A SIMPLE AND ADAPTIVE TWO-SAMPLE TEST IN HIGH DIMENSIONS BASED ON L^2 NORM

By Jin-Ting Zhang[‡,*], Jia Guo[‡,*], Bu Zhou[‡,*], and Ming-Yen Cheng[§,†]

National University of Singapore[*], *and National Taiwan University*[†]

Testing the equality of two means is a fundamental inference problem. For high-dimensional data, which are commonly encountered nowadays, the conventional Hotelling's T^2-test either performs poorly or even becomes inapplicable. Several modifications have been proposed to address this challenging issue and shown to perform well. However, most of them are based on asymptotic normality of the null distributions of their test statistics which inevitably requires strong assumptions on the covariance structure. We study this serious issue thoroughly and propose an L^2-norm based test that works under much milder conditions and even when there are fewer observations than the dimension. Specially, to cope with possible non-normality of the null distribution we employ the Welch–Satterthwaite χ^2-approximation. We derive a sharp upper bound on the approximation error and use it to theoretically justify that the χ^2-approximation is preferred to normal approximation, even when the null distribution is indeed asymptotically normal. Simple ratio-consistent estimators for the parameters in the χ^2-approximation are given. Most importantly, while existing tests based on asymptotic normality are not, our test is adaptive to singularity or near singularity of the unknown covariance which is commonly seen in high dimensions and is the main cause of non-normality. The approximate and asymptotic powers are also investigated. Simulation studies and a real data application show that our test outperforms a number of existing tests in terms of size control, while the powers are comparable when their sizes are comparable.

[‡] The work was financially supported by the National University of Singapore academic research grant R-155-000-164-112.

[§] Research was supported by the Ministry of Science and Technology grant 104-2118-M-002-005-MY3.

MSC 2010 subject classifications: Primary 62H15; secondary 62F03.

Keywords and phrases: High-dimensional data; Hotelling's T^2-test; χ^2-type mixtures; Welch–Satterthwaite χ^2-approximation.

1. Introduction.

In the recent decades, with rapid development in data collection and storage techniques high-dimensional data are frequently collected in many fields such as medicine, genomics, finance and so on. For example, in microarray gene expression studies usually thousands of gene expression levels are measured on each subject. In high-dimensional data analysis, traditional methods may not be always applicable and are usually subject to instability. This has stimulated an abundant literature on new methods and theories in various settings. In regression, feature screening and selection, and dimension reduction are the mainstream approaches [6, 8, 14, 15, 16, 17, 26, 25, 34, 40, 41]. Such ideas have been extended to other problems such as classification and discriminant analysis [9, 12, 13, 18, 39].

We study testing difference of two means, which is a fundamental problem in the inference. For example, the colon data set analyzed in Section 5 contains expression levels of 2000 genes on 22 normal and 40 tumor colon tissues. Of interest is to check if the normal and tumor tissues have the same mean expression levels. Since the data dimension 2000 is much larger than the total sample size 62, the problem is no longer a classical finite-dimensional but rather a high-dimensional one which can be mathematically described as follows. Suppose we have two independent high-dimensional samples:

$$\mathbf{y}_{i1}, \dots, \mathbf{y}_{in_i} \text{ are i.i.d. with } \mathrm{E}(\mathbf{y}_{i1}) = \boldsymbol{\mu}_i,\ \mathrm{Cov}(\mathbf{y}_{i1}) = \boldsymbol{\Sigma}, i = 1, 2, \tag{1.1}$$

where the dimension p of $\mathbf{y}_{i1}$ is very large and may be close to or even much larger than the total sample size $n = n_1 + n_2$. The goal is to test whether the two mean vectors are equal:

$$H_0: \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1: \quad \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{1.2}$$

When p is fixed and much smaller than n, it is well known that the classical Hotelling [24]'s T^2-test is most powerful. Its test statistic is defined as

$$T_H = \frac{n_1 n_2}{n} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^\top \hat{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2),$$

where $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are the sample mean vectors and $\hat{\boldsymbol{\Sigma}}$ is the pooled sample covariance matrix. However, even when p is less than n but tends to ∞ in proportion to n, although Hotelling's T^2-test is still well defined, it has very low power as $\hat{\boldsymbol{\Sigma}}$ is usually nearly singular in this high-dimensional setting [1]. To overcome this difficulty, the seminal work by Bai and Saranadasa [1] proposed a non-exact test statistic which is equivalent to

$$T_{BS} = \frac{n_1 n_2}{n} \|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\|^2 - \mathrm{tr}(\hat{\boldsymbol{\Sigma}}),$$

where and throughout $\|\mathbf{a}\|$ denotes the L^2-norm of a vector $\mathbf{a}$ and $\operatorname{tr}(\mathbf{A})$ is the trace of a matrix $\mathbf{A}$. Bai and Saranadasa [1] derived asymptotic normality of T_{BS} under H_0, and showed theoretically and with intensive simulation studies that when p tends to ∞ proportionally with n their test has much higher power than Hotelling's T^2-test. Note that study of the high-dimensional two-sample problem can be dated back to Dempster [10, 11]. Another contribution of Bai and Saranadasa [1] is deriving the asymptotic normality of Dempster's non-exact test statistic under H_0.

Srivastava and Du [30] proposed a scale-invariant test by replacing $\hat{\mathbf{\Sigma}}$ in T_H with $\operatorname{diag}(\hat{\mathbf{\Sigma}})$. Chen and Qin [5] modified T_{BS} using U-statistics to treat the unequal covariances case and to allow p to be larger than n. Their test statistic reduces to T_{BS} in the considered equal covariances situation. Under some regularity conditions, these authors showed that the null distributions of their test statistics are asymptotically normal, and they all based their tests on the asymptotic normality. In particular, they all impose some strong assumptions on $\mathbf{\Sigma}$ detailed in (2.11)–(2.13). Further, Srivastava and Du [30] assumed $\mathbf{\Sigma}$ is positive definite so that its diagonal elements are always positive. When those regularity conditions are not satisfied, the null distributions may not be approximately normal and could even depart from normality seriously. See the supplement [45] for some illustrations.

To overcome the above mentioned problem, we propose to use an L^2-norm based test statistic:

$$T_n = \frac{n_1 n_2}{n}\|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\|^2. \tag{1.3}$$

When the two samples (1.1) are normal, we show that the null distribution of T_n is a χ^2-type mixture [43], which is generally skewed although it may be asymptotically normal under some (strong) regularity conditions. Therefore, a normal approximation may not be always applicable. To achieve adaptivity to the changing shape of the null distribution, we suggest to use the Welch–Satterthwaite (W-S) χ^2-approximation [28, 38]. When the two samples (1.1) are non-normal, we show that under some regularity conditions the asymptotic null distribution of T_n is also a χ^2-type mixture. Therefore, we can still apply the W-S χ^2-approximation. In both cases, we give simple formulae for estimating the parameters in the χ^2-approximation ratio-consistently.

The W-S χ^2-approximation has been used for more than six decades in providing accurate and effective solutions to the classical Behrens–Fisher problems in the context of normal data with fixed and low dimensions [2, 19, 27]. One question arises naturally: is it still applicable in high (and possibly varying) dimensions and both in the contexts of normal and non-normal data? To our knowledge, this work is the first attempt to thoroughly address this problem both theoretically and numerically. After carefully examining effect of high dimensionality in estimation of the parameters, we show that the W-S χ^2-approximation indeed provides a simple and adaptive test for high-dimensional normal and non-normal data, without requiring strong assumptions on $\mathbf{\Sigma}$. This appealing property is not shared by existing normal approximation based tests such as those proposed by Bai and Saranadasa [1], Chen and Qin [5] and Srivastava and Du [30].

Intuitively, for our test statistic T_n given in (1.3) the W-S χ^2-approximation should be better than the normal approximation for the following reasons. While the latter is based on asymptotic normality, the former is a non-asymptotic approach and so is expected to work well in more general cases. Besides, although both of the methods are two-cumulant matched approaches, the normal distribution has a fixed bell shape while the χ^2-distribution can be either symmetric or skewed. These facts also explain why the W-S χ^2-approximation can easily cope with the changing shape of the null distribution while the normal approximation cannot. Theoretically speaking, we show that the W-S χ^2-approximation is preferred in terms of approximation accuracy. Therefore, we believe that in high-dimensional settings it deserves much of our attention from both theoretical and practical viewpoints.

The main contributions of this work are as follows. When the two samples (1.1) are normal, we derive a sharp uniform bound on the approximation error of the W-S χ^2-approximation. This is the first theoretical justification for the method which for decades has been widely used in the context of finite and low dimensional Behrens–Fisher problems. Further, the error bound indicates that in high dimensions the W-S χ^2-approximation is at least comparable (when the correlation is moderate) and is often much better than the normal approximation (when the correlation is either low or high). In addition, we give a necessary and sufficient condition for the null distribution of T_n to be asymptotically normal as p tends to ∞. And we show that our test is adaptive to the varying shape of the null distribution no matter whether the two samples are normal or not. That is, when the null distribution is asymptotically normal the degrees of freedom d in the χ^2-approximation will tend to ∞, and when d is finite the asymptotic normality cannot hold. We also show that none of the conditions (2.11)–(2.13) imposed by Bai and Saranadasa [1], Chen and Qin [5] and Srivastava and Du [30] would hold when d is finite. When the two samples (1.1) are non-normal, we show that the regularity conditions imposed by Bai and Saranadasa [1] imply d tends to ∞ and they could not hold when d is finite. Note that shape of the null distribution is mainly determined by the unknown complex covariance and our test is adaptive to this without requiring strong conditions but simply capturing it via the parameter d. In particular, d being finite basically corresponds to $\boldsymbol{\Sigma}$ being singular or nearly singular which is usually the case in high dimensions. The approximate and asymptotic powers of the proposed test are also studied. Finally, we demonstrate via extensive simulation studies that in terms of size control our test outperforms several existing tests including those suggested by Bai and Saranadasa [1] and Chen and Qin [5], and their powers are comparable when their sizes are comparable.

There exist other approaches to the considered problem. Wang et al. [37] proposed a jackknife empirical likelihood test which requires $p = o(\sqrt{n})$ and the implementation is complicated. Srivastava et al. [33] suggested a random projection T^2-test. It is an exact test when the two samples are normal; however, the computational burden is heavy and it involves some additional tuning parameters which further complicate the computation. To deal with heavy-tailed distributions, Wang et al. [36] gave a nonparametric extension of the U-statistic based test of Chen and Qin [5] using spatial sign. However, following Chen and Qin [5], they imposed condition (2.12) which is shown to be restrictive. Note that our test can also cope with skewed distributions. More importantly, it automatically adapts to singularity or near singularity of $\boldsymbol{\Sigma}$ while none of the existing tests does.

In summary, the proposed test is superior to existing tests in many senses. It is simple and fast to compute; this is a major advantage in practice, in particular when both n and p are large. It is widely applicable under various situations of n_1, n_2, p and $\boldsymbol{\Sigma}$ and even different distributions of the data. In addition, it enjoys both superior size accuracy and good power. Further, the methodology can be extended to the case where the two samples have unequal covariance matrices and other high-dimensional problems such as testing equality of covariance matrices [4, 46].

The methods when the data are normal and non-normal are described in Sections 2 and 3 respectively. Simulation results and application to the colon data are given in Sections 4 and 5 respectively. We give some concluding remarks in Section 6 and leave proofs of the main results to the Appendix.

2. Methodologies for normal data.

Throughout this section we assume that the two samples (1.1) are normal. In this case, we have

$$(n_1 n_2/n)^{\frac{1}{2}}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \sim \mathcal{N}_p(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}) \quad \text{where} \quad \boldsymbol{\mu}_n = (n_1 n_2/n)^{\frac{1}{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{2.1}$$

In order to test (1.2), we need to derive the null distribution of T_n defined in (1.3). For this purpose, we set $\mathbf{x}_{ij} = \mathbf{y}_{ij} - \boldsymbol{\mu}_i, j = 1, 2, \ldots, n_i; i = 1, 2$. Then we have $\bar{\mathbf{x}}_i = \bar{\mathbf{y}}_i - \boldsymbol{\mu}_i, i = 1, 2$. We now write

$$T_n = T_{n0} + 2S_n + \|\boldsymbol{\mu}_n\|^2, \tag{2.2}$$

where

$$T_{n0} = n_1 n_2 n^{-1} \|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|^2, \;\; S_n = n_1 n_2 n^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{2.3}$$

Note that T_{n0} has the same distribution as T_n under the null hypothesis.

Note that $\mathbf{\Sigma}$ has the following singular value decomposition:

$$\mathbf{\Sigma} = \sum_{r=1}^{p} \lambda_r \mathbf{u}_r \mathbf{u}_r^\top, \tag{2.4}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ are the corresponding orthonormal eigenvectors. Let χ_v^2 denote a central chi-squared distribution with v degrees of freedom. Let $\overset{d}{=}$ and $\xrightarrow{L}$ denote equality in distribution and convergence in distribution respectively. We have the following useful theorem.

Theorem 1. *For any fixed $n_1, n_2,$ and p,*

$$T_{n0} \overset{d}{=} \sum_{r=1}^{p} \lambda_r A_r, \quad \textit{where} \;\; A_1, A_2, \ldots, A_p \overset{i.i.d.}{\sim} \chi_1^2. \tag{2.5}$$

In addition, the first three cumulants of T_{n0} are $E(T_{n0}) = tr(\mathbf{\Sigma})$, $Var(T_{n0}) = 2tr(\mathbf{\Sigma}^2)$ and $E[T_{n0} - E(T_{n0})]^3 = 8tr(\mathbf{\Sigma}^3)$. Further, we have $T_{n0}/tr(\mathbf{\Sigma}) \xrightarrow{L} T_0$ as $p \to \infty$, where $T_0 \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r A_r$ with the λ_r's being the eigenvalues of $\lim_{p\to\infty} \mathbf{\Sigma}/tr(\mathbf{\Sigma})$.

