0$ in (1.1), for some small constants $\tau_1, \tau_2 > 0$ that $n^{-1+\tau_2} < \eta_0 \le n^{-\tau_1}$, we have
$$\mathbb{T} \Rightarrow \mathcal{N}(0, 2\mathsf{v}).$$

Proof. Under these assumptions, it is clear from Lemma 3.1 and (3.5) that $\varrho_1 = \varrho_2 \equiv \varrho, \gamma_-^1 = \gamma_-^2 \equiv \gamma_-$ and $\gamma_+^1 = \gamma_+^2 \equiv \gamma_+$. Moreover, by Lemma A.2 and the discussion in Remark A.1, we see that with high probability, $\gamma_- < \gamma < \gamma_+$. These also imply that $\mathsf{M}_x = \mathsf{M}_y$. The proof then follows from Theorem 3.1, the independence splitting in Algorithm 2.1 and $\mathbb{T} = (\mathbb{T}_x - \mathsf{M}_x) - (\mathbb{T}_y - \mathsf{M}_y)$. □

As a consequence of Corollary 3.1, we immediately obtain the asymptotic properties of our Algorithm 2.2 in terms of the decision ratio (DR) in (2.10).

Corollary 3.2. Suppose the assumptions of Corollary 3.1 hold. Then when n is sufficiently large, for DR in (2.10) of our Algorithm 2.2, when the null hypothesis $\boldsymbol{H}_0$ in (1.1) holds, we have that conditional on the datasets $(\mathcal{X}, \mathcal{Y})$

$$\mathsf{K} \times \mathsf{DR} \Rightarrow \mathcal{B}_{\mathsf{K},\alpha},$$

where $\mathcal{B}_{\mathsf{K},\alpha}$ is a Binomial random variable with size K and probability α.

Proof. Analogous to the proof of Corollary 3.1, by Lemma A.2 and the discussion in Remark A.1, we see that with high probability, for all $1 \le i \le \mathsf{K}$,

$$\lambda_1^i = \gamma_+ + o(1),\ \mu_1^i = \gamma_+ + o(1),\ \lambda_n^i = \gamma_- + o(1),\ \mu_n^i = \gamma_- + o(1),$$

and

$$\gamma_- < \gamma^i < \gamma_+.$$

In view of (2.7) and Definition 2.1, we conclude that with high probability

$$\mathcal{S}_0 = \emptyset,\ \mathcal{S}_1 = \{1, 2, \ldots, \mathsf{K}\}. \tag{3.7}$$

Moreover, according to Corollary 3.1, when conditional on the datasets, we see that $\{\mathsf{c}_i\}$ are asymptotically iid Bernoulli random variables with probability α. This immediately completes the proof. □

Corollary 3.2 states that when n is large, we can essentially characterize the asymptotic distribution of DR. Therefore, as discussed in Remark 2.3, we can use it as the statistic to test (1.1) (see (2.11)).

3.3. Power Analysis

In this section, we study the power of the statistic $\mathbb{T}$ in (2.4) under the alternative that

$$\boldsymbol{H}_a : \Sigma_1 \neq \Sigma_2. \tag{3.8}$$

For notional simplicity, denote the ESDs of Σ_1 and Σ_2 as π_1 and π_2 and their associated kth moments as

$$\mathfrak{m}_k(\Sigma_i) = \int x^k \pi_i(\mathrm{d}x),\ i = 1, 2. \tag{3.9}$$

We point out that Algorithm 2.2 only uses statistics (2.4) for ϵ-efficient splits. Thus, we first study the power of $\mathbb{T}$ for efficiently split datasets.

Theorem 3.2. Suppose Assumption 3.1 holds. Moreover, for some small $\epsilon > 0$, we assume that $(\mathcal{X}^s, \mathcal{Y}^s, \mathcal{Z}^s)$ generated from Algorithm 2.1 is an ϵ-efficient data splitting satisfying Definition 2.1. For some small constants $\tau_1, \tau_2 > 0$, we assume that $n^{-1+\tau_2} < \eta_0 \le n^{-\tau_1}$. Moreover, suppose that for sufficiently large n and any constant $\mathsf{c} = O(n^{-1} + \phi^{-1/2})$, we have that

$$\phi^{1/2}|\mathfrak{m}_1(\Sigma_1) - \mathfrak{m}_1(\Sigma_2)| + \mathsf{c}|\mathfrak{m}_2(\Sigma_1) - \mathfrak{m}_2(\Sigma_2)| \neq 0. \tag{3.10}$$

