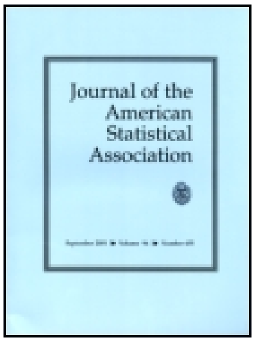

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

A Simple Two-Sample Test in High Dimensions Based on L^2-Norm

Jin-Ting Zhang, Jia Guo, Bu Zhou & Ming-Yen Cheng

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–17: Theory and Methods
https://doi.org/10.1080/01621459.2019.1604366

Taylor & Francis
Taylor & Francis Group

A Simple Two-Sample Test in High Dimensions Based on L^2-Norm

Jin-Ting Zhang[a], Jia Guo[b], Bu Zhou[c], and Ming-Yen Cheng[d]

[a]Department of Statistics and Applied Probability, National University of Singapore, Singapore; [b]School of Management, Zhejiang University of Technology, Hangzhou, China; [c]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China; [d]Department of Mathematics, Hong Kong Baptist University, Hong Kong

ABSTRACT

Testing the equality of two means is a fundamental inference problem. For high-dimensional data, the Hotelling's T^2-test either performs poorly or becomes inapplicable. Several modifications have been proposed to address this issue. 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. We study this problem thoroughly and propose an L^2-norm based test that works under mild conditions and even when there are fewer observations than the dimension. Specially, to cope with general nonnormality 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 justify that χ^2-approximation is preferred to normal approximation. Simple ratio-consistent estimators for the parameters in the χ^2-approximation are given. Importantly, our test can cope with singularity or near singularity of the covariance which is commonly seen in high dimensions and is the main cause of nonnormality. The power of the proposed test is also investigated. Extensive simulation studies and an application show that our test is at least comparable to and often outperforms several competitors in terms of size control, and the powers are comparable when their sizes are. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2018
Revised February 2019

KEYWORDS
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 (Cheng and Wu 2013; Cook and Li 2002; Fan and Li 2001; Fan and Lv 2008; Fan and Peng 2004; Fan and Song 2010; Lv and Fan 2009; Li et al. 2012; Tibshirani 1996; Xia 2007, 2008). And, these ideas have been extended to other problems such as classification and discriminant analysis (Cui, Li, and Zhong 2015; Fan and Fan 2008; Fan, Feng, and Tong 2012; Fan et al. 2015; Witten and Tibshirani 2011).

We study testing difference of two means, which is a fundamental problem in statistical inference. For example, the colon dataset analyzed in Section 4 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 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

$$\mathbf{y}_{i1},\ldots,\mathbf{y}_{in_i} \text{ are iid with } \mathrm{E}(\mathbf{y}_{i1})=\boldsymbol{\mu}_i,\ \mathrm{cov}(\mathbf{y}_{i1})=\Sigma,\ i=1,2, \tag{1}$$

where the dimension p of $\mathbf{y}_{i1}$, that is, the number of variables in $\mathbf{y}_{i1}$, 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

$$H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2 \quad \text{versus} \quad H_1:\ \boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2. \tag{2}$$

When p is fixed and much smaller than n, it is well known that the classical Hotelling's (1931) T^2-test is the most powerful invariant test (Anderson 2003). Its test statistic is defined as

$$T_{\mathrm{H}}=(n_1n_2/n)(\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2)^\top\hat{\Sigma}^{-1}(\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2),$$

where $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$ are the sample mean vectors and $\hat{\Sigma}$ 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 $\hat{\Sigma}$ is usually nearly singular in this high-dimensional setting (Bai and Saranadasa 1996). To overcome this difficulty, the seminal work by Bai and Saranadasa (1996) proposed a nonexact test statistic which is equivalent to

$$T_{\mathrm{BS}}=(n_1n_2/n)\|\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2\|^2-\mathrm{tr}(\hat{\Sigma}),$$

where and throughout $\|\mathbf{a}\|$ denotes the L^2-norm of a vector $\mathbf{a}$ and $\mathrm{tr}(\mathbf{A})$ is the trace of a matrix $\mathbf{A}$. Bai and Saranadasa (1996) derived asymptotic normality of T_{BS} under H_0, and showed theoretically and with extensive simulation studies

CONTACT Ming-Yen Cheng chengmingyen@hkbu.edu.hk Department of Mathematics, Hong Kong Baptist University, 224 Waterloo Road, Kowloon Tong, Hong Kong.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

that when p tends to ∞ proportionally to 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 (1958, 1960). Another contribution of Bai and Saranadasa (1996) is deriving the asymptotic normality of Dempster's nonexact test statistic under H_0. Srivastava and Du (2008) proposed a scale-invariant test by replacing $\hat{\Sigma}$ in $T_{\rm H}$ with $\mathrm{diag}(\hat{\Sigma})$. Chen and Qin (2010) modified $T_{\rm BS}$ using U-statistics to treat the unequal covariances case and to allow p to be larger than n. Under some regularity conditions, in particular the assumptions on Σ detailed in (8)–(10), these authors showed that the null distributions of their test statistics are asymptotically normal, and they all based their tests on the asymptotic normality. Srivastava and Du (2008) further assumed Σ is positive definite so that its diagonal elements are all positive. When these regularity conditions are violated, the null distributions may not be approximately normal and could even depart from normality seriously. See the supplement for some illustrations.

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

$$T_{n,p} = (n_1 n_2/n)\|\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\|^2. \tag{3}$$

Under some mild conditions, we show that the null distribution of $T_{n,p}$ tends to a χ^2-type mixture (Zhang 2005), which is generally skewed although it may be asymptotically normal under some additional regularity conditions. Therefore, 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 (Satterthwaite 1946; Welch 1947). In addition, we give simple formulas for ratio-consistent estimation of the parameters in the W–S χ^2-approximation.