2.1. *Approximate and asymptotic distributions of T_{n0}.* Theorem 1 indicates that the distribution of T_{n0} is nonnegative and generally skewed. Therefore, it is natural to apply the well-known W-S χ^2-approximation, also known as the Box χ^2-approximation [2], to approximate the distribution of T_{n0}. Its key idea is to approximate the distribution of T_{n0} by that of a random variable R of the following form

$$R = \beta \chi_d^2 \tag{2.6}$$

via matching the first two cumulants [43]. Therefore, the W-S χ^2-approximation is also called the two-cumulant matched χ^2-approximation.

Remark 1. *By (2.5), when $\lambda_r = \lambda_1, r \leq k$, and $\lambda_r = 0, r > k$, for some $1 \leq k \leq p$ we have $T_{n0} \overset{d}{=} \lambda_1 \chi_k^2$. This is one of the reasons why the two-cumulant matched χ^2-approximation using (2.6) works well for T_{n0}.*

Some simple algebra shows that the parameters β and d are given by

$$\beta = \operatorname{tr}(\mathbf{\Sigma}^2)/\operatorname{tr}(\mathbf{\Sigma}) \quad \text{and} \quad d = \operatorname{tr}^2(\mathbf{\Sigma})/\operatorname{tr}(\mathbf{\Sigma}^2). \tag{2.7}$$

We call d the approximate degrees of freedom of the two-cumulant matched χ^2-approximation. One advantage of this χ^2-approximation is that it is simple to implement provided that the unknown parameters β and d can be consistently estimated from the data. Another advantage, as shown later in this section, is that it is adaptive to the distribution shape of T_{n0} in the sense that when T_{n0} is asymptotically normal d tends to ∞, and when d is a finite number T_{n0} is not asymptotically normal. This appealing property is intuitively clear as the χ^2-approximation is a finite-sample method and thus it would work in a wide range of situations.

The skewness of T_{n0} can be expressed as

$$\frac{\mathrm{E}[T_{n0} - \mathrm{E}(T_{n0})]^3}{\operatorname{Var}^{3/2}(T_{n0})} = (8/d^*)^{1/2}, \text{ where } d^* = \frac{\operatorname{tr}^3(\mathbf{\Sigma}^2)}{\operatorname{tr}^2(\mathbf{\Sigma}^3)}. \tag{2.8}$$

Therefore, the skewness of T_{n0} will tend to 0 as d^* tends to ∞, and we may use d^* as a measure of the symmetry and normality of T_{n0}. In fact, as $p \to \infty$, "$d^* \to \infty$" is a necessary and sufficient condition for the asymptotic normality of T_{n0} as shown in the next theorem.

THEOREM 2. *As $p \to \infty$, T_{n0} is asymptotically normal if and only if $d^* \to \infty$.*

Theorem 2 says that we can approximate the distribution of T_{n0} by a normal distribution only when d^* is large. Indeed d^* has some connection to the three-cumulant matched χ^2-approximation [43] in which the distribution of T_{n0} is approximated by that of a random variable of the form

$$R^* = \beta_0^* + \beta_1^* \chi^2_{d^*}. \tag{2.9}$$

Here the parameters β_0^*, β_1^* and d^* are determined via matching the first three cumulants of R^* and T_{n0} [43] and can be written as

$$\beta_0^* = \operatorname{tr}(\mathbf{\Sigma}) - \frac{\operatorname{tr}^2(\mathbf{\Sigma}^2)}{\operatorname{tr}(\mathbf{\Sigma}^3)}, \ \beta_1^* = \frac{\operatorname{tr}(\mathbf{\Sigma}^3)}{\operatorname{tr}(\mathbf{\Sigma}^2)}, \ d^* = \frac{\operatorname{tr}^3(\mathbf{\Sigma}^2)}{\operatorname{tr}^2(\mathbf{\Sigma}^3)}. \tag{2.10}$$

We call d^* the approximate degrees of freedom of the three-cumulant matched χ^2-approximation. Notice that d^* given in (2.10) coincides with the parameter d^* given in the expression for the skewness of T_{n0} (2.8).

To show the asymptotic normality of their test statistics in high dimensions, various authors imposed different sufficient conditions. In the current context, Bai and Saranadasa [1]'s sufficient condition can be expressed as

$$\lambda_{\max}^2 = o[\mathrm{tr}(\boldsymbol{\Sigma}^2)], \quad \text{as } p \to \infty, \tag{2.11}$$

where $\lambda_{\max}$ denotes the largest eigenvalue of $\boldsymbol{\Sigma}$; Chen and Qin [5]'s key condition reduces to

$$\mathrm{tr}(\boldsymbol{\Sigma}^4) = o[\mathrm{tr}^2(\boldsymbol{\Sigma}^2)], \quad \text{as } p \to \infty; \tag{2.12}$$

while Srivastava and Du [30]'s sufficient condition is given by

$$\mathrm{tr}(\boldsymbol{\Sigma}^r)/p \to a_r \in (0, \infty), r = 1, 2, 3, \quad \text{as } p \to \infty. \tag{2.13}$$

In practice, it is challenging to justify any of the conditions (2.11)–(2.13). By contrast, our L^2-norm based test with the two-cumulant matched χ^2-approximation does not require any of these conditions.

We show in the following corollary that any of (2.11)–(2.13) is also sufficient for the asymptotic normality of T_{n0} to hold.

Corollary 1. *For any n and p, we have*

$$d^* \geq \left[\frac{\lambda_{\max}^2}{tr(\boldsymbol{\Sigma}^2)}\right]^{-1}, \; d^* \geq \left[\frac{tr(\boldsymbol{\Sigma}^4)}{tr^2(\boldsymbol{\Sigma}^2)}\right]^{-1}, \; \text{and } d^* = p\frac{[tr(\boldsymbol{\Sigma}^2)/p]^3}{[tr(\boldsymbol{\Sigma}^3)/p]^2}. \tag{2.14}$$

Thus, any of the conditions (2.11), (2.12) and (2.13) implies that, as $p \to \infty$, $d^* \to \infty$ and T_{n0} is asymptotically normal.

In some other situations, "$d^* \to \infty$" may not be valid as shown in the following corollary.

Corollary 2. *Assume that $tr(\boldsymbol{\Sigma}^r)/p^r \to b_r \in (0, \infty)$ as $p \to \infty$, $r = 1, 2, 3$. Then $d^* \to b_2^3/b_3^2 \in (0, \infty)$ as $p \to \infty$ and so the distribution of T_{n0} is not asymptotically normal.*

Under the conditions of Corollary 2, T_{n0} is not asymptotically normal. In general, when d^* is finite, especially when the value of d^* is small, normal approximation may not be appropriate for T_{n0}, nor for any of the test statistics suggested by [1], [5] and [30].

Note that both of the approximate degrees of freedom d and d^* depend only on the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ as we always have $\mathrm{tr}(\boldsymbol{\Sigma}^k) = \sum_{r=1}^p \lambda_r^k$ for all $k = 1, 2, \ldots$. Some interesting facts about the relationship between d, d^* and p are established in the following theorem.

THEOREM 3. *We have (a) $1 \leq d^* \leq d \leq p$; (b) $d^* = d = 1$ if and only if only the first eigenvalue λ_1 is nonzero; and (c) $d^* = d = p$ if and only if all the eigenvalues $\lambda_r, r = 1, 2, \ldots, p$ are the same.*

Theorem 3 indicates that both d^* and d are always finite if p is finite. Combining Theorem 3 and Corollary 1, we also observe that none of the conditions (2.11), (2.12) and (2.13) is satisfied when d is finite. Notice from (2.7) and (2.8) that when the first k ($k < p$) eigenvalues of $\mathbf{\Sigma}$ are the same and the remaining ones are all 0 we have $d^* = d = k$. In general, we expect that when the first few eigenvalues of $\mathbf{\Sigma}$ are much larger than the remaining ones both d^* and d will take on smaller values, and when all the eigenvalues are nearly the same both d^* and d will take on larger values. It follows that both d^* and d are small (large) when the data are highly (less) correlated.

The following corollary follows from Theorems 2 and 3.

COROLLARY 3. *When d is bounded d^* is always bounded and so both T_{n0} and R will be asymptotically non-normal; when $d^* \to \infty$ we always have $d \to \infty$ and so we have*

$$\frac{T_{n0} - tr(\mathbf{\Sigma})}{[2tr(\mathbf{\Sigma}^2)]^{1/2}} \xrightarrow{L} \mathcal{N}(0,1), \quad \textit{and} \quad \frac{R - tr(\mathbf{\Sigma})}{[2tr(\mathbf{\Sigma}^2)]^{1/2}} \xrightarrow{L} \mathcal{N}(0,1). \tag{2.15}$$

Corollary 3 shows that the approximation R is adaptive to the shape of the distribution of T_{n0} in the following sense. When T_{n0} is asymptotically normal d will tend to ∞ so that R is also asymptotically normal. When d is finite so that R is not asymptotically normal, d^* is also finite so that T_{n0} is also not asymptotically normal. This advantageous property is not shared by the tests by [1], [5] and [30] as they are all based on asymptotic normality of their test statistics.

Corollary 3 also indicates that we may use d to determine if normal approximation to the distribution of T_{n0} is adequate. However, that is not needed and indeed not recommended for the following reasons. Theoretically speaking, R is adaptive to the shape of the distribution of T_{n0}, and Remark 2 given latter says that R is generally better than normal approximation in terms of accuracy (even when d is large). Also, numerically Section 4 show that R is adequate even when d is as small as 1 or 2.

Next, we give the following interesting example.

EXAMPLE 1. *Let $\mathbf{\Sigma} = \sigma^2 [(1-\rho)\mathbf{I}_p + \rho \mathbf{J}_p]$ be a compound symmetric matrix where $0 \leq \rho \leq 1$ and $\sigma^2 > 0$. Then by some simple algebra we have*

$$d^* = \frac{[(1-\rho^2) + \rho^2 p]^3 p}{[(1-\rho)^2(1+2\rho) + 3(1-\rho)\rho^2 p + \rho^3 p^2]^2} \quad \textit{and} \quad d = \frac{p}{1 + (p-1)\rho^2}.$$

Thus, when ρ is a fixed constant in $(0,1)$, we have $d^* \to 1$ and $d \to 1/\rho^2$ as $p \to \infty$. When $\rho = 0$ and 1 we have $d^* = d = p$ and $d^* = d = 1$, respectively. Set $\rho = Cp^{-\tau}$ where $0 < C < \infty$ and $0 < \tau < \infty$ are fixed constants. Then, we always have $d \to \infty$ as $p \to \infty$, but we have $d^* \to \infty$ only when $\tau > 1/2$. In fact, we have $d^* \to 1$ and $d^* \to (1+C^2)^3/C^6$ respectively when $0 < \tau < 1/2$ and $\tau = 1/2$. Note that the condition "$\rho = Cp^{-\tau}$ with $\tau > 1/2$" essentially corresponds to $\mathbf{\Sigma} - \sigma^2\mathbf{I}_p \to 0$ as $p \to \infty$, which is extremely rare in practice. Table 1 gives the values of d^* and d for different values of ρ and p. Observe that for a fixed $\rho \geq 0.10$, d^* is decreasing while d is increasing with increasing p; and for a fixed p, both d^* and d are decreasing with increasing ρ. Especially, when ρ is large, both d^* and d are small, indicating that a normal approximation is not adequate.

TABLE 1

Values of d^* and d for different values of ρ and p when $\mathbf{\Sigma} = \sigma^2\,[(1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p]$.

p	ρ	0	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.99	1
50	d^*	50	49.05	7.11	1.96	1.31	1.12	1.06	1.03	1.01	1.00	1.00	1.00	1
	d	50	49.78	33.6	16.9	9.24	5.65	3.77	2.68	2.00	1.54	1.23	1.02	1
500	d^*	500	295.9	1.54	1.09	1.03	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1
	d	500	476.2	83.5	23.9	10.9	6.19	3.98	2.77	2.04	1.56	1.23	1.02	1
1000	d^*	1000	252.3	1.26	1.05	1.02	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1
	d	1000	909.2	91.0	24.4	11.0	6.22	3.99	2.77	2.04	1.56	1.23	1.02	1

We denote the probability density function and normalized version of a random variable X by f_X and $\tilde{X} = [X - \mathrm{E}(X)]/\sqrt{\mathrm{Var}(X)}$, respectively. The following theorem gives a uniform bound on the approximation error of R.