Then given some Type I error rate α, suppose the alternative (3.8) holds in the sense that

$$\phi^{1/2}|\mathfrak{m}_1(\Sigma_1) - \mathfrak{m}_1(\Sigma_2)| + \phi^{-1/2}|\mathfrak{m}_2(\Sigma_1) - \mathfrak{m}_2(\Sigma_2)| > C_\alpha \eta_0^{-2} n^{-1}, \tag{3.11}$$

where the constant $C_\alpha \equiv C_\alpha(n) \uparrow \infty$ as $n \to \infty$, we have that

$$\mathbb{P}\left(|\mathbb{T}| > \sqrt{2\mathsf{v}} z_{1-\alpha/2}\right) = 1. \tag{3.12}$$

Proof. See Section A.2. □

Remark 3.3. A few remarks are in order. First, (3.10) is a mild condition and can be easily satisfied. In fact, we can actually remove this condition when $\phi^{-1/2} \ll n^{-1}$, or equivalently, $p \gg n^3$. In such a setting, since $\mathfrak{m}_2(\Sigma_i), i = 1, 2$, are bounded from above, we have that $\mathsf{c}|\mathfrak{m}_2(\Sigma_2) - \mathfrak{m}_2(\Sigma_1)| = O(n^{-1})$. Consequently, (3.11) implies (3.10) so we can remove (3.10). Second, (3.11) is generally a weak alternative. It suggests that we should use a relatively larger η_0 in order to increase the power. Third, the condition (3.11) does not impose any explicit structural assumptions on the form of the difference $\Sigma_1 - \Sigma_2$. Finally, the condition (3.11) also relies on the ratio $\phi = p/n$. It demonstrates that as the ratio ϕ increases, weaker alternatives may be sufficient. For example, if $p \gg n^3$, (3.11) reads as

$$|\mathfrak{m}_1(\Sigma_1) - \mathfrak{m}_1(\Sigma_2)| > C_\alpha \phi^{-1/2} \eta_0^{-2} n^{-1},$$

which can be much weaker than those used in Cai, Liu, and Xia (2013), and Li and Chen (2012).

Theorem 3.2 also yields the results of the power analysis of our Algorithm 2.2 in terms of the decision ratio in (2.10).

Corollary 3.3. Suppose Assumption 3.1 and (3.11) hold. For given Type I error rate α, when n is sufficiently large, we have that conditional on the datasets $(\mathcal{X}, \mathcal{Y})$

$$\mathsf{DR} = 1 + o_{\mathbb{P}}(1).$$

Proof. If $\mathcal{S}_1 = \emptyset$, then by Step one of our Algorithm 2.2, we have that $\mathsf{DR} = 1$. Otherwise, together with Step four of our Algorithm 2.2 and (3.12), we can see that $\mathsf{DR} = 1 + o_{\mathbb{P}}(1)$. This completes our proof. □

Corollary 3.3 implies that when n is large and the weak local alternative (3.11) holds, DR will converge to 1 with high probability. Consequently, our Algorithm 2.2 will be able to reject the null hypothesis under the weak alternative as in (3.11) for any threshold $\delta < 1$.

4. Numerical Results

In this section, we conduct extensive Monte Carlo simulations and two real data analysis to show the accuracy and powerfulness of our proposed test procedure Algorithm 2.2. For illustrations, we compare our Algorithm 2.2 (Proposed) with five state-of-the-art methods: CLX2013 Cai, Liu, and Xia (2013), LC2012 Li and Chen (2012), SY2010 Srivastava and Yanagihara (2010), HC2018 He and Chen (2018), and ZLGY2020 Zheng et al. (2019). Section C.4 compares the computational complexity

of all methods, showing our method is generally more efficient. For users' convenience, all these methods can be implemented using our R package `UHDtst`.

4.1. Numerical Simulations

In this section, we check and compare the accuracy and power via Monte Carlo simulations.