For more than six decades the W–S χ^2-approximation has been used to provide accurate and effective solutions to the classical Behrens–Fisher problems in the context of normal data with fixed and low dimensions (Box 1954; Feiveson and Delaney 1968; Satterthwaite 1941). One question arises naturally: is it still applicable in high (and possibly varying and diverging) dimensions, and both in the contexts of normal and nonnormal data? To our knowledge, this work is one of the first few attempts to thoroughly study this problem both theoretically and numerically. After carefully examining the effect of high dimensionality in estimation of the parameters, we show that the W–S χ^2-approximation indeed yields a simple and widely applicable test for high-dimensional data, regardless they are normal or nonnormal and without requiring any strong regularity conditions or model assumptions on Σ. This appealing property is not shared by normal approximation based tests such as those proposed by Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008).

Intuitively, for our test statistic $T_{n,p}$ given in (3) the W–S χ^2-approximation is preferred to the normal approximation for the following reason. Although both methods are two-moment matched approaches, depending on its degrees of freedom the χ^2-approximation can be either bell shaped or skewed while the normal approximation has a fixed bell shape. Thus, the W–S χ^2-approximation may be able to cope with the changing shape of the null distribution while the normal approximation may not. Theoretically speaking, we give a uniform bound on its approximation error and use that to show it is preferred to normal approximation 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. 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. The error bound indicates that the W–S χ^2-approximation is at least comparable to (when the correlation is low) and can be better than the normal approximation (when the correlation is moderate or high). In addition, we show that our test can automatically deal with the varying shape of the null distribution of $T_{n,p}$ 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 diverge to ∞, and when d is bounded by a constant the asymptotic normality cannot hold. Further, we show that none of the conditions (8)–(10) on Σ imposed by Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008) would hold when d is bounded by a constant. Therefore, when d is bounded by a constant our test is applicable but those normal approximation based tests are not, and our test is always applicable when they are. Note that the shape of the null distribution of $T_{n,p}$ is mainly determined by the unknown complex covariance matrix Σ and our test automatically copes with this without requiring strong conditions but simply capturing it via the parameter d. In particular, d being bounded by a constant basically corresponds to Σ being singular or nearly singular which is usually the case in high dimensions. The power of the proposed test is also studied. Finally, we conducted extensive simulation studies to examine finite sample performances of our test and existing tests including those suggested by Bai and Saranadasa (1996), Chen and Qin (2010), and a number of others. Compared to normal approximation based tests, our test has comparable size and comparable power if normal approximation is applicable, otherwise our test still retains similar size control and power while they become too liberal. In addition, it happens often that the other considered competitors are either too liberal or too conservative. Note that when a test is too liberal, its good power has not much meaning as it tends to reject the null hypothesis even if it holds. Therefore, we can conclude that our test is generally more reliable and more widely applicable in high dimensions.

There exist other approaches to solving the considered problem. Wang, Peng, and Qi (2013) proposed a jackknife empirical likelihood test which requires $p = o(\sqrt{n})$ and the implementation is complicated. Srivastava, Li, and Ruppert (2015) 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 implementation. To deal with heavy-tailed distributions, Wang, Peng, and Li (2015) gave a nonparametric extension of the U-statistic based test introduced by Chen and Qin (2010) using spatial sign. Following Chen and Qin (2010) they imposed condition (9). Our test also can cope with heavy-tailed distributed data after some proper modifications.

Recently, several authors proposed a few tests that involve nonnormal asymptotic null distributions. Katayama, Kanoa, and Srivastava (2013) studied one-sample testing for normal data using standardized version of the test statistic suggested by Bai and Saranadasa (1996) or that by Srivastava and Du (2008). They focused on the case where the covariance matrix Σ is positive definite and has spiked eigenvalues, that is, some of the eigenvalues are exponential functions of the dimension p while the other eigenvalues are constants. The asymptotic null distribution is an unknown function of Σ, and is typically a weighted sum of a standard normal component and a centered χ^2-type mixture component with the weights determined by the spiked model parameters. Therefore, successful implementation of their test strongly depends on accurate specification of the spiked model for Σ and proper estimation of the associated parameters. The authors implemented their test using the W–S χ^2-approximation assuming that the spiked model is correctly specified and the parameters are known. They did briefly mention maximum likelihood method estimation of these parameters but it is unknown if the resulting parameter estimators are ratio-consistent, especially in the high-dimensional case. It remains to be seen how this test can be applied properly in data analysis where the structure of Σ is generally unknown and one may not know if a spiked model can fit Σ well. Aoshima and Yata (2018) considered a weighted distance-based two-sample test when Σ has a strongly or non-strongly spiked structure where the weight matrix (which depends on the spiked model parameters) is assumed to be known. The asymptotic null distribution of their test statistic is standard normal for the non strongly spiked case and is a χ^2-type mixture for the strongly spiked case. They focused on the case when only the first eigenvalue of Σ is more spiked than all the others. However, their simulation results indicate that when the weight matrix is estimated from the data the resulting test is too liberal to be useful in practice. Pauly, Ellenberger, and Brunner (2015) considered a one-sample test for normally distributed repeated measurements. The asymptotic null distribution of their test statistic is either the standard normal distribution or a centered χ^2-type mixture. They suggested using the three-cumulant matched χ^2-approximation (Zhang 2005) and estimating the parameters by U-statistics, which are computationally intensive and not translation-invariant. The resulting parameter estimators are shown to be ratio-consistent only under the null hypothesis, but in practice one may not know if the null hypothesis holds or not. Compared with these methods our test has one major advantage: we do not assume any particular structure on the unknown covariance matrix, for example, a spiked model, which in practice is difficult to validate and involves complicated parameters that are challenging to estimate; instead our test automatically deals with the changing shape of the null distribution which is mainly determined by the covariance structure. In addition, we do not assume the data are normal, and our parameter estimators are ratio-consistent under both null and alternative hypotheses.