THEOREM 4. *Let $\Delta = \lambda_{\max}^2/tr(\mathbf{\Sigma}^2)$ and $M = tr(\mathbf{\Sigma}^4)/tr^2(\mathbf{\Sigma}^2)$. Then, when $\Delta < 1/10$ and $d > 10$ we have*

$$\begin{aligned}\sup_x \left|f_{\tilde{T}_{n0}}(x) - f_{\tilde{R}}(x)\right| \;\leq\; & 0.1403\big\{[3 + \tfrac{3.8578}{(1-10\Delta)^{5/2}}]M + [3 + \tfrac{3.8578}{(1-10/d)^{5/2}}]/d \\ & +0.7040[(d^*)^{-1/2} - d^{-1/2}].\end{aligned}$$

REMARK 2. *By Theorem 1 (a) of [43], the error bound of the normal approximation to T_{n0} is $O[(d^*)^{-1/2}]$. By Theorem 4, the error bound of the two-cumulant matched χ^2-approximation R is $O(M) + O(d^{-1}) + O[(d^*)^{-1/2} - d^{-1/2}]$. Note that $O(M)$, $O(d^{-1})$ and $O[(d^*)^{-1/2} - d^{-1/2}]$ are of smaller orders or generally smaller than $O[(d^*)^{-1/2}]$. Thus it is theoretically justifiable that R is generally preferred to the normal approximation. In particular, when d^* and d are different (which is the case when the data are moderately or mildly correlated such as $0.1 \leq \rho \leq 0.7$ in Example 1) $O[(d^*)^{-1/2} - d^{-1/2}]$*

dominates $O(M)+O(d^{-1})$ *and it is smaller than although sometimes comparable to* $O[(d^*)^{-1/2}]$ *so that* R *is better than or comparable to the normal approximation. In addition, when* d^* *and* d *are close to each other (which occurs in the extreme cases where the data are either nearly independent or highly correlated such as* $\rho=0$ *or* $\rho \geq 0.8$ *in Example 1)* $O[(d^*)^{-1/2}-d^{-1/2}]$ *is close to* 0 *so that the error bound of* R *is roughly* $O(M)+O(d^{-1})$ *which is much smaller than* $O[(d^*)^{-1/2}]$*, showing that* R *is much better than the normal approximation. These conclusions are actually verified by simulation results presented in Section 4 and explain why our test has a much better size control than other approaches when* d *is small as demonstrated in Section 4.*

For decades the W-S χ^2-approximation has been a popular approach to univariate Behrens–Fisher problems but lacks theoretical justifications. Theorem 4 solves this long-term open problem. Note that the error bound given in Theorem 4 is much sharper than the error bound $O(d^{*-1/6})$ obtained by Chuang and Shih [7] for correlated χ^2-mixtures which cannot be used to argue that the W-S χ^2-approximation is better than normal approximation.

It is well known that when p is large the data at hand are usually highly correlated and so both d and d^* tend to take on small values. Therefore, in high dimensions it is usually the case that our test is still applicable while existing tests based on asymptotic normality, including those given by [1], [5] and [30], are not. Further, Remark 2 indicates that our approach is still superior even when the data are nearly independent and normal approximation is adequate.

We give the following remark before concluding this section.

Remark 3. *We prefer the two-cumulant matched* χ^2*-approximation* $R=\beta\chi^2_d$ *to the three-cumulant matched* χ^2*-approximation* $R^*=\beta_0^*+\beta_1^*\chi^2_{d^*}$ *for approximating the distribution of* T_{n0} *in high-dimensional settings, although* R^* *generally outperforms* R *in terms of convergence rate [7, 43]. The reasons are as follows. First of all,* R *is simpler and faster to compute as it only involves the two parameters* β *and* d *given in (2.7), while* R^* *involves the three parameters* β_0^*, β_1^* *and* d^* *given in (2.10). In particular, in high-dimensional settings it is rather challenging to estimate* β_0^*, β_1^* *and* d^* *consistently as they involve complicated terms such as* $tr(\mathbf{\Sigma}^3), tr^3(\mathbf{\Sigma}^2)$ *and* $tr^2(\mathbf{\Sigma}^3)$*. We may construct some ratio-consistent estimators for them using U-statistics or the techniques used by [5]. However, those formulae are complicated in their forms and time consuming to compute. In fact, we found that the computational time of Chen and Qin [5]'s test is about* 10 *times of that of our test; see the Supplement [45] for more details. Secondly, Zhang [44] showed via some simulations that when the coefficients of a* χ^2*-type mix-*

ture are all positive (which is the case for T_{n0}) R and R^* are comparable in terms of accuracy. In particular, when $d = d^*$ (which is the case when $\rho = 0$ or 1 in Example 1) R is comparable to or even better than R^*. Thirdly, R has the same range $[0, \infty)$ as T_{n0} while R^* has the range $[\beta_0^*, \infty)$ which differs from the range of T_{n0} unless $\beta_0^* = 0$. But, when we restrict β_0^* to be 0, R^* reduces to R. The last but not the least reason is that for non-Gaussian data, it is also not easy to find a simple formula for the third cumulant of T_n under the null hypothesis; see Remark 5 for further details.

2.2. *Implementation.* We need to estimate $\operatorname{tr}(\mathbf{\Sigma})$, $\operatorname{tr}^2(\mathbf{\Sigma})$ and $\operatorname{tr}(\mathbf{\Sigma}^2)$ ratio-consistently in order to apply the W-S χ^2-approximation. Let $\hat{\theta}_{n,p}$ be an estimator of $\theta_{n,p}$, a non-random quantity depending on n and p which may tend to ∞ as $n, p \to \infty$. We say $\hat{\theta}_{n,p}$ is ratio-consistent in probability for $\theta_{n,p}$ if $\hat{\theta}_{n,p}/\theta_{n,p} \xrightarrow{P} 1$ as $n, p \to \infty$, where $\xrightarrow{P}$ denotes convergence in probability. The usual unbiased estimator of $\mathbf{\Sigma}$ is the pooled sample covariance matrix $\hat{\mathbf{\Sigma}} = (n-2)^{-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^\top$. Since the two samples (1.1) are normal, $\hat{\mathbf{\Sigma}}$ follows the Wishart distribution $W_p[n-2, \mathbf{\Sigma}/(n-2)]$. By Lemma S.3 given in the Supplement [45], $\operatorname{tr}(\hat{\mathbf{\Sigma}})$,

$$\begin{aligned} \widehat{\operatorname{tr}^2(\mathbf{\Sigma})} &= \tfrac{(n-2)(n-1)}{(n-3)n}\left[\operatorname{tr}^2(\hat{\mathbf{\Sigma}}) - \tfrac{2}{n-1}\operatorname{tr}(\hat{\mathbf{\Sigma}}^2)\right] \text{ and} \\ \widehat{\operatorname{tr}(\mathbf{\Sigma}^2)} &= \tfrac{(n-2)^2}{(n-3)n}\left[\operatorname{tr}(\hat{\mathbf{\Sigma}}^2) - \tfrac{1}{n-2}\operatorname{tr}^2(\hat{\mathbf{\Sigma}})\right] \end{aligned} \tag{2.16}$$

are unbiased and ratio-consistent estimators of $\operatorname{tr}(\mathbf{\Sigma}), \operatorname{tr}^2(\mathbf{\Sigma})$ and $\operatorname{tr}(\mathbf{\Sigma}^2)$, respectively, uniformly in p. Then by (2.7),

$$\hat{\beta} = \widehat{\operatorname{tr}(\mathbf{\Sigma}^2)}/\operatorname{tr}(\hat{\mathbf{\Sigma}}) \text{ and } \hat{d} = \widehat{\operatorname{tr}^2(\mathbf{\Sigma})}/\widehat{\operatorname{tr}(\mathbf{\Sigma}^2)} \tag{2.17}$$

are natural estimators of β and d. For any nominal level $\alpha > 0$, let $\chi_d^2(\alpha)$ denote the upper 100α percentile of χ_d^2. Theorem 5 stated below shows that $\hat{\beta}, \hat{d}$ and $\hat{\beta}\chi_{\hat{d}}^2(\alpha)$ are ratio-consistent for β, d and $\beta\chi_d^2(\alpha)$, respectively. Based on this result, using (2.17) we can conduct our test via using the approximate critical value $\hat{\beta}\chi_{\hat{d}}^2(\alpha)$ or the approximate p-value $\Pr(\chi_{\hat{d}}^2 \geq T_n/\hat{\beta})$.

THEOREM 5. *As $n \to \infty$, we have $\hat{\beta}/\beta \xrightarrow{P} 1$, $\hat{d}/d \xrightarrow{P} 1$ and $\hat{\beta}\chi_{\hat{d}}^2(\alpha)/\left[\beta\chi_d^2(\alpha)\right] \xrightarrow{P} 1$, uniformly in p.*

With the ratio-consistent estimators of $\operatorname{tr}(\mathbf{\Sigma})$ and $\operatorname{tr}(\mathbf{\Sigma}^2)$ given in (2.16), we have the following corollary which is useful when we analyze the asymptotic power of our test.

Corollary 4. *Assume $d^* \to \infty$ as $p \to \infty$. Then we have, as $n, p \to \infty$,*

$$\big[T_{n0} - tr(\hat{\mathbf{\Sigma}})\big] \big/ \big[2\widehat{tr(\mathbf{\Sigma}^2)}\big]^{1/2} \xrightarrow{L} \mathcal{N}(0,1). \tag{2.18}$$

Remark 4. *In practice, n_1, n_2 and p are always finite although p can be large, e.g., $p = 1000$. Then both d^* and d are finite (although they can be large) and hence the two-cumulant matched χ^2-approximation R can always be used and should always be recommended in practice. Besides, Remark 2 shows that even when the normal approximation is adequate, R is still comparable or better in terms of size control. On the other hand, to show the adaptivity of our test and to derive its asymptotic power, we still need to investigate the asymptotic normality of T_{n0} when $d^* \to \infty$.*

2.3. *Approximate and asymptotic powers.* In this section, we investigate the approximate and asymptotic powers of our test when d^* is finite and infinite, respectively. Recall that we have the decomposition (2.2) with T_{n0} and S_n defined in (2.3) and $\mathrm{Var}(S_n) = \boldsymbol{\mu}_n^\top \mathbf{\Sigma} \boldsymbol{\mu}_n$ where $\boldsymbol{\mu}_n$ is defined in (2.1). Similar to [1] and [5], we consider the power of T_n under the following local alternatives:

$$\text{As } n, p \to \infty, \;\; \boldsymbol{\mu}_n^\top \mathbf{\Sigma} \boldsymbol{\mu}_n = o[\mathrm{tr}(\mathbf{\Sigma}^2)]. \tag{2.19}$$

In this case $T_n = \|\boldsymbol{\mu}_n\|^2 + T_{n0}[1 + o_p(1)]$ as $n \to \infty$.

In the following theorem we consider the approximate power of the proposed test when d tends to a finite number as $p \to \infty$.

Theorem 6. *Assume d tends to a finite number as $p \to \infty$ and $n_1/n \to \tau \in (0,1)$ as $n \to \infty$. Then under the local alternatives (2.19), as $n, p \to \infty$, the power of our test based on T_n can be expressed as*

$$\Pr\big[T_n/\hat{\beta} \geq \chi^2_{\hat{d}}(\alpha)\big] \approx \Pr\big[\chi^2_d \geq \chi^2_d(\alpha) - n\tau(1-\tau)\beta^{-1}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2\big]. \tag{2.20}$$

In particular, when the conditions of Corollary 2 are satisfied and when $p/n \to c \in (0,\infty)$ as $n, p \to \infty$, we have $\beta/n \to cb_2/b_1 \in (0,\infty)$ and $d \to b_1^2/b_2 \in (0,\infty)$.

We now consider the asymptotic power when $d^* \to \infty$ as $p \to \infty$. In this case, d also tends to ∞ and by corollaries 3 and 4 the approximate critical value of T_n is $\mathrm{tr}(\hat{\mathbf{\Sigma}}) + [2\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)}]^{1/2} z_\alpha$ and the associated power is $\Pr\big\{\big[T_n - \mathrm{tr}(\hat{\mathbf{\Sigma}})\big] \big/ \big[2\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)}\big]^{1/2} \geq z_\alpha$, where z_α is the $100(1-\alpha)$ percentile of $\mathcal{N}(0,1)$. Let $\Phi(\cdot)$ denote the cumulative distribution function of $\mathcal{N}(0,1)$.

THEOREM 7. *Assume that $d^* \to \infty$ as $p \to \infty$ and $n_1/n \to \tau \in (0,1)$ as $n \to \infty$. Then under the local alternatives (2.19), as $n, p \to \infty$, the asymptotic power of our test can be expressed as*

$$\Pr\left\{ \frac{T_n - tr(\hat{\mathbf{\Sigma}})}{\big[2\widehat{tr(\mathbf{\Sigma}^2)}\big]^{1/2}} \geq z_\alpha \right\} = \Phi\bigg(-z_\alpha + \frac{n\tau(1-\tau)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\big[2\widehat{tr(\mathbf{\Sigma}^2)}\big]^{1/2}} \bigg) + o(1). \tag{2.21}$$

Note that when $d^* \to \infty$, the asymptotic power (2.21) of our test is the same as that of the test given by Bai and Saranadasa [1].

3. Methodologies for non-normal data. When the two samples (1.1) are non-normal, we show later in Theorem 8 that under some regularity conditions the asymptotic null distribution of T_n given in (1.3) is still a χ^2-type mixture. Therefore, we can still apply the two-cumulant χ^2-approximation (2.6) to the distribution of T_{n0} given in (2.3) which has the same distribution as T_n under the null hypothesis.

In the Supplement [45] we show that

$$\mathrm{E}(T_{n0}) = \mathrm{tr}(\mathbf{\Sigma}), \qquad \mathrm{Var}(T_{n0}) = 2\mathrm{tr}(\mathbf{\Sigma}^2) + \delta, \tag{3.1}$$

where

$$\begin{aligned} \delta &= \left(\tfrac{n_2}{n}\right)^2 \tfrac{\kappa_{1,11}}{n_1} + \left(\tfrac{n_1}{n}\right)^2 \tfrac{\kappa_{2,11}}{n_2}, \\ \kappa_{i,11} &= \mathrm{E}\|\mathbf{y}_{i1} - \boldsymbol{\mu}_i\|^4 - \mathrm{tr}^2(\mathbf{\Sigma}) - 2\mathrm{tr}(\mathbf{\Sigma}^2), i = 1, 2. \end{aligned} \tag{3.2}$$

Note that $\kappa_{i,11}$ is an important quantity used to measure the non-normality of $\mathbf{y}_{i1}, i = 1, 2$, [23]. As in the previous section, we approximate the distribution of T_{n0} using that of $R = \beta\chi^2_d$ via matching the first two cumulants of T_{n0} and R. The parameters β and d are then given by

$$\beta = \big[\mathrm{tr}(\mathbf{\Sigma}^2) + \delta/2\big]/\mathrm{tr}(\mathbf{\Sigma}), \quad d = \mathrm{tr}^2(\mathbf{\Sigma})/\big[\mathrm{tr}(\mathbf{\Sigma}^2) + \delta/2\big]. \tag{3.3}$$

When the two samples (1.1) are normal, the above two formulae reduce to those given in (2.7).