4.1.1. Simulation Setup

As in the second condition of Assumption 3.1, our two samples $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_j\}$ are generated according to $\boldsymbol{x}_i = \Sigma_1^{1/2}\boldsymbol{x}_i$ and $\boldsymbol{y}_j = \Sigma_2^{1/2}\boldsymbol{y}_j$, where $\boldsymbol{x}_i = (x_{is})$ and $\boldsymbol{y}_j = (y_{jt})$ satisfy (3.2) and (3.3). For the iid entries x_{is} and y_{jt}, we consider two different distributions: the standard Gaussian distribution $\mathcal{N}(0,1)$ with vanishing fourth cumulant and the two-point distribution that $\mathbb{P}(x = \sqrt{2}) = 1/3$ and $\mathbb{P}(x = -\sqrt{2}/2) = 2/3$ whose fourth cumulant is -1.5.

We formulate the null hypothesis $\boldsymbol{H}_0$ for the two population covariance matrices as

$$\boldsymbol{H}_0 : \Sigma_1 = \Sigma_2 \equiv \Sigma^*. \tag{4.1}$$

In the simulations, we will consider three different cases based on (4.1) as follows.

(Case I). We consider model two of Cai, Liu, and Xia (2013). For Σ^* in the null hypothesis (4.1), we consider the Toeplitz matrix $\Sigma^* = (\sigma_{ij}^*)$, where $\sigma_{ij}^* = 0.5^{|i-j|}$. For the alternative (3.8), we consider

$$\boldsymbol{H}_a : \Sigma_1 = \Sigma^*,\ \Sigma_2 = D^{1/2}\Sigma^* D^{1/2}.$$

Here $D = \mathrm{diag}(d_{ii})$, where d_{ii}'s are generated from $\mathrm{Unif}(0.5, 2.5)$.

(Case II). We consider case one of Li and Chen (2012). For Σ^* in (4.1), we consider $\Sigma^* = I$. For the alternative, we consider that

$$\boldsymbol{H}_a : \Sigma_1 = \Sigma^*,\ \Sigma_2 = \Sigma^* + \Delta.$$

Here for some constant $\vartheta > 0$, Δ is a banded matrix that

$$\Delta_{ij} = \vartheta^2 \cdot \mathbf{1}_{i=j} + \vartheta \cdot \mathbf{1}_{|i-j|=1}.$$

In other words, Σ_2 can be regarded as the covariance matrix of a p-dimensional realization of a stationary MA(1) process driven by $\mathcal{N}(0,1)$ random variables with parameter ϑ.

(Case III). For Σ^* in (4.1), we set $\Sigma^* = QDQ^\top$, where Q is some orthogonal matrix and $D = \mathrm{diag}(d_{ii})$ with d_{ii}'s being generated from $\mathrm{Unif}(3,6)$. For the alternative, we consider

$$\boldsymbol{H}_a : \Sigma_1 = \Sigma^*,\ \Sigma_2 = \Sigma^* + \varepsilon I_p, \tag{4.2}$$

where $\varepsilon > 0$ is some constant and I_p is the $p \times p$ identity matrix.

4.1.2. Simulation Results

In this section, we report and discuss the numerical results based on extensive Monte Carlo simulations. We conduct the simulations following the settings in Section 4.1.1. For the dimension and sample sizes, we consider $p = 6000$ and various combinations $(n_1, n_2) = (100,100), (100,150), (100,800), (100,1000)$. We report our results in Tables 1 and 2 and Figure 3. We elaborate our results in more details as follows.

Tables 1 and 2 summarize the results of the empirical size and power of our proposed method in Algorithm 2.2 and the other five methods in the literature (Cai, Liu, and Xia 2013; He and Chen 2018; Li and Chen 2012; Srivastava and Yanagihara 2010) for Gaussian samples and two-point samples, respectively. First, we conclude that across all the simulation settings, our proposed method (i.e., Proposed) is accurate and powerful. It also outperforms all the other methods in terms of both size and power. Second, LC2012 is reasonably accurate for all the simulation settings but lose their power in Case III. Third, due to the multiple testing procedure, HC2018 is powerful across all the settings but at the expense of being inaccurate. Fourth, SY2010 only works for Case II and is invalid for Cases I and III both in size and power. Fifth, when the samples are Gaussian and n_1 and n_2 are comparably large, CLX2013 works in Case I and II but loses its power in Case III. Moreover, if either n_1 and n_2 are incomparable or the samples follow two-point distribution, CLX2013 will be no longer accurate. Finally, ZLGY2020, comparing to CLX2013, succeeds in controlling the empirical size in the unbalanced cases. However, its power may fall under LC2012 in this ultra-high dimensional regime.