In summary, our test is simple to implement 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 Σ, and even different distributions of the data. In addition, it enjoys superior size accuracy and good power when the covariance matrix is singular or nearly singular which happens often in the high-dimensional setting. 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 several mean vectors or several covariance matrices. For example, Zhang, Guo, and Zhou (2017) extended the L^2-norm based test proposed in an earlier version of this work to linear hypothesis testing in high-dimensional one-way MANOVA. They showed that the null distribution of the test statistic for normal data is a χ^2-type mixture but, unlike the current article, the associated result for nonnormal data (see Theorem 2) has not been established. In addition, the fundamental uniform bound on the approximation error of the W–S χ^2-approximation (see Theorem 6) has not been established before.

The methods and main results are described in Section 2. Simulation results and application to the colon data are given in Sections 3 and 4, respectively. We give some concluding remarks in Section 5 and leave proofs of the main results to the Appendix. Histograms of simulated T_{BS}, additional numerical comparisons, and some lemmas and their proofs are given in the supplement.

2. Main Results

To test (2), we need to derive the null distribution of $T_{n,p}$ defined in (3). For this purpose, we set $\mathbf{x}_{ij} = \mathbf{y}_{ij} - \boldsymbol{\mu}_i, j = 1, \ldots, n_i; i = 1, 2$. Then we have $\bar{\mathbf{x}}_i = \bar{\mathbf{y}}_i - \boldsymbol{\mu}_i, i = 1, 2$. We now write

$$T_{n,p} = T_{n,p,0} + 2S_{n,p} + \|\boldsymbol{\mu}_{n,p}\|^2, \tag{4}$$

where $T_{n,p,0} = (n_1n_2/n)\|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|^2$, $S_{n,p} = (n_1n_2/n)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, and $\boldsymbol{\mu}_{n,p} = \sqrt{n_1n_2/n}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Clearly, $T_{n,p}$ equals $T_{n,p,0}$ under the null hypothesis.

Let $\lambda_{p,1}, \ldots, \lambda_{p,p}$ be the eigenvalues of Σ in the descending order. Let χ^2_v denote a central chi-square distribution with v degrees of freedom. Let $\stackrel{d}{=}$ and $\xrightarrow{\mathcal{L}}$ denote equality in distribution and convergence in distribution, respectively. Let $A_1, \ldots, A_r, \ldots$ be iid χ^2_1 random variables. Based on central limit theorem and continuous mapping theorem, we first have the following useful theorem.

Theorem 1. Assume that as $n \to \infty$, we have $n_1/n \to \tau \in (0, 1)$. Then, for any fixed finite p, we have $T_{n,p,0} \xrightarrow{\mathcal{L}} T_{p,0}$ as $n \to \infty$, where here and throughout $T_{p,0}$ is defined as

$$T_{p,0} \stackrel{d}{=} \sum_{r=1}^{p} \lambda_{p,r} A_r. \tag{5}$$

In addition, the first three cumulants of $T_{p,0}$ are $\mathrm{E}(T_{p,0}) = \mathrm{tr}(\Sigma)$, $\mathrm{var}(T_{p,0}) = 2\,\mathrm{tr}(\Sigma^2)$, and $\mathrm{E}[T_{p,0} - E(T_{p,0})]^3 = 8\,\mathrm{tr}(\Sigma^3)$.

Note that when the two samples (1) are normal, we always have $T_{n,p,0} \stackrel{d}{=} T_{p,0}$ for any n and fixed p. In theory, one may also be interested in the limiting distribution of $T_{n,p,0}$ when both n and p tend to infinity.

Set $\rho_{p,r} = \lambda_{p,r}/\left[\operatorname{tr}(\Sigma^2)\right]^{1/2}, r = 1, 2, \ldots, p$, which are the eigenvalues of $\Sigma/\left[\operatorname{tr}(\Sigma^2)\right]^{1/2}$ in the descending order. We impose the following assumptions:

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

Assumption 2. Assume $\mathrm{E}(z_{ijk}^4) = 3 + \Delta < \infty$ where z_{ijk} is the kth component of $\mathbf{z}_{ij}$, Δ is some constant, and $\mathrm{E}(z_{ij1}^{\nu_1} \ldots z_{ijp}^{\nu_p}) = 0$ (or 1) when there is one ν_k that equals 1 (or there are two ν_k's that are 2) whenever $\nu_1 + \cdots + \nu_p = 4$, where $\nu_1, \ldots, \nu_p$ are nonnegative integers.

Assumption 3. $n_1/n \to \tau \in (0, 1)$.

Assumption 4. There exist real numbers $\rho_{\infty,r}, r = 1, 2, \ldots$, such that $\lim_{p\to\infty} \rho_{p,r} = \rho_{\infty,r}, r = 1, 2, \ldots$, uniformly and $\lim_{p\to\infty} \sum_{r=1}^p \rho_{p,r} = \sum_{r=1}^\infty \rho_{\infty,r} < \infty$.

Assumptions 1 and 2 are also imposed by Bai and Saranadasa (1996) and Chen and Qin (2010). They specify a factor model for high-dimensional data analysis. Assumption 3 is a standard regularity assumption in two-sample problems, which guarantees that n_1 and n_2 go to infinity proportionally. Assumption 4 ensures existence of the limits of the eigenvalues of $\Sigma/\left[\operatorname{tr}(\Sigma^2)\right]^{1/2}$ as $p \to \infty$ and that the limit and summation operations in $\lim_{p\to\infty} \sum_{r=1}^p \rho_{p,r}$ are exchangeable. Note that Bai and Saranadasa (1996) also assumed that $p/n \to c \in (0, \infty)$, but this assumption was not used in the proofs of their main results. We do not need such an assumption either.