REMARK 5. *By (3.1) and (3.2), the second cumulant (variance) of T_{n0} already involves the fourth central moments. Then we can expect that the third cumulant of T_{n0} involves even higher order central moments. Therefore, it would be difficult to find an explicit formula for it, not mentioning how to estimate it ratio-consistently. This is another reason why we prefer to use the two-cumulant matched χ^2-approximation.*

To conduct the proposed test, we need to estimate the unknown parameters β and d given in (3.3) ratio-consistently. Under some regularity conditions, Himeno and Yamada [23] showed that for $i = 1, 2$, based on the i-th sample only, the unbiased and ratio-consistent estimators of $\mathrm{tr}(\mathbf{\Sigma}), \mathrm{tr}(\mathbf{\Sigma}^2), \mathrm{tr}^2(\mathbf{\Sigma})$, and $\kappa_{i,11}$ are given by $\mathrm{tr}(\hat{\mathbf{\Sigma}}_i)$ and

$$\begin{aligned}
a_i &= \tfrac{n_i-1}{n_i(n_i-2)(n_i-3)}[(n_i-1)(n_i-2)\mathrm{tr}(\hat{\mathbf{\Sigma}}_i^2) + \mathrm{tr}^2(\hat{\mathbf{\Sigma}}_i) - n_iQ_i],\\
b_i &= \tfrac{n_i-1}{n_i(n_i-2)(n_i-3)}[2\mathrm{tr}(\hat{\mathbf{\Sigma}}_i^2) + (n_i^2-3n_i+1)\mathrm{tr}^2(\hat{\mathbf{\Sigma}}_i) - n_iQ_i],\\
\hat{\kappa}_{i,11} &= \tfrac{-1}{(n_i-2)(n_i-3)}[2(n_i-1)^2\mathrm{tr}(\hat{\mathbf{\Sigma}}_i^2) + (n_i-1)^2\mathrm{tr}^2(\hat{\mathbf{\Sigma}}_i) - n_i(n_i+1)Q_i],
\end{aligned}$$

where $\hat{\mathbf{\Sigma}}_i = (n_i-1)^{-1}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\bar{\mathbf{y}}_i)(\mathbf{y}_{ij}-\bar{\mathbf{y}}_i)^\top$ and $Q_i = (n_i-1)^{-1}\sum_{j=1}^{n_i}\|\mathbf{y}_{ij}-\bar{\mathbf{y}}_i\|^4, i = 1, 2$. Thus, based on the two samples (1.1) together, the unbiased and ratio-consistent estimators of $\mathrm{tr}(\mathbf{\Sigma}), \mathrm{tr}(\mathbf{\Sigma}^2)$ and $\mathrm{tr}^2(\mathbf{\Sigma})$ are given by

$$\begin{aligned}
\mathrm{tr}(\hat{\mathbf{\Sigma}}) &= [(n_1-1)\mathrm{tr}(\hat{\mathbf{\Sigma}}_1) + (n_2-1)\mathrm{tr}(\hat{\mathbf{\Sigma}}_2)]/(n-2),\\
\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)} &= [(n_1-1)a_1 + (n_2-1)a_2]/(n-2),\\
\widehat{\mathrm{tr}^2(\mathbf{\Sigma})} &= [(n_1-1)b_1 + (n_2-1)b_2]/(n-2).
\end{aligned} \tag{3.4}$$

Plugging these ratio-consistent estimators into (3.2) and (3.3), we have

$$\hat{\delta} = \left(\frac{n_2}{n}\right)^2\frac{\hat{\kappa}_{1,11}}{n_1} + \left(\frac{n_1}{n}\right)^2\frac{\hat{\kappa}_{2,11}}{n_2}, \quad \hat{\beta} = \frac{\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)} + \hat{\delta}/2}{\mathrm{tr}(\hat{\mathbf{\Sigma}})}, \quad \hat{d} = \frac{\widehat{\mathrm{tr}^2(\mathbf{\Sigma})}}{\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)} + \hat{\delta}/2}.$$

Thus, Theorem 5 continues to hold when the data are non-normal.

As in the normal data case, in practice the two-cumulant matched χ^2-approximation can always be used without further assumptions because n_1, n_2 and p are always finite so that $\hat{\beta}$ and $\hat{d}$ are also finite. On the other hand, to study the adaptivity and asymptotic power of our test we need to investigate the asymptotic distributions of T_{n0}, T_n and R under some additional conditions, e.g. those by Bai and Saranadasa [1] as given below.

BS Assumptions:

A1. $\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \mathbf{\Gamma}\mathbf{z}_{ij}, j = 1, ..., n_i; i = 1, 2$ where $\mathbf{\Gamma} : p \times p$ satisfies $\mathbf{\Gamma}\mathbf{\Gamma}^\top = \mathbf{\Sigma}$ and $\mathbf{z}_{ij}$'s are i.i.d. p-vectors, with $\mathrm{E}(\mathbf{z}_{ij}) = \mathbf{0}$, and $\mathrm{Cov}(\mathbf{z}_{ij}) = \mathbf{I}_p$.

A2. Assume $\mathrm{E}z_{ijk}^4 = 3 + \Delta < \infty$ where z_{ijk} is the k-th component of $\mathbf{z}_{ij}$, Δ is some constant, and $\mathrm{E}(z_{ij1}^{v_1}z_{ij2}^{v_2}...z_{ijp}^{v_p}) = 0$ (or 1) when there is one $v_k = 1$ (or there are two $v_k = 2$) whenever $v_1 + \cdots + v_m = 4$.

A3. As $n, p \to \infty$, $p/n \to c \in (0, \infty)$ and $n_1/n \to \tau \in (0, 1)$.

A4. Condition (2.11) holds, i.e., $\lambda_{\max}^2 = o[\mathrm{tr}(\mathbf{\Sigma}^2)]$ as $p \to \infty$.

We first obtain the asymptotic distribution of T_{n0} under Assumptions A1, A2, and A3 in the following theorem.

THEOREM 8. *Let $A_1, A_2, \ldots, A_r, \ldots$ be i.i.d. χ_1^2 random variables. For any fixed finite p, we have $T_{n0} \xrightarrow{L} T_0$ as $n \to \infty$ with $n_1/n \to \tau \in (0,1)$ where $T_0 \overset{d}{=} \sum_{r=1}^{p} \lambda_r A_r$ with λ_r's being the eigenvalues of $\mathbf{\Sigma}$. Further, as $n, p \to \infty$, under Assumptions A1, A2, and A3, we have $T_{n0}/tr(\mathbf{\Sigma}) \xrightarrow{L} T_0$ where $T_0 \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r A_r$ with λ_r's being the eigenvalues of $\lim_{p\to\infty} \mathbf{\Sigma}/tr(\mathbf{\Sigma})$.*

Theorem 8 indicates that the asymptotic null distribution of T_n (1.3) is generally a χ^2-type mixture which is nonnegative and usually skewed. Therefore, normal approximation to the null distribution of T_n is not always applicable even when the sample sizes are large. In the following theorem we show that with the additional assumption A4, i.e. (2.11), both T_{n0} and R are asymptotically normal.

THEOREM 9. *Under the BS assumptions, as $n, p \to \infty$, we have $d \to \infty$ so that*

$$[T_{n0} - tr(\mathbf{\Sigma})]/[2tr(\mathbf{\Sigma}^2)]^{1/2} \xrightarrow{L} \mathcal{N}(0,1), \quad [R - tr(\mathbf{\Sigma})]/[2tr(\mathbf{\Sigma}^2)]^{1/2} \xrightarrow{L} \mathcal{N}(0,1).$$

Theorem 9 indicates that under the BS assumptions, R adapts to the asymptotic normality of T_{n0} as $n, p \to \infty$. Besides, when d tends to a finite number the BS assumptions are unlikely to hold as otherwise d tends to ∞ as $n, p \to \infty$. From the proof of Theorem 9 given in the Appendix, it is seen that Assumption A4 is unlikely to hold when d is bounded.

We have the following corollary which parallels Corollary 4.

COROLLARY 5. *Under the BS assumptions and H_0, as $n, p \to \infty$ we have the asymptotic normality (2.18) with $tr(\hat{\mathbf{\Sigma}})$ and $\widehat{tr(\mathbf{\Sigma}^2)}$ given in (3.4).*

We conclude this section via briefly investigating the approximate and asymptotic powers of T_n under the local alternatives (2.19). We have the following two theorems which parallel Theorems 6 and 7.

THEOREM 10. *Assume that as $n, p \to \infty$, d tends to a finite number and as $n \to \infty$, $n_1/n \to \tau \in (0,1)$. Then under the local alternatives (2.19), as $n, p \to \infty$, the approximate power of our test can still be expressed as the RHS of (2.20). In particular, when the conditions of Corollary 2 are satisfied and when $p/n \to c \in (0,\infty)$ and $\kappa_{i,11} = O[tr(\mathbf{\Sigma}^2)]$ as $n, p \to \infty$, we have $\beta/n \to cb_2/b_1 \in (0,\infty)$ and $d \to b_1^2/b_2 \in (0,\infty)$.*

THEOREM 11. *Under the BS assumptions and the local alternatives (2.19), as $n, p \to \infty$, the asymptotic power of our test is given by the RHS of (2.21).*

4. Simulation studies.

We conducted intensive simulation studies to compare the proposed test with some existing tests for the two-sample problems in high dimensions. For easy reference, we denote it as L2N (or L2D) when the two samples (1.1) are assumed to be normal (or non-normal) so that the two-cumulant matched χ^2-approximation developed in Section 2 (or Section 3) is applied. We compared L2N, L2D and some existing tests under various simulation settings in terms of size and power, aiming to see if L2N and L2D work well when the two samples (1.1) are actually normal or non-normal and how they perform as compared with those existing tests. To measure the overall performance of a test in maintaining the nominal size, we define its average relative error as ARE $= 100M^{-1}\sum_{j=1}^{M}|\hat{\alpha}_j - \alpha|/\alpha$, where α is the nominal size (e.g., 5%) and $\hat{\alpha}_j, j = 1, 2, \ldots, M$, denote the empirical sizes under M simulation settings. A smaller value of ARE indicates better performance of a test in terms of size control.

4.1. *Simulation 1.*

In this simulation study, we shall compare L2N and L2D against the tests proposed by Bai and Saranadasa [1] and Chen and Qin [5], denoted as BS and CQ, respectively. In each simulation run, we generate the two samples (1.1) using $\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}^{1/2}\mathbf{z}_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2$ where $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \delta\mathbf{h}$ and $\boldsymbol{\Sigma} = \sigma^2\left[(1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p\right]$ with the i.i.d. random variables $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2}, \ldots, z_{ijp})^\top$, $j = 1, 2, \ldots, n_i; i = 1, 2$ generated from the following three models:

Model 1. $z_{ijt}, t = 1, \ldots, p, \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$.

Model 2. $z_{ijt} = w_{ijt}/\sqrt{2}, t = 1, \ldots, p$ with $w_{ijt}, t = 1, \ldots, p, \overset{i.i.d.}{\sim} t_4$.

Model 3. $z_{ijt} = (w_{ijt} - 1)/\sqrt{2}, t = 1, \ldots, p$ with $w_{ijt}, t = 1, \ldots, p, \overset{i.i.d.}{\sim} \chi^2_1$.

Note that the tuning parameters $\delta, \mathbf{h}$ and ρ control the mean vector difference $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and the correlation. Also, the power of a test will increase as δ increases and the correlation will increase as ρ increases. For simplicity and without loss of generality, we set $\boldsymbol{\mu}_1 = \mathbf{0}$, $\mathbf{h} = \mathbf{u}/\|\mathbf{u}\|$ and $\sigma^2 = 1$ where $\mathbf{u} = (1, 2, \ldots, p)^\top$. To compare the performance of the considered tests with small, moderate and large values of d, we consider three cases of dimension $p = 50, 500, 1000$, three cases of sample sizes $[n_1, n_2] = [30, 50]$, $[120, 200]$ and $[240, 400]$, and three cases of correlation $\rho = 0.1, 0.5$ and 0.9. The empirical sizes and powers of the tests are obtained from $N = 10000$ simulation runs.

Table 2 displays the empirical sizes of the four tests with the last row displaying their ARE values associated with the three values of ρ. We may draw several useful conclusions from Table 2. The first one is that under a given setting, the empirical sizes of L2N and L2D are roughly the same, and they range from 4.64% to 6.86% and are below 6% in most of the cases. This shows that L2N and L2D are generally comparable and can be used

TABLE 2
Simulation 1: Empirical sizes (in percentages).