Before concluding this section, to show the mildness of the condition (3.11) and the powerfulness of our method, in Figure 3, using Case III with the alternative (4.2), we report how the simulated power changes with ε. It can be concluded that our proposed will achieve power one even for very weak alternatives, while all the other methods either are powerless or require much larger ε to have nontrivial power.

Additional simulations for scenarios with smaller gaps between Σ_1 and Σ_2 and cases with smaller sample sizes (n_1, n_2) further demonstrate that our method performs well compared to other methods in various settings. Details can be found in Section C.3 of the supplement.

4.2. Real Data Analysis

In this section, we consider the analysis of two gene expression datasets using our proposed method and compare it with the methods developed in Cai, Liu, and Xia (2013), He and Chen (2018), Li and Chen (2012), Srivastava and Yanagihara (2010), and Zheng et al. (2019). The first dataset is the clinical prostate cancer dataset Singh et al. (2002)[1] and the second one is the adult T-cell acute lymphocytic leukemia (ALL) dataset (Chiaretti et al. 2004).[2] We will see from the analysis below that while some of these methods (including ours) work for the first dataset, only our proposed method works for the second dataset.

[1]The dataset can be downloaded from *https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68907*

[2]The dataset can be loaded from the R package `ALL`; see *https://bioconductor.org/packages/release/data/experiment/html/ALL.html*

Table 1. Comparison of simulated Type I error and power for Gaussian samples.

		Methods/setting	(100, 100)	(100, 150)	(100,800)	(100,1000)
Empirical size	Case I	SY2010	0	0	0.007	0.017
		LC2012	0.049	0.049	0.07	0.067
		CLX2013	0.038	0.046	0.24	0.36
		HC2018	0.014	0.017	0.01	0.003
		ZLGY2020	0.047	0.043	0.063	0.057
		Proposed	0.045	0.047	0.051	0.048
	Case II	SY2010	0.035	0.033	0.06	0.06
		LC2012	0.054	0.062	0.05	0.05
		CLX2013	0.043	0.031	0.223	0.243
		HC2018	0.013	0.007	0.01	0
		ZLGY2020	0.049	0.061	0.052	0.043
		Proposed	0.048	0.049	0.052	0.05
	Case III	SY2010	0	0	0.003	0.01
		LC2012	0.048	0.049	0.057	0.067
		CLX2013	0.052	0.046	0.19	0.223
		HC2018	0.024	0.004	0.003	0.007
		ZLGY2020	0.067	0.076	0.093	0.053
		Proposed	0.047	0.05	0.051	0.051
Empirical power	Case I	SY2010	0	0	0.873	0.917
		LC2012	1	1	1	1
		CLX2013	1	1	1	1
		HC2018	1	1	1	1
		ZLGY2020	1	1	1	1
		Proposed	1	1	1	1
	Case II	SY2010	1	1	1	1
		LC2012	1	1	1	1
		CLX2013	0.947	1	1	1
		HC2018	1	1	1	1
		ZLGY2020	1	1	1	1
		Proposed	1	1	1	1
	Case III	SY2010	0	0	0.013	0.023
		LC2012	0.218	0.286	0.45	0.463
		CLX2013	0.067	0.057	1	1
		HC2018	1	1	1	1
		ZLGY2020	0.028	0.046	0.34	0.437
		Proposed	1	1	1	1

NOTE: Here we choose the type error $\alpha = 0.05$ and consider the setups in Section 4.1.1 for four different combinations of (n_1, n_2) with $p = 6000$. For Case II, we choose $\vartheta = 0.5$ for the alternative and for Case III, we choose $\varepsilon = 1$ for the alternative. In our R package `UTDtst`, the functions `TwoSampleTest`, `LC2012`, `CLX2013`, `SY2010`, `HC2018`, and `ZLGY2020` implement our proposed method, LC2012, CLX2013, SY2010, HC2018, and ZLGY2020, respectively. We report the results based on 1000 repetitions.

4.2.1. Prostate Cancer Data

The prostate cancer dataset (Singh et al. 2002) focuses on the gene expression patterns associated with clinical behaviors of prostate cancer. This study employed microarray expression analysis to discern the global biological variations that might be linked to the common pathological characteristics of prostate cancer.