Theorem 2. Under Assumptions 1–4, as $n, p \to \infty$, we have

$$[T_{n,p,0} - \operatorname{tr}(\Sigma)]/\left[2\operatorname{tr}(\Sigma^2)\right]^{1/2} \xrightarrow{\mathcal{L}} \zeta, \quad \text{and}$$
$$[T_{p,0} - \operatorname{tr}(\Sigma)]/\left[2\operatorname{tr}(\Sigma^2)\right]^{1/2} \xrightarrow{\mathcal{L}} \zeta, \tag{6}$$

where $\zeta \stackrel{d}{=} \sum_{r=1}^\infty \rho_{\infty,r}(A_r - 1)/\sqrt{2}$. Therefore, as $n, p \to \infty$, we have

$$\sup_x \left|\Pr(T_{n,p,0} \le x) - \Pr(T_{p,0} \le x)\right| \to 0. \tag{7}$$

To show the asymptotic normality of their test statistics, besides Assumptions 1–3, Bai and Saranadasa (1996) imposed the following key condition

$$\lambda_{p,\max}^2 = o[\operatorname{tr}(\Sigma^2)], \quad \text{as } p \to \infty, \tag{8}$$

where $\lambda_{p,\max}$ denotes the largest eigenvalue of Σ; Chen and Qin (2010) required that

$$\operatorname{tr}(\Sigma^4) = o[\operatorname{tr}^2(\Sigma^2)], \quad \text{as } p \to \infty; \tag{9}$$

while Srivastava and Du (2008) assumed that

$$\operatorname{tr}(\Sigma^\ell)/p \to a_\ell \in (0, \infty), \ell = 1, 2, 3, \quad \text{as } p \to \infty. \tag{10}$$

Theorem 3. Under Assumptions 1–3 and any of the conditions (8)–(10), as $n, p \to \infty$, we have

$$[T_{n,p,0} - \operatorname{tr}(\Sigma)]/\left[2\operatorname{tr}(\Sigma^2)\right]^{1/2} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{and}$$
$$[T_{p,0} - \operatorname{tr}(\Sigma)]/\left[2\operatorname{tr}(\Sigma^2)\right]^{1/2} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \tag{11}$$

and so we have the same uniform convergence of $\left|\Pr(T_{n,p,0} \le x) - \Pr(T_{p,0} \le x)\right|$ to 0 as given in (7).

Suppose Assumptions 1–3 hold. Then Theorem 2 shows that $T_{n,p,0}$ and $T_{p,0}$ have the same nonnormal limit when Assumption 4 is satisfied, and Theorem 3 shows that $T_{n,p,0}$ and $T_{p,0}$ have the same normal limit when any of the conditions (8)–(10) is true. These results show that in either case the distribution of $T_{n,p,0}$ is close to that of $T_{p,0}$ when n and p are large.

2.1. W–S χ^2-Approximation to the Distribution of $T_{p,0}$

Theorems 1–3 suggest that, under appropriate conditions, we can approximate the distribution of $T_{n,p,0}$ by that of $T_{p,0}$ when p is fixed finite and $n \to \infty$, and when both $n, p \to \infty$. Note that $T_{p,0}$ is a χ^2-type mixture whose coefficients are the eigenvalues of Σ. These eigenvalues are generally unknown and are often difficult to estimate consistently, especially when p is large or even much larger than n. Therefore, Imhof's (1961) approach to computing the exact distribution of a χ^2-type mixture with a few known coefficients cannot be directly applied to find the distribution of $T_{p,0}$. Here we suggest applying the well-known W–S χ^2-approximation, also known as the Box χ^2-approximation (Box 1954), to approximate the distribution of $T_{p,0}$. The key idea is to approximate the distribution of $T_{p,0}$ by that of a random variable R of the following form

$$R \stackrel{d}{=} \beta\chi_d^2, \tag{12}$$

where β and d are unknown parameters which can be determined via matching the means and variances (or equivalently the first two cumulants) of $T_{p,0}$ and R. We call d the approximate degrees of freedom of the W–S χ^2-approximation. By the definition of $T_{p,0}$ given in (5), we have $\mathrm{E}(T_{p,0}) = \operatorname{tr}(\Sigma)$ and $\operatorname{var}(T_{p,0}) = 2\operatorname{tr}(\Sigma^2)$ where we use the fact that

$$\sum_{r=1}^p \lambda_{p,r}^\ell = \operatorname{tr}(\Sigma^\ell), \quad \ell = 1, 2, \ldots. \tag{13}$$

By the definition of R given in (12), we have $\mathrm{E}(R) = \beta d$ and $\operatorname{var}(R) = 2\beta^2 d$. Equating the means and variances of $T_{p,0}$ and R, respectively, we have

$$\beta = \operatorname{tr}(\Sigma^2)/\operatorname{tr}(\Sigma) \quad \text{and} \quad d = \operatorname{tr}^2(\Sigma)/\operatorname{tr}(\Sigma^2). \tag{14}$$

Note that the degrees of freedom d defined in the above may not be an integer. Throughout this article, a χ_ν^2 distribution is treated as a Gamma$(\nu/2, 1/2)$ distribution when ν is a noninteger. Note also that the skewness of R is $\sqrt{8/d}$ which tends to zero when $d \to \infty$.

The above χ^2-approximation has several appealing properties. First of all, it is simple to implement provided that the unknown parameters β and d can be consistently estimated based on the data. Second, to estimate β and d there is no need to estimate the eigenvalues of Σ consistently, instead we only need to estimate $\operatorname{tr}(\Sigma), \operatorname{tr}(\Sigma^2)$, and $\operatorname{tr}^2(\Sigma)$ consistently. Third, as shown in Section 2.2, the W–S χ^2-approximation automatically mimics the shape of the distribution of $T_{p,0}$ in the sense described in Section 1. That is, d will tend to ∞ (and R will be bell shaped) when $T_{p,0}$ is asymptotically normal, and $T_{p,0}$ will not tend to normal when d is bounded by a constant. Finally, the χ^2-approximation to the distribution of $T_{p,0}$ is at least comparable to and often outperforms the normal approximation in terms of approximation accuracy both theoretically (see Section 2.3) and numerically (see Section 3).