			$\rho = 0.1$				$\rho = 0.5$				$\rho = 0.9$			
Model	p	$\boldsymbol{n}$	L2N	L2D	BS	CQ	L2N	L2D	BS	CQ	L2N	L2D	BS	CQ
1	50	$[30, 50]$	5.47	5.58	6.32	5.93	6.05	5.97	7.30	7.10	5.31	5.36	7.12	7.47
		$[120, 200]$	5.29	5.26	6.13	6.26	5.25	5.27	6.75	6.67	4.77	4.79	6.42	6.83
		$[240, 400]$	5.40	5.38	6.20	6.34	5.04	5.05	6.33	6.60	4.64	4.65	6.54	7.03
	500	$[30, 50]$	6.58	6.53	7.08	7.26	5.89	5.95	7.30	7.32	5.53	5.62	7.29	7.39
		$[120, 200]$	5.59	5.59	6.07	6.79	5.70	5.68	7.10	6.54	5.23	5.23	7.05	6.56
		$[240, 400]$	5.91	5.91	6.36	6.42	5.45	5.43	6.73	6.67	4.68	4.70	6.46	6.40
	1000	$[30, 50]$	6.82	6.76	7.18	7.19	5.50	5.64	6.83	7.41	6.03	6.06	7.51	7.25
		$[120, 200]$	6.06	6.04	6.45	6.65	5.33	5.31	6.91	6.89	5.32	5.33	7.15	6.65
		$[240, 400]$	6.13	6.14	6.51	6.54	4.95	4.94	6.35	6.75	5.07	5.06	6.82	6.68
2	50	$[30, 50]$	4.79	4.97	5.42	6.11	5.87	6.00	7.34	7.43	5.74	5.81	7.30	7.69
		$[120, 200]$	5.01	5.13	5.81	5.94	5.67	5.70	7.06	7.15	4.86	4.82	6.73	7.04
		$[240, 400]$	5.29	5.33	6.23	5.71	5.45	5.47	7.02	6.96	5.13	5.13	7.03	6.72
	500	$[30, 50]$	6.00	6.07	6.49	7.24	6.01	6.01	7.34	7.14	5.89	5.91	7.73	7.65
		$[120, 200]$	6.09	6.11	6.60	6.87	5.50	5.48	6.81	7.19	4.85	4.81	6.61	7.04
		$[240, 400]$	6.08	6.10	6.66	6.69	5.41	5.39	7.08	7.36	4.76	4.75	6.58	6.93
	1000	$[30, 50]$	6.86	6.82	7.29	6.87	6.09	6.10	7.44	7.19	5.76	5.67	7.54	7.31
		$[120, 200]$	5.68	5.70	6.10	6.29	5.83	5.85	7.13	6.71	5.36	5.37	7.15	6.83
		$[240, 400]$	6.22	6.22	6.63	6.61	5.12	5.13	6.64	6.79	5.41	5.40	7.27	6.52
3	50	$[30, 50]$	5.07	5.16	5.88	6.08	5.50	5.50	6.90	7.20	5.65	5.65	7.23	7.23
		$[120, 200]$	4.94	4.92	5.69	6.17	5.62	5.63	6.95	7.17	5.28	5.24	7.06	6.93
		$[240, 400]$	5.33	5.35	6.29	6.45	5.56	5.56	6.98	6.82	4.98	5.01	6.94	7.29
	500	$[30, 50]$	5.84	5.89	6.18	7.21	5.97	6.04	7.22	7.80	5.61	5.61	7.32	7.55
		$[120, 200]$	5.97	5.99	6.55	6.69	5.78	5.78	7.04	7.01	5.68	5.71	7.49	6.89
		$[240, 400]$	5.90	5.91	6.32	6.23	5.74	5.73	7.24	6.89	5.37	5.35	7.21	6.83
	1000	$[30, 50]$	6.59	6.64	6.99	7.19	5.78	5.80	7.29	7.95	5.47	5.52	7.03	7.16
		$[120, 200]$	6.18	6.22	6.55	6.83	5.36	5.38	6.85	6.73	5.13	5.12	7.01	7.15
		$[240, 400]$	6.12	6.11	6.60	6.73	5.62	5.60	6.98	6.99	4.77	4.82	6.83	6.63
ARE			16.85	17.07	27.84	31.33	11.96	12.23	39.93	41.06	7.90	8.01	41.05	40.48

TABLE 3
Simulation 1: Empirical powers (in percentages).

				$\rho = 0.1$				$\rho = 0.5$				$\rho = 0.9$			
Model	p	$\boldsymbol{n}$	δ	L2N	L2D	BS	CQ	L2N	L2D	BS	CQ	L2N	L2D	BS	CQ
1	50	$[30, 50]$	1.0	36.86	36.91	39.18	39.01	13.87	13.90	16.62	16.79	9.57	9.55	12.00	12.83
		$[120, 200]$	0.6	50.65	50.61	53.29	53.72	16.36	16.33	19.21	20.15	11.46	11.41	14.37	14.39
		$[240, 400]$	0.4	46.09	46.09	48.70	48.86	15.44	15.42	18.21	17.87	10.24	10.23	13.04	13.69
	500	$[30, 50]$	3.2	53.85	53.82	55.13	53.66	15.04	15.11	17.32	17.24	10.20	10.24	12.37	12.46
		$[120, 200]$	1.5	46.66	46.64	48.25	47.88	13.00	13.01	15.51	15.50	8.45	8.47	10.72	11.86
		$[240, 400]$	1.2	57.92	57.90	59.45	59.37	14.87	14.86	17.85	17.36	10.02	10.04	13.04	12.27
	1000	$[30, 50]$	4.5	53.40	53.38	54.61	55.13	14.30	14.37	16.80	16.50	9.87	9.83	11.86	12.23
		$[120, 200]$	2.0	43.69	43.67	45.04	44.97	11.84	11.85	14.26	14.46	8.07	8.06	10.63	11.21
		$[240, 400]$	1.5	47.57	47.56	49.24	49.42	12.62	12.63	15.16	15.76	8.64	8.63	11.45	11.12
2	50	$[30, 50]$	1.0	35.57	36.35	38.11	39.85	14.19	14.23	16.96	16.94	9.86	9.96	12.05	12.57
		$[120, 200]$	0.6	50.73	51.00	53.13	53.28	17.58	17.64	20.28	20.29	11.11	11.17	14.00	14.74
		$[240, 400]$	0.4	46.10	46.33	48.92	48.98	14.96	14.96	18.12	18.92	10.46	10.49	13.25	13.27
	500	$[30, 50]$	3.2	52.69	52.91	54.06	53.68	14.55	14.57	17.10	17.17	9.58	9.61	12.25	12.33
		$[120, 200]$	1.5	47.58	47.60	49.00	48.59	12.60	12.62	15.06	15.26	8.92	8.96	11.49	11.19
		$[240, 400]$	1.2	57.87	57.89	59.40	58.70	14.56	14.58	17.24	16.79	10.28	10.29	13.03	12.53
	1000	$[30, 50]$	4.5	52.70	52.78	53.93	55.38	14.17	14.19	16.54	17.40	10.23	10.24	12.49	12.64
		$[120, 200]$	2.0	43.41	43.46	44.82	45.62	11.84	11.87	14.40	14.77	8.01	8.08	10.55	11.01
		$[240, 400]$	1.5	48.20	48.26	49.60	48.98	12.76	12.78	15.18	15.52	9.21	9.21	11.99	11.46
3	50	$[30, 50]$	1.0	35.15	36.04	37.73	40.21	13.43	13.51	15.94	16.11	9.77	9.70	12.09	12.65
		$[120, 200]$	0.6	50.60	50.83	53.01	53.73	17.20	17.28	20.31	19.68	10.91	10.91	13.86	14.35
		$[240, 400]$	0.4	46.23	46.36	48.95	49.93	15.82	15.88	18.85	18.75	10.86	10.83	13.65	13.64
	500	$[30, 50]$	3.2	53.16	53.38	54.67	53.81	14.85	14.90	17.07	17.22	10.03	9.91	12.47	13.30
		$[120, 200]$	1.5	47.24	47.34	48.97	48.96	12.43	12.46	15.09	15.20	8.84	8.87	11.49	11.36
		$[240, 400]$	1.2	58.39	58.41	59.82	59.34	14.82	14.81	17.53	17.10	9.99	9.99	12.67	12.57
	1000	$[30, 50]$	4.5	54.19	54.24	55.52	55.01	13.92	13.95	16.52	17.10	10.03	10.04	12.31	12.46
		$[120, 200]$	2.0	44.34	44.37	45.75	44.99	12.16	12.19	14.41	13.81	7.99	8.05	10.44	10.37
		$[240, 400]$	1.5	47.91	47.91	49.54	49.01	12.41	12.40	15.08	14.93	8.92	8.88	11.50	10.95

regardless if the data are normal or non-normal. The second conclusion is that BS and CQ are generally comparable, and they are generally liberal with their empirical sizes ranging from 5.42% to 7.95% and being around 7% in a large number of cases. The situation gets worse as the value of ρ increases. This is not surprising because the estimated approximate degrees of freedom of L2N and L2D shown in Table 4 become smaller as ρ increases, showing that the normal approximation used by BS and CQ is less adequate. The last but not the least conclusion is that in all of the considered settings L2N and L2D outperform BS and CQ in terms of size control, as indicated by their ARE values given in Table 2. It is also interesting to note that L2N and L2D have better size control even when $\rho = 0.1$ in which case d is very large (see Table 1) and BS and CQ are reasonable.

TABLE 4
Simulation 1: Estimated approximate degrees of freedom of L2N and L2D.

			$\rho = 0.1$		$\rho = 0.5$		$\rho = 0.9$	
Model	p	$\boldsymbol{n}$	L2N	L2D	L2N	L2D	L2N	L2D
1	50	$[30, 50]$	34.1	34.2	4.0	4.0	1.2	1.2
		$[120, 200]$	33.7	33.7	3.8	3.8	1.2	1.2
		$[240, 400]$	33.6	33.6	3.8	3.8	1.2	1.2
	500	$[30, 50]$	89.3	89.3	4.2	4.2	1.2	1.2
		$[120, 200]$	84.9	84.9	4.0	4.0	1.2	1.2
		$[240, 400]$	84.2	84.2	4.0	4.0	1.2	1.2
	1000	$[30, 50]$	98.2	98.3	4.2	4.2	1.2	1.2
		$[120, 200]$	92.6	92.6	4.0	4.0	1.2	1.2
		$[240, 400]$	91.9	91.9	4.0	4.0	1.2	1.2
2	50	$[30, 50]$	31.5	32.4	3.9	3.9	1.2	1.2
		$[120, 200]$	32.7	33.1	3.8	3.8	1.2	1.2
		$[240, 400]$	33.1	33.3	3.8	3.8	1.2	1.2
	500	$[30, 50]$	85.9	87.0	4.2	4.2	1.2	1.2
		$[120, 200]$	83.9	84.2	4.0	4.0	1.2	1.2
		$[240, 400]$	83.8	84.0	4.0	4.0	1.2	1.2
	1000	$[30, 50]$	96.3	97.1	4.2	4.2	1.2	1.2
		$[120, 200]$	92.3	92.5	4.0	4.0	1.2	1.2
		$[240, 400]$	91.5	91.7	4.0	4.0	1.2	1.2
3	50	$[30, 50]$	31.2	32.2	3.9	3.9	1.2	1.2
		$[120, 200]$	32.9	33.1	3.8	3.8	1.2	1.2
		$[240, 400]$	33.2	33.4	3.8	3.8	1.2	1.2
	500	$[30, 50]$	86.7	87.5	4.2	4.2	1.2	1.2
		$[120, 200]$	84.4	84.6	4.0	4.0	1.2	1.2
		$[240, 400]$	83.9	84.0	4.0	4.0	1.2	1.2
	1000	$[30, 50]$	96.8	97.4	4.2	4.2	1.2	1.2
		$[120, 200]$	92.4	92.5	4.0	4.0	1.2	1.2
		$[240, 400]$	91.7	91.8	4.0	4.0	1.2	1.2

Table 3 displays the empirical powers of the four tests. We can see that L2N and L2D have similar empirical powers, showing that they are comparable regardless whether the two samples (1.1) are normal or non-normal. This is consistent with what we observed from Table 2 that their empirical sizes are comparable. In addition, observe that the empirical powers of L2N and L2D are comparable to those of BS and CQ when the four tests have comparable empirical sizes (when $\rho = 0.1$), and L2N and L2D have slightly lower powers than BS and CQ when $\rho = 0.5$ or 0.9. However, since BS and CQ are generally more liberal and generally have worse size control than L2N and L2D (see Table 2), it is not surprising that they generally have slightly higher powers. Finally, we can see that under various settings the empirical powers of the four tests become smaller with the value of ρ increasing. This is reasonable because with ρ increasing the noise is increasing.

Table 4 displays the estimated approximate degrees of freedom of L2N and L2D under the considered settings. First of all, observe that the estimated approximate degrees of freedom of the two tests are about the same under each setting. This explains why the two tests are comparable in terms of both size accuracy and power as shown by tables 2 and 3. Secondly, it is seen that with the sample sizes increasing, the estimated approximate degrees of freedom of L2N and L2D become closer to their true approximate degrees of freedom listed in Table 1. This is consistent with the conclusion drawn from Theorem 5 which holds for both normal and non-normal data. Thirdly, notice that with the value of ρ increasing, the estimated approximate degrees of freedom of L2N and L2D become smaller. This shows that the normal approximation becomes less adequate as the correlation increases. Therefore, we expect that BS and CQ are less accurate when the correlation is larger, which can also be observed from Table 2.

TABLE 5
Simulation 2: Empirical sizes (in percentages).