The dataset categorizes observations into distinct groups. More specifically, it has 12,600 columns of gene expressions and comprises samples from two groups: a normal group with 50 samples and a tumor group with 52 samples. We point out that this dataset has been also used for analysis in Cai, Liu, and Xia (2013). However, to avoid some computational issue, they only select the top 5000 columns (genes) with the largest t-values in the sense of group means. In what follows, we will conduct our analysis on both this subsample with 5000 genes and all the samples with 12,600 genes.

We conduct the two sample covariance tests both within groups and between groups. More concretely, for within group test, we consider the normal group and divide the 50 samples into two subgroups with sample sizes $n_1 = 30, n_2 = 20$. For between group test, we use all 50 samples for normal group and all 52 samples for tumor group, that is, $n_1 = 50, n_2 = 52$. The results are summarized in Table 3 and we can make the following conclusions.

First, all of our proposed method, LC2012, HC2018, CLX2013, and ZLGY2020 will be able to accept the null hypothesis for the within group test and reject the null hypothesis for the between group test for both datasets with different numbers of genes. Second, SY2010 is able to accept the null hypothesis for the between group test but has no power to reject the null hypothesis.

4.2.2. Acute Lymphoblastic Leukemia Data

The second dataset (Chiaretti et al. 2004) contains gene expression of adult T-cell acute lymphocytic leukemia (ALL) of patients with different biological indices. This study focuses on the relation between overall gene expressions and molecular biology types, helping to reveal the mechanism between different ALL gene expressions and their responses to therapy and survival.

The dataset contains 128 patients and their genes with length of 12,625. There are six types of molecular biology in total. Here we only select the two groups with largest numbers of patients,

Table 2. Comparison of simulated Type I error and power for two-point samples.

		Methods/setting	(100, 100)	(100, 150)	(100,800)	(100,1000)
Empirical size	Case I	SY2010	0	0	0.01	0.01
		LC2012	0.048	0.050	0.043	0.046
		CLX2013	0.176	0.158	0.457	0.513
		HC2018	0.01	0.014	0.003	0.007
		ZLGY2020	0.054	0.062	0.06	0.054
		Proposed	0.049	0.051	0.048	0.047
	Case II	SY2010	0.051	0.033	0.117	0.077
		LC2012	0.046	0.059	0.066	0.05
		CLX2013	0.321	0.306	0.95	0.997
		HC2018	0.011	0.013	0.013	0.004
		ZLGY2020	0.058	0.06	0.083	0.078
		Proposed	0.051	0.052	0.048	0.048
	Case III	SY2010	0	0	0.02	0.02
		LC2012	0.044	0.051	0.047	0.06
		CLX2013	0.037	0.035	0.2	0.233
		HC2018	0.019	0.01	0.006	0.008
		ZLGY2020	0.058	0.058	0.05	0.047
		Proposed	0.051	0.051	0.047	0.048
Empirical power	Case I	SY2010	0	0	0.83	0.843
		LC2012	1	1	1	1
		CLX2013	1	1	1	1
		HC2018	1	1	1	1
		ZLGY2020	1	1	1	1
		Proposed	1	1	1	1
	Case II	SY2010	1	1	1	1
		LC2012	1	1	1	1
		CLX2013	1	1	1	1
		HC2018	1	1	1	1
		ZLGY2020	1	1	1	1
		Proposed	1	1	1	1
	Case III	SY2010	0	0	0	0.01
		LC2012	0.237	0.302	0.413	0.39
		CLX2013	0.038	0.051	1	1
		HC2018	1	1	1	1
		ZLGY2020	0.028	0.047	0.374	0.403
		Proposed	1	1	1	1

NOTE: Here we choose the type error $\alpha = 0.05$ and consider the setups in Section 4.1.1 for four different combinations of (n_1, n_2) with $p = 6000$. For Case II, we choose $\vartheta = 0.5$ for the alternative and for Case III, we choose $\varepsilon = 1$ for the alternative. In our R package `UTDtst`, the functions `TwoSampleTest`, `LC2012`, `CLX2013`, `SY2010`, `HC2018`, and `ZLGY2020` implement our proposed method, LC2012, CLX2013, SY2010, HC2018, and ZLGY2020, respectively. We report the results based on 1000 repetitions.