2.2. Asymptotic Distributions of $T_{p,0}$ and the W–S χ^2-Approximation

It is of interest to see if the W–S χ^2-approximation R defined in (12) can automatically mimic the shape of the χ^2-type mixture distribution of $T_{p,0}$ when $p \to \infty$. To study this problem, we first need to answer the question "Under what conditions the distribution of $T_{p,0}$ tends to normal?" First, we note that the skewness of $T_{p,0}$ is given by

$$\mathrm{E}[T_{p,0} - \mathrm{E}(T_{p,0})]^3 / \mathrm{var}^{3/2}(T_{p,0}) = \left(8/d^*\right)^{1/2}, \quad \text{where } d^* = \mathrm{tr}^3(\Sigma^2)/\,\mathrm{tr}^2(\Sigma^3). \tag{15}$$

When the distribution of $T_{p,0}$ tends to normal, its skewness must tend to 0, implying that d^* must tend to ∞. That is to say "$d^* \to \infty$" is a necessary condition for $T_{p,0}$ to tend to normal in distribution. Theorem 4 shows that "$d^* \to \infty$" is a necessary and sufficient condition for $T_{p,0}$ to tend to normal in distribution.

Theorem 4. As $p \to \infty$, the distribution of $T_{p,0}$ tends to normal if and only if $d^* \to \infty$.

By (13), both d and d^* depend only on the eigenvalues of Σ. Some interesting facts about the relationship between d, d^*, and p are established in the following theorem.

Theorem 5. We have (a) $1 \le d^* \le d \le p$; (b) $d^* = d = 1$ if and only if only the first eigenvalue $\lambda_{p,1}$ is nonzero; and (c) $d^* = d = p$ if and only if all the eigenvalues $\lambda_{p,r}, r = 1, \ldots, p$, are the same.

Part (a) of Theorem 5 indicates that d^* and d are always bounded if p is a given number, and d^* and d can tend to infinity only when $p \to \infty$. Notice from (14) and (15) that when the first k $(k < p)$ eigenvalues of Σ 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 Σ 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 is known that when the p variables are highly correlated the first few eigenvalues of Σ will be much larger than the remaining ones, and when they are less correlated all the eigenvalues are nearly the same. Therefore, d^* and d are small (or large) when the variables are highly (or less) correlated. See Example 1 for an illustration.

As mentioned earlier, to obtain critical values or p-values of their tests via using normal approximation, Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008) imposed regularity assumptions to show the asymptotic normality of their test statistics, including their key conditions given in (8)–(10). In practice, it may be difficult to justify any of these conditions, however. In fact, from Theorems 4 and 5 and the proof of Theorem 3, we know that when d is bounded by a constant none of the conditions (8)–(10) would hold and so these existing tests are not applicable. For example, Assumption 4 implies that, as $p \to \infty$, we have

$$d = \mathrm{tr}^2(\Sigma)/\,\mathrm{tr}(\Sigma^2) \to \Big[\sum_{r=1}^{\infty} \rho_{\infty,r}\Big]^2 < \infty. \tag{16}$$

By comparison, in situations where d tends to a finite limit, d^* also tends to a finite limit, and the distributions of $T_{p,0}$ will not tend to normal as $p \to \infty$, so our approach of using the distribution of R to approximate that of $T_{p,0}$ is still appropriate. The following corollary identifies another situation where d tends to a finite limit.

Corollary 1. Assume that, as $p \to \infty$, we have $\mathrm{tr}(\Sigma^\ell)/p^\ell \to b_\ell \in (0, \infty)$, $\ell = 1, 2, 3$. Then, as $p \to \infty$, we have $d^* \to b_2^3/b_3^2 \in (0, \infty)$, $d \to b_1^2/b_2 \in (0, \infty)$, and the distributions of $T_{p,0}$ and R are skewed and do not tend to normal.

Note that the conditions specified in Corollary 1 look similar to but are indeed different from condition (10). In fact, from the proof of Theorem 3 we can see that condition (10) implies $d^* \to \infty$.

By Theorems 4 and 5, we have the following corollary.

Corollary 2. When d is bounded, d^* is also bounded, and so neither of the distributions of $T_{p,0}$ and R will tend to normal. When $d^* \to \infty$, we have asymptotic normality of $T_{p,0}$ given in (11), and we also have $d \to \infty$ which implies $[R - \mathrm{tr}(\Sigma)]/[2\,\mathrm{tr}(\Sigma^2)]^{1/2} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Corollary 2 shows that the distribution of R automatically mimics the shape of the distribution of $T_{p,0}$ in the sense described in Sections 1 and 2.1. Therefore, in practice there is no need to check whether the distribution of $T_{p,0}$ is asymptotically normal or not. This appealing property is not shared by any of those tests by Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008) as implied by part (a) of Theorem 5 and discussed earlier.

Corollary 2 also indicates that we may use the value of (an estimate of) d to determine if the normal approximation to the distribution of $T_{p,0}$ is adequate. This helps explain why the proposed L^2-norm based test with the W–S χ^2-approximation is preferred to the tests by Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008) as we will do in our numerical studies.

In Theorem 3, we show that the conditions (8)–(10) required by the tests by Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008) are sufficient for the distribution of $T_{p,0}$ to tend to normal. We emphasize that this is only for the purpose of showing the adaptivity of the W–S χ^2-approximation R to the distribution of $T_{p,0}$, and we do not need to check these conditions to see if the normal approximation to the distribution of $T_{p,0}$ is adequate. In the implementation of the proposed L^2-norm based test, we recommend to always use the W–S χ^2-approximation and its adaptivity property provides a good justification for this approach. Further, Remark 1 given in Section 2.3 says the W–S χ^2-approximation is at least comparable to and is often better than the normal approximation in terms of accuracy (even when d is large). Last but not least, Section 3 shows numerically that the W–S χ^2-approximation is adequate even when d is as small as 1 or 2.

We close this section via giving the following interesting example.