			$\rho = 0.1$				$\rho = 0.5$				$\rho = 0.9$			
Model	p	$\boldsymbol{n}$	L2N	L2D	BS	CQ	L2N	L2D	BS	CQ	L2N	L2D	BS	CQ
1	50	$[30, 50]$	5.89	5.87	6.87	6.91	5.10	5.05	6.12	6.20	5.06	5.08	6.69	6.69
		$[120, 200]$	5.50	5.51	6.64	6.61	5.22	5.21	6.61	6.61	5.44	5.40	7.23	7.27
		$[240, 400]$	5.25	5.27	6.19	6.18	5.40	5.42	6.59	6.59	4.99	4.96	6.61	6.49
	500	$[30, 50]$	5.11	5.07	5.40	5.38	4.85	4.85	5.27	5.38	5.55	5.47	6.46	6.44
		$[120, 200]$	5.13	5.11	5.62	5.55	5.28	5.30	5.76	5.75	5.65	5.66	6.71	6.76
		$[240, 400]$	4.96	4.96	5.37	5.46	5.19	5.18	5.64	5.69	5.34	5.35	6.36	6.33
	1000	$[30, 50]$	5.33	5.35	5.56	5.51	5.49	5.46	5.76	5.87	5.55	5.53	6.14	6.28
		$[120, 200]$	4.97	4.98	5.29	5.31	5.34	5.36	5.74	5.75	5.38	5.39	6.26	6.26
		$[240, 400]$	4.83	4.82	5.15	5.17	5.18	5.18	5.57	5.58	5.10	5.11	5.96	6.01
2	50	$[30, 50]$	5.01	5.31	6.00	6.51	5.11	5.47	6.51	7.12	5.42	5.54	6.97	7.17
		$[120, 200]$	4.81	4.93	5.83	6.10	5.29	5.35	6.47	6.66	5.00	5.05	6.75	6.76
		$[240, 400]$	5.17	5.26	6.26	6.39	5.27	5.32	6.51	6.60	4.84	4.88	6.52	6.46
	500	$[30, 50]$	4.27	4.56	4.54	5.39	4.87	5.09	5.31	5.95	6.02	6.09	7.06	7.11
		$[120, 200]$	4.65	4.82	5.05	5.47	5.14	5.20	5.65	6.04	5.61	5.61	6.55	6.60
		$[240, 400]$	5.01	5.12	5.46	5.81	4.92	4.93	5.38	5.39	4.96	4.97	6.07	5.97
	1000	$[30, 50]$	4.24	4.52	4.40	5.28	4.85	5.00	5.14	6.06	5.45	5.53	6.21	6.43
		$[120, 200]$	4.84	4.99	5.06	5.41	4.82	4.88	5.06	5.42	5.16	5.17	5.95	6.03
		$[240, 400]$	5.05	5.10	5.31	5.46	5.23	5.26	5.62	5.75	5.74	5.75	6.45	6.50
3	50	$[30, 50]$	4.29	4.29	5.28	5.95	5.23	5.16	6.40	6.76	5.57	5.47	7.06	7.20
		$[120, 200]$	4.97	5.02	6.38	6.69	5.16	5.18	6.58	6.82	5.28	5.31	6.95	7.05
		$[240, 400]$	4.99	5.02	6.07	6.24	5.73	5.73	6.91	6.92	5.20	5.22	6.82	6.94
	500	$[30, 50]$	4.71	5.00	5.03	5.82	5.25	5.42	5.78	6.24	5.64	5.63	6.42	6.64
		$[120, 200]$	4.57	4.62	4.89	5.16	5.70	5.74	6.23	6.34	5.38	5.34	6.38	6.36
		$[240, 400]$	4.72	4.78	5.20	5.28	5.55	5.55	6.20	6.23	5.35	5.35	6.34	6.38
	1000	$[30, 50]$	4.31	4.60	4.51	5.31	4.78	5.00	5.12	5.61	5.51	5.47	6.16	6.24
		$[120, 200]$	5.16	5.22	5.50	5.80	4.88	4.90	5.28	5.45	5.83	5.86	6.76	6.81
		$[240, 400]$	4.81	4.86	5.14	5.34	5.32	5.35	5.70	5.80	5.24	5.24	5.87	5.83
ARE			5.68	4.81	12.09	15.18	5.34	5.50	17.71	21.91	7.91	8.01	30.16	31.12

4.2. *Simulation 2.* In this simulation study, we continue to use the setup of Simulation 1 except we now use a new covariance matrix $\mathbf{\Sigma}$ defined as

$$\mathbf{\Sigma} = \mathbf{DRD}, \ \mathbf{R} = (\rho^{|i-j|})_{i,j=1,2,\ldots,p},$$

where $\mathbf{D} = \text{diag}(\mathbf{h})$ and $\mathbf{h}$ is as in Simulation 1. Such a covariance matrix was also used by [31] and [42]. Note that the tuning parameter ρ here plays a somewhat different role from the one used in Simulation 1, but its value is also strongly related to the correlation of the simulated data. That is, when the value of ρ is small (large) the simulated high-dimensional data are less (highly) correlated. The empirical sizes of L2N, L2D, BS and CQ are displayed in Table 5. Notice that again L2N and L2D are comparable and they generally outperform BS and CQ in terms of size control.

Table 6
Simulation 3: Empirical sizes (in percentages).

Dependence			Partial				Full			
Model	p	$\boldsymbol{n}$	L2N	L2D	BS	CQ	L2N	L2D	BS	CQ
1	50	$[30, 50]$	5.37	5.31	6.47	6.51	5.82	5.81	7.59	7.55
		$[120, 200]$	5.04	5.04	6.18	6.33	4.53	4.54	6.31	6.18
		$[240, 400]$	5.29	5.31	6.49	6.56	5.23	5.23	7.20	7.21
	500	$[30, 50]$	5.32	5.38	5.76	5.70	5.45	5.46	7.20	7.23
		$[120, 200]$	5.71	5.73	6.05	6.07	5.15	5.18	6.77	6.91
		$[240, 400]$	5.43	5.41	5.87	5.85	5.08	5.09	6.82	6.87
	1000	$[30, 50]$	5.44	5.44	5.62	5.59	5.57	5.53	7.01	7.07
		$[120, 200]$	5.25	5.29	5.52	5.48	5.04	5.04	6.85	6.97
		$[240, 400]$	5.45	5.44	5.84	5.87	5.01	5.00	6.86	6.91
2	50	$[30, 50]$	5.16	5.15	6.39	6.42	5.44	5.40	7.14	7.20
		$[120, 200]$	5.14	5.14	6.21	6.23	5.27	5.28	7.06	7.14
		$[240, 400]$	4.85	4.87	6.14	6.11	5.15	5.16	7.11	7.16
	500	$[30, 50]$	5.13	5.15	5.49	5.53	5.69	5.71	7.55	7.56
		$[120, 200]$	5.18	5.19	5.57	5.65	4.76	4.75	6.59	6.61
		$[240, 400]$	5.02	5.03	5.56	5.57	5.06	5.06	7.00	7.00
	1000	$[30, 50]$	5.27	5.27	5.51	5.61	5.64	5.71	7.17	7.37
		$[120, 200]$	5.38	5.36	5.65	5.63	4.89	4.92	6.72	6.76
		$[240, 400]$	5.21	5.22	5.64	5.58	5.37	5.37	7.19	7.23
ARE			5.49	5.54	17.73	18.10	6.43	6.47	40.16	41.03

4.3. *Simulation 3.* In this simulation study, we consider the moving average model considered by [5] in which the t-th component of $\mathbf{y}_{ij} = [y_{ij1}, y_{ij2}, \ldots, y_{ijp}]^\top$ is generated in the following way:

$$y_{ijt} = \mu_{it} + \rho_1 z_{ijt} + \rho_2 z_{ij(t+1)} + \cdots + \rho_p z_{ij(t+p-1)},$$

with μ_{it} being the t-th component of $\boldsymbol{\mu}_i, i = 1, 2$, and z_{ijt}, $t = 1, 2, \ldots, p$, $j = 1, 2, \ldots, n_i$, $i = 1, 2$, being i.i.d. random variables. The following two models are used for the innovations $\{z_{ijt}\}$:

Model 1. $z_{ijt}, t = 1, 2, \ldots, p, \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$.

Model 2. $z_{ijt} = (w_{ijt} - 4)/2$ with $w_{ijt}, t = 1, 2, \ldots, p, \overset{i.i.d.}{\sim}$ Gamma$(4, 1)$.

For the dependence between the components of $\mathbf{y}_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2$, we consider the following two cases:

1. Partial dependence: $\rho_1 = 2.883$, $\rho_2 = 2.794$, $\rho_3 = 2.849$, $\rho_t = 0$ for $t = 4, \ldots, p$, and are kept fixed throughout the simulation study.
2. Full dependence: The ρ_t's are all generated from $U[2, 3]$ and kept fixed.

Note that the simulated two high-dimensional samples are less (highly) correlated in the partial (full) dependence case. As in Simulations 1 and 2, the number of simulation runs is 10000. The resulting empirical sizes of L2N, L2D, BS, and CQ are displayed in Table 6. In terms of size control, again L2N and L2D are comparable and they generally outperform BS and CQ.

Table 7
Simulation 4: Empirical sizes (in percentages).

ρ	p	$\boldsymbol{n}$	L2N	L2D	SK	CLX	GCBL$_m$	GCBL$_l$
0.1	50	$[30, 50]$	5.3	5.4	3.7	33.3	9.6	93.5
		$[120, 200]$	5.6	5.6	3.3	6.9	9.1	14.2
		$[240, 400]$	5.2	5.2	3.7	3.5	9.4	10.5
	500	$[30, 50]$	6.5	6.5	4.1	21.7	7.8	100.0
		$[120, 200]$	5.7	5.7	4.8	13.2	5.7	100.0
		$[240, 400]$	4.0	4.0	4.2	9.4	5.9	100.0
	1000	$[30, 50]$	4.5	4.7	2.6	14.1	9.7	100.0
		$[120, 200]$	6.0	6.0	3.2	15.5	5.3	100.0
		$[240, 400]$	6.2	6.2	5.0	9.3	6.0	100.0
0.5	50	$[30, 50]$	5.7	5.7	3.4	28.7	13.6	83.8
		$[120, 200]$	6.5	6.5	2.6	6.4	14.3	18.8
		$[240, 400]$	5.3	5.3	2.9	3.9	12.7	13.3
	500	$[30, 50]$	5.0	5.1	2.3	17.0	6.1	100.0
		$[120, 200]$	5.9	5.9	3.1	10.2	4.7	100.0
		$[240, 400]$	5.2	5.2	3.8	6.4	5.8	100.0
	1000	$[30, 50]$	4.6	4.6	2.6	13.8	7.8	100.0
		$[120, 200]$	4.2	4.2	2.5	11.6	4.4	100.0
		$[240, 400]$	6.1	6.1	4.3	8.2	6.0	100.0
0.9	50	$[30, 50]$	4.9	4.9	1.5	51.2	39.2	74.9
		$[120, 200]$	4.4	4.4	1.4	23.3	39.6	41.6
		$[240, 400]$	6.2	6.2	2.0	17.7	38.8	39.2
	500	$[30, 50]$	5.8	5.9	2.0	80.2	14.9	100.0
		$[120, 200]$	6.4	6.5	2.6	76.9	15.9	100.0
		$[240, 400]$	6.4	6.5	2.3	71.8	14.2	99.7
	1000	$[30, 50]$	5.6	5.3	1.6	71.7	10.1	100.0
		$[120, 200]$	6.5	6.5	2.5	89.7	11.5	100.0
		$[240, 400]$	5.3	5.3	2.1	83.9	11.2	100.0
ARE			15.41	15.41	40.67	496.07	152.67	1521.85

4.4. *Simulation 4.* In this simulation study, we compare L2N and L2D against the tests proposed by Srivastava and Kubokawa [32], Cai et al. [3] and Gregory et al. [22], denoted as SK, CLX and GCBL, respectively. The data are generated as in Simulation 2 but we focus on normal data only and the number of simulation runs is reduced to 1000 as CLX is very time-consuming to compute. Since we assume the two samples (1.1) have the same covariance matrix, we use the equal covariances version of CLX only. We consider two variants of GCBL: moderate-p GCBL and large-p GCBL, denoted as GCBL_m and GCBL_l, respectively. Following [22], we choose "Parzen" window with lag window size $2\sqrt{p}/3$ for both GCBL_m and GCBL_l.

The empirical sizes of the considered tests are given in Table 7. It is seen that L2N and L2D are again the winners in terms of size control as indicated by the ARE values of the six tests. SK has very conservative empirical sizes, especially when the correlation parameter ρ is large, showing that SK may yield misleading results when the data are moderately or highly correlated. On the other hand, CLX, GCBL_m, and GCBL_l generally have very liberal empirical sizes especially when ρ is large or the sample sizes are small, showing that they may also yield misleading results in such cases.

TABLE 8
Simulation 5: Empirical sizes (in percentages).

ρ	p	L2N	L2D	SD	SKK	FZWZ
	50	5.39	5.40	5.13	6.26	6.06
0.1	500	6.53	6.50	5.39	6.46	6.96
	1000	6.79	6.73	4.98	5.99	6.86
	50	5.91	5.90	3.02	3.36	7.07
0.5	500	5.86	5.87	1.16	1.26	7.04
	1000	5.75	5.73	0.65	0.83	6.79
	50	5.70	5.67	1.10	1.21	6.82
0.9	500	5.79	5.80	0.16	0.17	6.84
	1000	5.26	5.33	0.01	0.04	6.47
ARE		17.73	17.62	54.31	59.64	35.36

4.5. *Simulation 5.* In this simulation study, we compare L2N and L2D against the tests proposed by Srivastava and Du [30], Srivastava et al. [31], and Feng et al. [20], denoted as SD, SKK, and FZWZ, respectively. The data are generated as in Simulation 1 but we only consider the normal data case with $[n_1, n_2] = [30, 50]$ because FZWZ is computationally very intensive.

The associated empirical sizes are displayed in Table 8. Observe that L2N and L2D are again the winners in terms of size control. Note also that SD and SKK are too conservative when the data are moderately or highly correlated (when $\rho = 0.5, 0.9$) although they are fine when the data are less correlated (when $\rho = 0.1$). The empirical sizes of FZWZ are generally around 7% which are much more liberal than those of our tests.