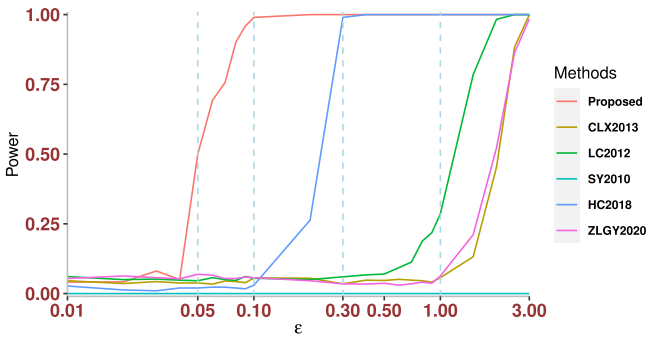

Figure 3. Comparison of power of different methods under the alternative (4.2) of Case III. We use Gaussian samples with $n_1 = 100, n_2 = 150, p = 6000$ and report based on 1000 repetitions.

NEG with size 74 and BCR/ABL with size 37. We conduct our study on the whole gene sequence that $p = 12{,}625$.

We conduct the two sample covariance tests both within groups and between groups. More concretely, for within group test, we consider the NEG group and divide the 74 samples into two subgroups with sample sizes $n_1 = 30, n_2 = 44$. For between group test, we use all 74 samples from the NEG group and all 37 samples from BCR/ABL group, that is, $n_1 = 74, n_2 = 37$. The results are summarized in Table 4.

Table 3. Comparison of results for the prostate cancer data.

Methods	Within group		Between groups	
	5000 data	12,600 data	5000 data	12,600 data
SY2010	Accept	Accept	Accept	Accept
LC2012	Accept	Accept	Reject	Reject
CLX2013	Accept	Accept	Reject	Reject
HC2018	Accept	Accept	Reject	Reject
ZLGY2020	Accept	Accept	Reject	Reject
Proposed	Accept	Accept	Reject	Reject

NOTE: Here 5000 data only uses $p = 5000$ genes as in Cai, Liu, and Xia (2013) and 12,600 data contains all genes.

Table 4. Comparison of results for the ALL data.

Methods	Within group	Between groups
SY2010	Accept	Accept
LC2012	Accept	Accept
CLX2013	Reject	Reject
HC2018	Reject	Accept
ZLGY2020	Reject	Reject
Proposed	Accept	Reject

We can see that for this dataset, only our proposed method works while SY2010, LC2012, and HC2018 fail to reject the null hypothesis for between group test, while CLX2013 and ZLGY2020 reject the null hypothesis for within group test.

5. Discussions

In this article, we consider the test of equality of two population covariance matrices (see (1.1)) of ultra-high dimensional (see (1.2)) random vectors. We propose a novel and adaptive test procedure which (i). does not require specific assumption (e.g., comparable or balancing, etc.) on the sizes of two samples; (ii). does not need quantitative or structural assumptions of the population covariance matrices; (iii). does not need the parametric distributions or the detailed knowledge of the moments of the two populations.

Our approach, outlined in Algorithm 2.2, has three key components. First, a data splitting procedure (Algorithm 2.1) ensures independence of data used in our test statistic from data used for selecting the location and bandwidth parameters. Second, we construct statistics based on a subset of eigenvalues from sample covariance matrices, with the subset determined by automatically selected location and bandwidth parameters using (2.2) and Algorithm B.1. This adaptive selection captures essential eigenvalues for distinguishing null and alternative hypotheses. Third, we compute a summary statistic (2.10) and a threshold δ via a calibration procedure detailed in Algorithm B.2.

The proposed methodology is highly inspired and justified by our theoretical development in random matrix theory. We establish the asymptotic distributions of the statistics used in our method and conduct the power analysis. We justify that our method is powerful under very weak alternatives. We also conduct extensive numerical simulations and show that our method significantly outperforms the existing ones developed in Cai, Liu, and Xia (2013), He and Chen (2018), Li and Chen (2012), and Srivastava and Yanagihara (2010), both in terms of size and power. Analysis of two real datasets is also carried out to demonstrate the usefulness and superior performance of our proposed methodology.