Example 1. Let $\Sigma = \sigma^2\left[(1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p\right]$ be a compound symmetric matrix where $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\mathbf{J}_p$ is the $p \times p$ matrix of ones, $0 \leq \rho \leq 1$ and $\sigma^2 > 0$. By simple algebra we have $d^* = \frac{[(1-\rho^2)+\rho^2 p]^3 p}{[(1-\rho)^2(1+2\rho)+3(1-\rho)\rho^2 p+\rho^3 p^2]^2}$, and $d = \frac{p}{1+(p-1)\rho^2}$. 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 $\rho = 0$ (or $\rho = 1$) we have $d^* = d = p$ (or $d^* = d = 1$). Suppose $\rho = Cp^{-\tau}$ where $0 < C < \infty$ and $0 < \tau < \infty$ are fixed constants. Then, we have $d \to \infty$ as $p \to \infty$, and we have $d^* \to \infty$ when $\tau > 1/2$, $d^* \to 1$ when $0 < \tau < 1/2$ and $d^* \to (1 + C^2)^3/C^6$ when $\tau = 1/2$. Note that the condition "$\rho = Cp^{-\tau}$ with $\tau > 0$" essentially corresponds to $\Sigma - \sigma^2\mathbf{I}_p \to \mathbf{0}$ as $p \to \infty$, which occurs extremely rarely 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^* decreases while d increases as p increases; and for a fixed p, both d^* and d decrease as ρ increases. Especially, when ρ is moderate or large both d^* and d are small, indicating that a normal approximation is not adequate.

2.3. Accuracy of the W–S Approximation to the Distribution of $T_{p,0}$

The W–S χ^2-approximation to the distribution of a χ^2-type mixture is widely used in the literature of Behrens–Fisher problems due to its accuracy and simplicity. However, its theoretical justification has not been well addressed before. In this section, we study the accuracy of the W–S χ^2-approximation to the distribution of $T_{p,0}$ described in Section 2.1. For this purpose, 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 the W–S χ^2-approximation to the distribution of $T_{p,0}$.

Theorem 6. Let $\Delta = \lambda^2_{p,\max}/\operatorname{tr}(\Sigma^2)$ and $M = \operatorname{tr}(\Sigma^4)/\operatorname{tr}^2(\Sigma^2)$. Then, when $\Delta < 1/10$ and $d > 10$ we have

$$\sup_x |f_{\tilde{T}_{p,0}}(x) - f_{\tilde{R}}(x)| \leq 0.1403\left\{\left[3 + \frac{3.8578}{(1-10\Delta)^{5/2}}\right]M + \left[3 + \frac{3.8578}{(1-10/d)^{5/2}}\right]\frac{1}{d}\right\} + 0.7040\left[\frac{1}{\sqrt{d^*}} - \frac{1}{\sqrt{d}}\right].$$

Remark 1. By Theorem 1(a) of Zhang (2005), the uniform error bound of the normal approximation to the distribution of $T_{p,0}$ is $O(1/\sqrt{d^*})$. By Theorem 6, the uniform error bound of the W–S χ^2-approximation is $O(M)+O(1/d)+O(1/\sqrt{d^*}-1/\sqrt{d})$. Note that the $O(M)$, $O(1/d)$, and $O(1/\sqrt{d^*} - 1/\sqrt{d})$ terms are all of smaller orders than $O(1/\sqrt{d^*})$ when $d^* \to \infty$, and they are at least comparable to and generally smaller than $O(1/\sqrt{d^*})$ when d^* is bounded. Thus, it is theoretically justifiable that the W–S χ^2-approximation is at least comparable and generally preferred to the normal approximation in terms of the upper bounds. In particular, when d^* and d are different (which is the case when the p variables are moderately or mildly correlated such as $0.1 \leq \rho \leq 0.7$ in Example 1), $O(1/\sqrt{d^*} - 1/\sqrt{d})$ dominates $O(M) + O(1/d)$ and is smaller than although sometimes comparable to $O(1/\sqrt{d^*})$, thus the W–S χ^2-approximation 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 variables are either nearly independent or highly correlated such as $\rho = 0$ or $\rho \geq 0.8$ in Example 1), $O(1/\sqrt{d^*} - 1/\sqrt{d})$ is close to 0 so that the error bound given in Theorem 6 is roughly $O(M)+O(1/d)$ which is much smaller than $O(1/\sqrt{d^*})$, thus the W–S χ^2-approximation is much better than the normal approximation. These conclusions are confirmed by the simulation results presented in Section 3 and explain why our test has a much better size control than other approaches when d is small as demonstrated in Section 3.

Remark 2. For decades the W–S χ^2-approximation has been a popular approach to the univariate Behrens–Fisher problems but lacks theoretical justifications. Theorem 6 solves this long-term open problem. Note that the error bound given in Theorem 6 is much sharper than the error bound $O(d^{*-1/6})$ obtained by Chuang and Shih (2012) for correlated χ^2-mixtures which cannot be used to argue that the W–S χ^2-approximation is better than the normal approximation.

Remark 3. It is well known that when p is large the variables 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 Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008), are not. Further, Remark 1 indicates that our approach is still superior even when the data are nearly independent and the normal approximation is adequate.

2.4. Implementation of the Proposed L^2-Norm Based Test

The proposed L^2-norm based test can be implemented easily provided that the parameters β and d in the W–S χ^2-approximation (12) are estimated ratio-consistently. Let $\hat{\theta}_{n,p}$ be an estimator of $\theta_{n,p}$, a nonrandom 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 $\xrightarrow{P} 1$ means convergence in probability. In practice, the group sample sizes n_1 and n_2, and the dimension p are always fixed. To take this into account, we implement the proposed L^2-norm based test via approximating the distribution of $T_{n,p,0}$ using the random variable R defined in (12) with the parameters β and d determined via matching the means and variances of $T_{n,p,0}$ and R, respectively.