5. Applications to the colon data. In this section we apply L2N, L2D, and BS (we do not include other competitors mentioned in Section 4 to save space) to the colon data set which is publicly available at *http://microarray.princeton.edu/oncology/affydata/index.html*. The aim is to test if the mean vector of 2000 gene expressions of the normal colon tissues ($n_1 = 22$) is significantly different from that of the tumor colon tissues ($n_2 = 40$). Table 9 shows the results of L2N, L2D, and BS on this data set. Observe that L2N and L2D have similar p-values and similar estimated parameters. This is consistent with what we observed from the simulation studies presented in Section 4. Although the three tests all suggest a strong rejection of the null hypothesis, the p-value of BS is less reliable. This is because both the estimated approximate degrees of freedom of L2N and L2D are less than 7, indicating that the normal approximation used by BS is not adequate. Therefore, BS is not recommended at least in this example.

TABLE 9
Two-sample testing for the colon data.

Method	Statistic	P-value	$\hat{\beta}$	$\hat{d}$
L2N	1.34×10^9	6.26×10^{-4}	5.47×10^7	6.5
L2D	1.34×10^9	9.83×10^{-4}	5.80×10^7	6.3
BS	4.94	4.00×10^{-7}	-	-

6. Concluding remarks.

We propose and study an L^2-norm based test with the two-cumulant matched χ^2-approximation for two-sample high-dimensional problems where the dimension can be much larger than the total sample size. Unlike existing modifications of the classical Hotelling's T^2-test, the proposed test is adaptive to the shape of the null distribution of the test statistic and hence has a good size control in a wide range of situations. The key to this success is that the χ^2-approximation we employ does not require any structural assumptions on the covariance matrix, nor it assumes restrictive conditions on the covariance such as (2.11)–(2.13) by [1], [5] and [30], which rarely hold in high dimensions.

Besides the new theoretical insights into the advantages of the proposed L2N and L2D tests, the five simulation studies presented in Section 4 demonstrate that they are comparable in terms of size and power regardless whether the distributions of the two samples are normal or non-normal. In terms of size control, they are always the winners against the nine considered existing tests: BS, CQ, SK, CLX, GCBL$_m$, GCBL$_l$, SD, SKK, and FZWZ. In particular, the competitors are either too conservative or too liberal when the data are moderately or highly correlated while our tests L2N and L2D perform well under various simulation settings and various covariance structures. In addition, our tests are easy to understand, simple to implement and fast to compute. We refer to the Supplement [45] for computational speed comparisons. Therefore, they have great potentials in applications to real world problems, in particular in the new era of big data. Although we provide different formulae for estimation of the parameters in the χ^2-approximation depending on whether the data are normal or non-normal, we show both theoretically and numerically they are always comparable in all the considered simulation configurations. Therefore, we suggest to simply use the version for normal data in practice. Finally, we mention that the ideas of the proposed test can be extended to other high-dimensional testing problems, including testing equality of means under unequal covariance matrices, testing equality of covariance matrices [4, 46], testing regression means, and so on. Further studies in such directions are interesting and are ongoing.

APPENDIX: Technical proofs

Proof of Theorem 1. By (2.1) and (2.4), we have $\sqrt{n_1 n_2/n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \overset{d}{=} \sum_{r=1}^p \lambda_r^{1/2}\zeta_r \mathbf{u}_r$ where $\zeta_r, r = 1, 2, \ldots, p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. It follows that $T_n \overset{d}{=} \sum_{r=1}^p \lambda_r A_r$ where $A_r = \zeta_r^2 \sim \chi_1^2, r = 1, 2, \ldots, p$. The expressions of the first three cumulants $\mathcal{K}_1 = \mathrm{E}(T_{n0}), \mathcal{K}_2 = \mathrm{Var}(T_{n0})$ and $\mathcal{K}_3 = \mathrm{E}[T_{n0} - \mathrm{E}(T_{n0})]^3$ then follow immediately from Eq. (4) of [43]. We now prove the case when $p \to \infty$. Without loss of generality, we assume $\mathrm{tr}(\mathbf{\Sigma}) = 1$ for all p since otherwise we work with $T_{n0}/\mathrm{tr}(\mathbf{\Sigma})$. By the Karhunen–Loève expansion [35], we have $\lim_{p\to\infty} \sqrt{n_1 n_2/n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \overset{d}{=} \sum_{r=1}^\infty \lambda_r^{1/2}\zeta_r \mathbf{u}_r$ where $\zeta_r, r = 1, 2, \ldots, \infty \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Thus as $p \to \infty$, we have $T_{n0} \xrightarrow{L} T_0$ where $T_0 \overset{d}{=} \sum_{r=1}^\infty \lambda_r A_r$ with $\lambda_1, \lambda_2, \ldots$ being the eigenvalues of $\lim_{p\to\infty} \mathbf{\Sigma}$. □

Proof of Theorem 2. By (2.5), T_{n0} is a central χ^2-type mixture with all of the coefficients nonnegative. The theorem then follows from Lemma S.1 of [45] immediately with d^* defined in (2.8). □

Proof of Corollary 1. Since all the eigenvalues of $\mathbf{\Sigma}$ are nonnegative, $\mathrm{tr}(\mathbf{\Sigma}^3) = \sum_{r=1}^p \lambda_r^3 \le \lambda_{\max} \sum_{r=1}^p \lambda_r^2 = \lambda_{\max}\mathrm{tr}(\mathbf{\Sigma}^2)$ and hence

$$d^* = \mathrm{tr}^3(\mathbf{\Sigma}^2)/\mathrm{tr}^2(\mathbf{\Sigma}^3) \ge [\lambda_{\max}^2/\mathrm{tr}(\mathbf{\Sigma}^2)]^{-1} \tag{A.1}$$

as desired. Next, by the Cauchy–Schwarz inequality, we have $\mathrm{tr}^2\left(\mathbf{\Sigma}^3\right) \le \mathrm{tr}\left(\mathbf{\Sigma}^4\right)\mathrm{tr}\left(\mathbf{\Sigma}^2\right)$ and hence $d^* \ge \left[\mathrm{tr}(\mathbf{\Sigma}^4)/\mathrm{tr}^2(\mathbf{\Sigma}^2)\right]^{-1}$ as desired. Finally, we can write $d^* = p[\mathrm{tr}(\mathbf{\Sigma}^2)/p]^3/[\mathrm{tr}\left(\mathbf{\Sigma}^3\right)/p]^2$. Thus, under any of the conditions (2.11), (2.12) and (2.13), we have $d^* \to \infty$ as $p \to \infty$. Then, by Theorem 2, T_{n0} is asymptotically normal as $p \to \infty$. □

Proof of Corollary 2. Under the given conditions, as $p \to \infty$, we have $d^* = [\mathrm{tr}(\mathbf{\Sigma}^2)/p^2]^3/[\mathrm{tr}(\mathbf{\Sigma}^3)/p^3]^2 \to b_2^3/b_3^2 \in (0,\infty)$. The corollary then follows directly from Theorem 2. □

Proof of Theorem 3. Since $\lambda_r, r = 1, 2, \ldots, p$, are all nonnegative,

$$\mathrm{tr}^2(\mathbf{\Sigma}^k) = \big(\sum_{r=1}^p \lambda_r^k\big)^2 \ge \sum_{r=1}^p \lambda_r^{2k} = \mathrm{tr}(\mathbf{\Sigma}^{2k}), k = 1, 2, \ldots. \tag{A.2}$$

Together with (2.14) this shows that $d^* \geq \mathrm{tr}^2(\mathbf{\Sigma}^2)/\mathrm{tr}(\mathbf{\Sigma}^4) \geq 1$. By the Cauchy–Schwarz inequality, we have $\mathrm{tr}^2\left(\mathbf{\Sigma}^2\right) \leq \mathrm{tr}\left(\mathbf{\Sigma}^3\right)\mathrm{tr}\left(\mathbf{\Sigma}\right)$. Thus,

$$d^* = \mathrm{tr}^3\left(\mathbf{\Sigma}^2\right)/\mathrm{tr}^2\left(\mathbf{\Sigma}^3\right) \leq \mathrm{tr}^3\left(\mathbf{\Sigma}^2\right)/\left[\mathrm{tr}^4\left(\mathbf{\Sigma}^2\right)/\mathrm{tr}^2\left(\mathbf{\Sigma}\right)\right] = d. \tag{A.3}$$

To show the last inequality of (a), note that $[\mathrm{tr}(\mathbf{\Sigma})/p]^2 = (\sum_{r=1}^p \lambda_r/p)^2 \leq (\sum_{r=1}^p \lambda_r^2/p) = \mathrm{tr}(\mathbf{\Sigma}^2)/p$. We then have $d \leq p$. Assertion (a) is now proved.

To show assertion (b), note that when only λ_1 is nonzero we have $\mathrm{tr}(\mathbf{\Sigma}^k) = \lambda_1^k, k = 1, 2, \ldots$, so that we have $d^* = d = 1$. Conversely, when $d = 1$ we have $(\sum_{r=1}^p \lambda_r)^2 = \sum_{r=1}^p \lambda_r^2$ and thus $\sum_{r\neq s} \lambda_r\lambda_s = 0$. Since $\lambda_1, \lambda_2, \ldots, \lambda_p$ are nonnegative and in descending order, we have $\lambda_r\lambda_s = 0$ for all $r \neq s$. However, $\mathbf{\Sigma} = \mathbf{0}$ if all of the eigenvalues of $\mathbf{\Sigma}$ are zero. This case does not make any sense in practice. So let us assume that λ_1 is positive. Then we have $\lambda_r = 0, r = 2, \ldots, p$. Assertion (b) is then proved.

We now show assertion (c). When all the eigenvalues of $\mathbf{\Sigma}$ are the same, we have $\mathbf{\Sigma} = \lambda_1\mathbf{I}_p$. Hence $\mathrm{tr}(\mathbf{\Sigma}^k) = p\lambda_1^k, k = 1, 2, \ldots$, so we have $d^* = d = p$. Conversely, when $d^* = d = p$ we have $\mathrm{tr}^2(\mathbf{\Sigma})/\mathrm{tr}(\mathbf{\Sigma}^2) = p$ and it follows that $(\sum_{r=1}^p \lambda_r/p)^2 = \sum_{r=1}^p \lambda_r^2/p$ which implies $\lambda_1 = \cdots = \lambda_p$. □

Proof of Corollary 3. By (a) of Theorem 3, when d is bounded, so is d^* so that $R = \beta\chi_d^2$ is not asymptotically normal. By Theorem 2, T_{n0} is also not asymptotically normal. On the other hand, when $d^* \to \infty$, we have $d \to \infty$ so that $[R - \mathrm{tr}(\mathbf{\Sigma})]/[2\mathrm{tr}(\mathbf{\Sigma}^2)]^{1/2} = (\chi_d^2 - d)/(2d)^{1/2} \xrightarrow{L} \mathcal{N}(0, 1)$. By Theorem 2, we also have $[T_{n0} - \mathrm{tr}(\mathbf{\Sigma})]/[2\mathrm{tr}(\mathbf{\Sigma}^2)]^{1/2} \xrightarrow{L} \mathcal{N}(0, 1)$. □

Proof of Theorem 4. Theorem 4 is a special case of Lemma S.2 of [45] when $v = 1$ and $\lambda_1, \ldots, \lambda_p$ are taken as the eigenvalues of $\mathbf{\Sigma}$. □

Proof of Theorem 5. When the two samples (1.1) are normal, by (2.17) and Lemma S.3 of [45], as $n \to \infty$, we have uniformly for all p,

$$\frac{\hat{\beta}}{\beta} = \frac{\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)}/\mathrm{tr}(\mathbf{\Sigma}^2)}{\mathrm{tr}(\hat{\mathbf{\Sigma}})/\mathrm{tr}(\mathbf{\Sigma})} \xrightarrow{P} 1, \quad \frac{\hat{d}}{d} = \frac{\widehat{\mathrm{tr}^2\left(\mathbf{\Sigma}\right)}/\mathrm{tr}^2(\mathbf{\Sigma})}{\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)}/\mathrm{tr}(\mathbf{\Sigma}^2)} \xrightarrow{P} 1.$$

It follows that $\hat{\beta}\chi_{\hat{d}}^2(\alpha)$ is ratio-consistent for $\beta\chi_d^2(\alpha)$ uniformly for all p. □

Proof of Corollary 4. By Lemma S.3 of [45], the proof is obvious since $\mathrm{tr}(\hat{\mathbf{\Sigma}})$ and $\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)}$ are ratio-consistent of $\mathrm{tr}(\mathbf{\Sigma})$ and $\mathrm{tr}(\mathbf{\Sigma}^2)$, respectively. □

Proof of Theorem 6. Under the given conditions and the local alternative (2.19), by (2.2) and Theorem 5, we have

$$\begin{aligned}\Pr\left[T_n/\hat{\beta} \geq \chi_{\hat{d}}^2(\alpha)\right] &\approx \Pr\left[\tfrac{T_{n0}}{\beta} \geq \tfrac{\hat{\beta}\chi_{\hat{d}}^2(\alpha)}{\beta\chi_d^2(\alpha)}\chi_d^2(\alpha) - \tfrac{n\tau(1-\tau)\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|^2}{\beta}\right] \\ &\approx \Pr\left[\chi_d^2 \geq \chi_d^2(\alpha) - n\tau(1-\tau)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/\beta\right],\end{aligned}$$

as desired. Notice that when the conditions of Corollary 2 are satisfied $\mathrm{tr}(\boldsymbol{\Sigma}^r)/p^r \to b_r \in (0,\infty), r = 1,2,3$. Therefore, when $p/n \to c \in (0,\infty)$ as $n, p \to \infty$, we have $\beta/n \to cb_2/b_1 \in (0,\infty)$ and $d \to b_1^2/b_2 \in (0,\infty)$. □

Proof of Theorem 7. Under the given conditions and the condition (2.19), by (2.2) and Theorem 2, we have

$$\begin{aligned}\Pr\left\{\frac{T_n - \mathrm{tr}(\hat{\boldsymbol{\Sigma}})}{\left[2\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right]^{1/2}} \geq z_\alpha\right\} &= \Pr\left\{\frac{T_{n0} - \mathrm{tr}(\hat{\boldsymbol{\Sigma}})}{\left[2\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right]^{1/2}} \geq z_\alpha - \frac{n\tau(1-\tau)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\left[2\widehat{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right]^{1/2}}\right\} + o(1) \\ &= \Phi\left\{-z_\alpha + \frac{n\tau(1-\tau)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\left[2\mathrm{tr}(\boldsymbol{\Sigma}^2)\right]^{1/2}}\right\} + o(1). \qquad \Box\end{aligned}$$

Proof of Theorem 8. Set $\mathbf{w}_n = \sqrt{n_1 n_2/n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. For any fixed finite p, by the central limit theorem, as $n \to \infty$ with $n_1/n \to \tau \in (0,1)$, we have $\mathbf{w}_n \xrightarrow{L} \mathbf{w}$ where $\mathbf{w} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. By the continuous mapping theorem, we have $T_{n0} = \|\mathbf{w}_n\|^2 \xrightarrow{L} T_0$ where $T_0 = \|\mathbf{w}\|^2 \overset{d}{=} \sum_{r=1}^p \lambda_r A_r$.