Several further works can be considered following the current article's spirit. First, besides the two-sample covariance matrix test, people are also interested in high-dimensional two-sample mean tests under various settings, as seen in Chen and Qin (2010), Chen, Li, and Zhong (2019), and Xue and Yao (2020). Proposing an adaptive, accurate, and powerful test for two-sample means under the ultra-high dimensional setup (1.2) is important. Second, we assume that the eigenvalues of Σ_1 and Σ_2 are bounded from above and below away from zero, which is realistic in many applications. However, in applications where a factor model is more beneficial, spiked covariance matrix models with a few larger or divergent eigenvalues may be considered (Fan, Guo, and Zheng 2022; Ke, Ma, and Lin 2023; Zhang et al. 2023). Generalizing our results and methods to the spiked model would be interesting. Third, since our algorithm involves multiple data splitting, it is worth exploring the implementation of Algorithm 2.2 in a parallel or distributed fashion (Dobriban and Owen 2019; Dobriban and Sheng 2021). Finally, exploring bootstrap extensions to the proposed testing framework is also an interesting direction for potential future works.

Supplementary Materials

In the supplement, we provide the details of the technical proof, the automated procedures for selecting the tuning parameters and additional numerical studies.

Acknowledgments

The authors would like to thank the editor, associate editor, and the anonymous referees for their insightful comments, which have significantly improved the article.

Disclosure Statement

The authors are listed alphabetically (and roughly equally contributed) and there are no competing interests to declare.

Funding

XCD is partially supported by NSF DMS-2113489 and DMS-2306439.

References

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis* (3rd ed.), Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley-Interscience. [1]

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009), "Corrections to LRT on Large-Dimensional Covariance Matrix by RMT," *The Annals of Statistics*, 37, 3822–3840. [1,2]

Bai, Z., and Silverstein, J. W. (2010), *Spectral Analysis of Large Dimensional Random Matrices* (2nd ed.), Springer Series in Statistics, New York: Springer. [6]

Bloemendal, A., Knowles, A., Yau, H.-T., and Yin, J. (2016), "On the Principal Components of Sample Covariance Matrices," *Probability Theory and Related Fields*, 164, 459–552. [6]

Cai, T., Liu, W., and Xia, Y. (2013), "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings," *Journal of the American Statistical Association*, 108, 265–277. [2,3,7,8,9,10,11]

Chen, S. X., Li, J., and Zhong, P-S. (2019), "Two-Sample and ANOVA Tests for High Dimensional Means," *The Annals of Statistics*, 47, 1443–1474. [11]

Chen, S. X., and Qin, Y.-L. (2010), "A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing," *The Annals of Statistics*, 38, 808–835. [11]

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010), "Tests for High-Dimensional Covariance Matrices," *Journal of the American Statistical Association*, 105, 810–819. [6]

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004), "Gene Expression Profile of Adult t-cell Acute Lymphocytic Leukemia Identifies Distinct Subsets of Patients with Different Response to Therapy and Survival," *Blood*, 103, 2771–2778. [8,9]

Ding, X., and Wang, Z. (2023), "Global and Local CLTs for Linear Spectral Statistics of General Sample Covariance Matrices When the Dimension is much Larger than the Sample Size with Applications," arXiv preprint arXiv:2308.08646. [2,6]

Ding, X., and Yang, F. (2018), "A Necessary and Sufficient Condition for Edge Universality at the Largest Singular Values of Covariance Matrices," *The Annals of Applied Probability*, 28, 1679–1738. [6]

Dobriban, E., and Owen, A. B. (2019), "Deterministic Parallel Analysis: An Improved Method for Selecting Factors and Principal Components," *Journal of the Royal Statistical Society*, Series B, 81, 163–183. [6,11]

Dobriban, E., and Sheng, Y. (2021), "Distributed Linear Regression by Averaging," *The Annals of Statistics*, 49, 918–943. [11]

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of the American Statistical Association*, 97, 77–87. [1]

Fan, J., Guo, J., and Zheng, S. (2022), "Estimating Number of Factors by Adjusted Eigenvalues Thresholding," *Journal of the American Statistical Association*, 117, 852–861. [11]

Gupta, A. K., and Tang, J. (1984), "Distribution of Likelihood Ratio Statistic for Testing Equality of Covariance Matrices of Multivariate Gaussian Models," *Biometrika*, 71, 555–559. [1]