We first consider the case where the two samples (1) are normal. In this case, for any n and fixed p, we have $T_{n,p,0} \stackrel{d}{=} T_{p,0}$ and the parameters β and d that matches the first two moments of $T_{n,p,0}$ and $R \stackrel{d}{=} \beta\chi^2_d$ are as given in (14). It follows that finding ratio-consistent estimators of β and d is equivalent to finding ratio-consistent estimators of $\operatorname{tr}(\Sigma)$, $\operatorname{tr}^2(\Sigma)$, and $\operatorname{tr}(\Sigma^2)$ which can be achieved in the following way. The usual unbiased estimator of Σ is the pooled sample covariance matrix $\hat{\Sigma} = (n-2)^{-1}\sum_{i=1}^2\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^\top$. Since the two samples (1) are normal, $\hat{\Sigma}$ follows the Wishart distribution $W_p[n-2, \Sigma/(n-2)]$. By Lemma S.3 given in the supplement, uniformly in p, $\operatorname{tr}(\hat{\Sigma})$,

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

p	ρ	0	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	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

$$\begin{aligned}\widehat{\mathrm{tr}^2(\Sigma)} &= [(n-2)(n-1)]/[(n-3)n] \\ &\quad\times\left[\mathrm{tr}^2(\hat\Sigma) - 2\,\mathrm{tr}(\hat\Sigma^2)/(n-1)\right], \text{ and} \\ \widehat{\mathrm{tr}(\Sigma^2)} &= (n-2)^2/[(n-3)n]\left[\mathrm{tr}(\hat\Sigma^2) - \mathrm{tr}^2(\hat\Sigma)/(n-2)\right]\end{aligned} \tag{17}$$

are unbiased and ratio-consistent estimators of $\mathrm{tr}(\Sigma), \mathrm{tr}^2(\Sigma)$, and $\mathrm{tr}(\Sigma^2)$, respectively. Then, plugging these estimators into the formulas for β and d given in (14), we obtain ratio-consistent estimators of β and d defined as

$$\hat\beta = \widehat{\mathrm{tr}(\Sigma^2)}/\,\mathrm{tr}(\hat\Sigma) \text{ and } \hat d = \widehat{\mathrm{tr}^2(\Sigma)}/\widehat{\mathrm{tr}(\Sigma^2)}. \tag{18}$$

For given significance level $\alpha > 0$, let $\chi^2_d(\alpha)$ denote the upper 100α percentile of the χ^2_d distribution. Theorem 7 shows that $\hat\beta, \hat d$, and $\hat\beta\chi^2_{\hat d}(\alpha)$ are ratio-consistent for β, d, and $\beta\chi^2_d(\alpha)$, respectively. Based on this result, we can conduct the proposed L^2-norm based test via using the approximate critical value $\hat\beta\chi^2_{\hat d}(\alpha)$ or the approximate p-value $\Pr(\chi^2_{\hat d} \geq T_{n,p}/\hat\beta)$.

Theorem 7. Assume that the two samples (1) are normally distributed. Then, as $n \to \infty$, we have

$$\hat\beta/\beta \xrightarrow{P} 1, \quad \hat d/d \xrightarrow{P} 1, \quad \text{and} \quad \hat\beta\chi^2_{\hat d}(\alpha)/\left[\beta\chi^2_d(\alpha)\right] \xrightarrow{P} 1, \tag{19}$$

uniformly in p.

We now consider the case when the two samples (1) are nonnormal. In this case, Theorems 1 and 2 show that under some regularity conditions the distribution of $T_{n,p,0}$ can be still approximated by a χ^2-type mixture. Therefore, we can still apply the W–S χ^2-approximation (12) to the distribution of $T_{n,p,0}$. Again, for any given high-dimensional dataset, n and p are always fixed. To take this into account, we implement the proposed L^2-norm based test via matching the means and variances of $T_{n,p,0}$ and R directly. In the supplement we show that

$$\mathrm{E}(T_{n,p,0}) = \mathrm{tr}(\Sigma), \quad \text{and} \quad \mathrm{var}(T_{n,p,0}) = 2\,\mathrm{tr}(\Sigma^2) + \delta, \tag{20}$$

where

$$\begin{aligned}\delta &= (n_2/n)^2\,\kappa_{1,11}/n_1 + (n_1/n)^2\,\kappa_{2,11}/n_2, \\ \kappa_{i,11} &= \mathrm{E}(\|\mathbf{y}_{i1} - \boldsymbol{\mu}_i\|^4) - \mathrm{tr}^2(\Sigma) - 2\,\mathrm{tr}(\Sigma^2), \; i = 1, 2.\end{aligned} \tag{21}$$

Here $\kappa_{i,11}$ is a quantity used to measure the nonnormality of $\mathbf{y}_{i1}$, $i = 1, 2$, (Himeno and Yamada 2014). By equating the means and variances of $T_{n,p,0}$ and R, respectively, we have

$$\beta = [\mathrm{tr}(\Sigma^2) + \delta/2]/\,\mathrm{tr}(\Sigma), \quad d = \mathrm{tr}^2(\Sigma)/[\mathrm{tr}(\Sigma^2) + \delta/2]. \tag{22}$$

Note that when the two samples (1) are normal, the above two formulas reduce to those given in (14).

To estimate the unknown parameters β and d given in (22) ratio-consistently, we adopt the following results of Himeno and Yamada (2014). Under some regularity conditions, based on the ith sample only, the unbiased and ratio-consistent estimators of $\mathrm{tr}(\Sigma), \mathrm{tr}(\Sigma^2), \mathrm{tr}^2(\Sigma)$, and $\kappa_{i,11}$ are given by $\mathrm{tr}(\hat\Sigma_i)$ and

$$\begin{aligned}a_i &= (n_i-1)[n_i(n_i-2)(n_i-3)]^{-1}[(n_i-1)(n_i-2)\,\mathrm{tr}(\hat\Sigma_i^2) \\ &\quad + \mathrm{tr}^2(\hat\Sigma_i) - n_iQ_i], \\ b_i &= (n_i-1)[n_i(n_i-2)(n_i-3)]^{-1}[2\,\mathrm{tr}(\hat\Sigma_i^2) \\ &\quad + (n_i^2 - 3n_i + 1)\,\mathrm{tr}^2(\hat\Sigma_i) - n_iQ_i], \\ \hat\kappa_{i,11} &= -[(n_i-2)(n_i-3)]^{-1}[2(n_i-1)^2\,\mathrm{tr}(\hat\Sigma_i^2) \\ &\quad + (n_i-1)^2\,\mathrm{tr}^2(\hat\Sigma_i) - n_i(n_i+1)Q_i],\end{aligned}$$

where $\hat\Sigma_i = (n_i-1)^{-1}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^\top$ 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) together, the unbiased and ratio-consistent estimators of $\mathrm{tr}(\Sigma), \mathrm{tr}(\Sigma^2)$ and $\mathrm{tr}^2(\Sigma)$ are, respectively, given by

$$\begin{aligned}\mathrm{tr}(\hat\Sigma) &= [(n_1-1)\,\mathrm{tr}(\hat\Sigma_1) + (n_2-1)\,\mathrm{tr}(\hat\Sigma_2)]/(n-2), \\ \widehat{\mathrm{tr}(\Sigma^2)} &= [(n_1-1)a_1 + (n_2-1)a_2]/(n-2), \quad \text{and} \\ \widehat{\mathrm{tr}^2(\Sigma)} &= [(n_1-1)b_1 + (n_2-1)b_2]/(n-2).\end{aligned} \tag{23}$$