We now prove the case when $n, p \to \infty$ via the characteristic function $[\psi_X(t) = \mathrm{E}(e^{itX})]$ method. Without loss of generality, we assume that $\mathrm{tr}(\boldsymbol{\Sigma}) = 1$ for all p since otherwise we can consider $T_{n0}/\mathrm{tr}(\boldsymbol{\Sigma})$. Recall that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p$ are the eigenvectors associated with the decreasing-ordered eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $\boldsymbol{\Sigma}$. We have $\mathbf{w}_n = \sum_{r=1}^p \xi_{n,r}\mathbf{u}_r$ where $\xi_{n,r} = \mathbf{w}_n^\top \mathbf{u}_r$. It is known that $\xi_{n,r}, r = 1,2,\ldots,p$ are uncorrelated and $\mathrm{E}(\xi_{n,r}) = 0$ and $\mathrm{Var}(\xi_{n,r}) = \lambda_r, r = 1,2,\ldots$. Further, by the proofs of (3.1) and (3.2) given in the Supplement [45], we have

$$\mathrm{E}(\xi_{n,r}^4) = 3\lambda_r^2 + \left(\frac{n_2}{n}\right)^2 \frac{\mathrm{E}(\mathbf{x}_{11}^\top \mathbf{u}_r)^4 - 3\lambda_r^2}{n_1} + \left(\frac{n_1}{n}\right)^2 \frac{\mathrm{E}(\mathbf{x}_{21}^\top \mathbf{u}_r)^4 - 3\lambda_r^2}{n_2},$$

where $\mathbf{x}_{i1} = \mathbf{y}_{i1} - \boldsymbol{\mu}_i, i = 1,2$ as defined before. Under Assumptions A1, A2 and A3, by some simple algebra, we have $\mathrm{E}(\mathbf{x}_{i1}^\top \mathbf{u}_r)^4 \leq (3+\Delta)\lambda_r^2, i = 1,2$. Thus, we have

$$\mathrm{E}(\xi_{n,r}^4) \leq 3\left[1 + \left(\frac{n_2}{n}\right)^2 \frac{\Delta}{3n_1} + \left(\frac{n_1}{n}\right)^2 \frac{\Delta}{3n_2}\right]\lambda_r^2, r = 1,2,\ldots.$$

Note that $T_{n0} = \sum_{r=1}^p \xi_{n,r}^2$. Set $T_{n0,q} = \sum_{r=1}^q \xi_{n,r}^2$. Then we have

$$|\psi_{T_{n0}}(t) - \psi_{T_{n0,q}}(t)| \leq |t|\left[\mathrm{E}(T_{n0} - T_{n0,q})^2\right]^{1/2} \leq |t|\left[2\sum_{r=q+1}^p \mathrm{E}(\xi_{n,r}^4)\right]^{1/2},$$

which is valid for all large p. As $p \to \infty$, we have $T_{n0} = \sum_{r=1}^\infty \xi_{n,r}^2$ and the above result still holds with the upper bound $|t|[2\sum_{r=q+1}^p \mathrm{E}(\xi_{n,r}^4)]^{1/2}$

replaced by $|t|[2\sum_{r=q+1}^{\infty}\mathrm{E}(\xi_{n,r}^4)]^{1/2}$ and with λ_r's being the eigenvalues of $\lim\limits_{p\to\infty}\mathbf{\Sigma}$. Let t be fixed. Since $\sum_{r=1}^{\infty}\lambda_r^2 \le (\sum_{r=1}^{\infty}\lambda_r)^2 = 1$, for any $\epsilon > 0$, there exist Q and N_1, both depending on t and ϵ, such that as $n_1, n_2 \ge N_1$, we have $|\ \ _{T_{n0}}(t) - \psi_{T_{n0,Q}}(t)| \le |t|\big\{6[1+o(1)]\sum_{r=Q+1}^{\infty}\lambda_r^2\big\}^{1/2} \le \epsilon$. For the fixed Q, by the central limit theorem we have $T_{n0,Q} \xrightarrow{L} T_{0,Q}$ where $T_{0,Q} \overset{d}{=} \sum_{r=1}^{Q}\lambda_r A_r$ since as $n \to \infty$, $\xi_{n,r} \xrightarrow{L} \mathcal{N}(0, \lambda_r)$ and $\xi_{n,r}$'s are asymptotically independent. That is, there exists N_2, depending on t and ϵ, such that as $n_1, n_2 > N_2$, we have $|\psi_{T_{n0,Q}}(t) - \psi_{T_{0,Q}}(t)| \le \epsilon$. Note that $T_0 = \sum_{r=1}^{\infty}\lambda_r A_r$. We have

$$|\psi_{T_{0,Q}}(t) - \psi_{T_0}(t)| \le |t|\Big[2\sum_{r=Q+1}^{\infty}\lambda_r^2\mathrm{E}(A_r^2)\Big]^{1/2} \le |t|\Big[6\sum_{r=Q+1}^{\infty}\lambda_r^2\Big]^{1/2} \le \epsilon.$$

It follows that as $n_1, n_2 \ge \min(N_1, N_2)$, we have $|\psi_{T_{n0}}(t) - \psi_{T_0}(t)| \le 3\epsilon$. The theorem follows as we can let $\epsilon \to 0$. □

Proof of Theorem 9. Under the BS Assumptions A1 and A2, by some simple algebra, we have $\kappa_{i,11} = \Delta\sum_{r=1}^{m}\gamma_r^2, i = 1, 2$ where γ_r is the r-th diagonal entry of $\mathbf{\Gamma}^\top\mathbf{\Gamma}$. It follows that $\kappa_{i,11} = O[\mathrm{tr}(\mathbf{\Sigma}^2)], i = 1, 2$ since $\Delta < \infty$ and $\sum_{r=1}^{m}\gamma_r^2 \le \mathrm{tr}(\mathbf{\Sigma}^2)$. Therefore, as $n, p \to \infty$, we have $\delta = o[\mathrm{tr}(\mathbf{\Sigma}^2)]$ and

$$d = \mathrm{tr}^2(\mathbf{\Sigma})/[\mathrm{tr}(\mathbf{\Sigma}^2) + \delta/2] = \mathrm{tr}^2(\mathbf{\Sigma})\mathrm{tr}(\mathbf{\Sigma}^2)^{-1}[1 + o(1)]. \tag{A.4}$$

By (A.3) and Corollary 1, we have $\mathrm{tr}^2(\mathbf{\Sigma})/\mathrm{tr}(\mathbf{\Sigma}^2) \ge \mathrm{tr}^3(\mathbf{\Sigma}^2)/\mathrm{tr}^2(\mathbf{\Sigma}^3) \ge \mathrm{tr}(\mathbf{\Sigma}^2)/\lambda_{\max}^2$. Therefore, $d \to \infty$ and R is asymptotically normal as $n, p \to \infty$. We refer to [1] for the proof of asymptotic normality of T_{n0}. □

Proof of Corollary 5. Under the BS assumptions and H_0, by Theorem 9 and ratio-consistency of $\mathrm{tr}(\hat{\mathbf{\Sigma}})$ and $\widehat{\mathrm{tr}\left(\mathbf{\Sigma}^2\right)}$, (2.18) follows immediately. □

Proof of Theorem 10. The proof is about the same lines as those in the proof of Theorem 6 except now we need to use the assumption $\kappa_{i,11} = O[\mathrm{tr}(\mathbf{\Sigma}^2)]$ to prove that expression (A.4) holds. The details are then omitted.

Proof of Theorem 11. The proof is similar to that of Theorem 7.

SUPPLEMENTARY MATERIAL

Supplement A: Additional examples and some lemmas
(doi: xx.xxxx/xx-xxxx). Histograms of simulated T_{BS}, and some lemmas and their proofs.

REFERENCES

[1] Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica*, 6(2):311–329.

[2] Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Stat.*, 25(2):290–302.

[3] Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(2):349–372.

[4] Chen, S., Zhang, L., and Zhong, P. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.*, 105:810–819.

[5] Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, 38(2):808–835.

[6] Cheng, M.-Y. and Wu, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc.*, 108:1421–1434.

[7] Chuang, L.-L. and Shih, Y.-S. (2012). Approximated distributions of the weighted sum of correlated chi-squared random variables. *J. Statist. Plann. Inference*, 142(2):457–472.

[8] Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.*, 30:455–474.

[9] Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Amer. Statist. Assoc.*, 110:630–641.

[10] Dempster, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Stat.*, 29(4):995–1010.

[11] Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, 16:41–50.

[12] Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Ann. Statist.*, 36(6):2605–2637.

[13] Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(4):745–771.

[14] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360.

[15] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70:849–911.

[16] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, 32(3):928–961.

[17] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38:3567–3604.

[18] Fan, Y., Kong, Y., Li, D., and Zheng, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist.*, 43:1243–1272.

[19] Feiveson, A. H. and Delaney, F. C. (1968). *The distribution and properties of a weighted sum of chi squares*, volume 4575. National Aeronautics and Space Administration.

[20] Feng, L., Zou, C., Wang, Z., and Zhu, L. (2015). Two sample Behrens–Fisher problem for high-dimensional data. *Statist. Sinica*, 25:1297–1312.

[21] Gregory, K. B. (2014). *highD2pop: Two-Sample Tests for Equality of Means in High Dimension.* R package version 1.0.

[22] Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015). A two-sample test for equality of means in high dimension. *J. Amer. Statist. Assoc.*, 110(510):837–849.

[23] Himeno, T. and Yamada, T. (2014). Estimations for some functions of covariance matrix in high dimension under non-normality and its applications. *J. Multivariate Anal.*, 130:27–44.

[24] Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Stat.*,

2(3):360–378.

[25] Li, G., Peng, H., Zhang, J., and Zhu, L. (2012). Robust rank correlation based screening. *Ann. Statist.*, 40:1846–1877.

[26] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, 37:3498–3528.

[27] Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6(5):309–316.

[28] Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biom. Bull.*, 2(6):110–114.

[29] Schott, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *J. Multivariate Anal.*, 98(9):1825–1839.

[30] Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.*, 99(3):386–402.

[31] Srivastava, M. S., Katayama, S., and Kano, Y. (2013). A two sample test in high dimensional data. *J. Multivariate Anal.*, 114:349–358.

[32] Srivastava, M. S. and Kubokawa, T. (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *J. Multivariate Anal.*, 115:204–216.

[33] Srivastava, R., Li, P., and Ruppert, D. (2015). Raptt: An exact two-sample test in high dimensions using random projections. *J. Comput. Graph. Statist.*, 25(3):954–970.

[34] Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58:267–288.

[35] Wahba, G. (1990). *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, SIAM, Philadelphia.

[36] Wang, L., Peng, B., and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.*, 110(512):1658–1669.

[37] Wang, R., Peng, L., and Y., Q. (2013). Jackknife empirical likelihood test for equality of two high dimensional means. *Statist. Sinica*, 23:667–690.

[38] Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

[39] Witten, D. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73:753–772.

[40] Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.*, 35(6):2654–2690.

[41] Xia, Y. (2008). A multiple-index model and dimension reduction. *J. Amer. Statist. Assoc.*, 103:1631–1640.

[42] Yamada, T. and Himeno, T. (2015). Testing homogeneity of mean vectors under heteroscedasticity in high-dimension. *J. Multivariate Anal.*, 139:7–27.

[43] Zhang, J.-T. (2005). Approximate and asymptotic distributions of chi-squared-type mixtures with applications. *J. Amer. Statist. Assoc.*, 100(469):273–285.

[44] Zhang, J.-T. (2013). *Analysis of variance for functional data*. CRC Press.

[45] Zhang, J.-T., Guo, J., Zhou, B., and Cheng, M.-Y. (2017). Supplement to "a simple and adaptive two-sample test in high dimensions based on l^2 norm. *Ann. Statist.*

[46] Zhang, R., Peng, L., and Wang, R. (2013). Tests for covariance matrix with fixed or divergent dimension. *Ann. Statist.*, 41:2075–2096.

J.-T. Zhang, J. Guo, and B. Zhou
Department of Statistics & Applied Probability
National University of Singapore
3 Science Drive 2
Singapore 117546
E-mail: stazjt@nus.edu.sg; jia.guo@u.nus.edu; bu.zhou@u.nus.edu

M.-Y. Cheng
Department of Mathematics
National Taiwan University
Taipei 106, Taiwan
E-mail: cheng@math.ntu.edu.tw