He, J., and Chen, S. X. (2018), "High-Dimensional Two-Sample Covariance Matrix Testing via Super-Diagonals," *Statistica Sinica*, 28, 2671–2696. [2,3,6,7,8,11]

Hu, R., Qiu, X., and Glazko, G. (2010), "A New Gene Selection Procedure based on the Covariance Distance," *Bioinformatics*, 26, 348–354. [1]

Ke, Z. T., Ma, Y., and Lin, X. (2023), "Estimation of the Number of Spiked Eigenvalues in a Covariance Matrix by Bulk Eigenvalue Matching Analysis," *Journal of the American Statistical Association*, 118, 374–392. [6,11]

Li, J., and Chen, S. X. (2012), "Two Sample Tests for High-Dimensional Covariance Matrices," *The Annals of Statistics*, 40, 908–940. [2,3,7,8,11]

Li, Y., Schnelli, K., and Xu, Y. (2021), "Central Limit Theorem for Mesoscopic Eigenvalue Statistics of Deformed Wigner Matrices and Sample Covariance Matrices," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57, 506–546. [2]

Manly, B. F. J., and Rayner, J. C. W. (1987), "The Comparison of Sample Covariance Matrices Using Likelihood Ratio Tests," *Biometrika*, 74, 841–847. [1]

Nagao, H. (1973), "On Some Test Criteria for Covariance Matrix," *The Annals of Statistics*, 1, 700–709. [1]

O'Brien, P. C. (1992), "Robust Procedures for Testing Equality of Covariance Matrices," *Biometrics*, 48, 819–827. [1]

Perlman, M. D. (1980), "Unbiasedness of the Likelihood Ratio Tests for Equality of Several Covariance Matrices and Equality of Several Multivariate Normal Populations," *The Annals of Statistics*, 8, 247–263. [1]

Schott, J. R. (2007), "A Test for the Equality of Covariance Matrices When the Dimension is Large Relative to the Sample Sizes," *Computational Statistics & Data Analysis*, 51, 6535–6542. [1]

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., et al. (2002), "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, 1, 203–209. [8,9]

Srivastava, M. S., and Yanagihara, H. (2010), "Testing the Equality of Several Covariance Matrices with Fewer Observations than the Dimension," *Journal of Multivariate Analysis*, 101, 1319–1329. [2,3,7,8,11]

Sugiura, N., and Nagao, H. (1968), "Unbiasedness of Some Test Criteria for the Equality of One or Two Covariance Matrices," *The Annals of Mathematical Statistics*, 39, 1686–1692. [1]

Xue, K., and Yao, F. (2020), "Distribution and Correlation-Free Two-Sample Test of High-Dimensional Means," *The Annals of Statistics*, 48, 1304–1328. [11]

Yang, F. (2019), "Edge Universality of Separable Covariance Matrices," *Electronic Journal of Probability*, 24, 1–57. [6]

Yao, J., Zheng, S., and Bai, Z. (2015), *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press. [1,2]

Zhang, B., Pan, G., Yao, Q., and Zhou, W. (2023), "Factor Modeling for Clustering High-Dimensional Time Series," *Journal of the American Statistical Association*, 119, 1252–1263. [11]

Zhang, Q., Hu, J., and Bai, Z. (2017), "Optimal Modification of the LRT for the Equality of Two High-Dimensional Covariance Matrices," arXiv preprint arXiv:1706.06774. [2]

Zheng, S. (2012), "Central Limit Theorems for Linear Spectral Statistics of Large Dimensional *F*-matrices," *Annales de l'IHP Probabilités et statistiques*, 48, 444–476. [1]

Zheng, S., Bai, Z., and Yao, J. (2015), "Substitution Principle for CLT of Linear Spectral Statistics of High-Dimensional Sample Covariance Matrices with Applications to Hypothesis Testing," *The Annals of Statistics*, 43, 546–591. [2]

Zheng, S., Lin, R., Guo, J., and Yin, G. (2019), "Testing Homogeneity of High-Dimensional Covariance Matrices," *Statistica Sinica*, 30, 35–53. [2,7,8]

Zou, T., Lin, R., Zheng, S., and Tian, G.-L. (2021), "Two-Sample Tests for High-Dimensional Covariance Matrices Using both Difference and Ratio," *Electronic Journal of Statistics*, 15, 135–210. [2]