Plugging these estimators into (21) and (22), we have the following estimators for δ, β, and d

$$\begin{aligned}\hat\delta &= (n_2/n)^2\,\hat\kappa_{1,11}/n_1 + (n_1/n)^2\,\hat\kappa_{2,11}/n_2\,, \\ \hat\beta &= \left[\widehat{\mathrm{tr}(\Sigma^2)} + \hat\delta/2\right]/\,\mathrm{tr}(\hat\Sigma), \quad \text{and} \\ \hat d &= \widehat{\mathrm{tr}^2(\Sigma)}/\left[\widehat{\mathrm{tr}(\Sigma^2)} + \hat\delta/2\right].\end{aligned} \tag{24}$$

Then, when the data are nonnormal, the ratio-consistency results given in (19) of Theorem 7 continue to hold with β and d as defined in (22) and $\hat\beta$ and $\hat d$ as defined in (24).

In the above, we describe two approaches to estimating β and d consistently. The formulas of the first approach are obtained with the normality assumption; they are simple to understand and easy to calculate. The formulas of the second approach are obtained and can be used without the normality assumption, but they are harder to understand and time-consuming to calculate. In practice, it is often difficult to check if the two samples (1) are normal or not, especially when the dimension p is large. We can simply use the second approach as it works for normal data as well ($\delta = 0$ and $\hat\delta$ would be small if normality holds). On the other hand, the following results show that when the data satisfy Assumptions 1–3, we can also use the first approach regardless if the data are normal or not.

Theorem 8. Under Assumptions 1–3, as $p \to \infty$, we have $\delta = o[\mathrm{tr}(\Sigma^2)]$,

$$\beta = \left[\mathrm{tr}(\Sigma^2)/\,\mathrm{tr}(\Sigma)\right][1+o(1)], \quad \text{and} \\ d = \left[\mathrm{tr}^2(\Sigma)/\,\mathrm{tr}(\Sigma^2)\right][1+o(1)]. \tag{25}$$

Theorem 8 shows that the formulas for β and d given in (14) prescribed for normal data would hold approximately for nonnormal data if Assumptions 1–3 are satisfied. That is, in this case, the effect of nonnormality can be ignored asymptotically. We show further in the following theorem that the effect of nonnormality on $\hat{\beta}$ and $\hat{d}$ defined in (18) for normal data can be ignored asymptotically as well.

Theorem 9. Suppose Assumptions 1–3 hold. Then, when the data are nonnormal, as $n \to \infty$, the ratio-consistency results given in (19) for $\hat{\beta}$ and $\hat{d}$ defined in (18), and the resulting $\hat{\beta}\chi^2_{\hat{d}}(\alpha)$, still hold uniformly in p.

Theorem 9 shows that the first approach can also be used even when the two samples (1) are nonnormal provided that Assumptions 1–3 are satisfied. This conclusion is actually confirmed by the simulation studies and the data example presented in Sections 3 and 4, respectively.

2.5. *Power of the Proposed L^2-Norm Based Test*

In this section, we investigate the power of our L^2-norm based test. Recall that we have the decomposition of $T_{n,p}$ given in (4), and note that $\mathrm{var}(S_{n,p}) = \boldsymbol{\mu}_{n,p}^\top \Sigma \boldsymbol{\mu}_{n,p}$ where $\boldsymbol{\mu}_{n,p}$ is defined in (4). Similar to Bai and Saranadasa (1996) and Chen and Qin (2010), we consider power of our test based on $T_{n,p}$ under the following local alternative

$$\boldsymbol{\mu}_{n,p}^\top \Sigma \boldsymbol{\mu}_{n,p} = o[\mathrm{tr}(\Sigma^2)] \quad \text{as } n,p \to \infty. \tag{26}$$

In this case $T_{n,p} = \|\boldsymbol{\mu}_{n,p}\|^2 + T_{n,p,0}[1+o_p(1)]$ as $n,p \to \infty$.

Based on Theorems 7 and 9, we consider the estimators $\hat{\beta}$ and $\hat{d}$ defined in (18) with $\mathrm{tr}(\hat{\Sigma})$, $\widehat{\mathrm{tr}^2(\Sigma)}$, and $\widehat{\mathrm{tr}(\Sigma^2)}$ defined in (17). We consider the power of our test under Assumptions 1–3 and any of the conditions (8)–(10). In this case, by Theorem 3 both $T_{n,p,0}$ and $T_{p,0}$ are asymptotically normal as $n,p \to \infty$, and by the proof of Theorem 3, as $p \to \infty$, we have $d^* \to \infty$ and hence $d \to \infty$. Let $\Phi(\cdot)$ and z_α denote the cumulative distribution function and the $100(1-\alpha)$ percentile of $\mathcal{N}(0,1)$, respectively. We have the following theorem.

Theorem 10. Under Assumptions 1–3 and any of the conditions (8)–(10), and the local alternative (26), as $n,p \to \infty$, the power of our test can be expressed as

$$\Pr\left[T_{n,p} \geq \hat{\beta}\chi^2_{\hat{d}}(\alpha)\right] = \Phi\big\{-z_\alpha + n\tau(1-\tau)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/ \\ [2\,\mathrm{tr}(\Sigma^2)]^{1/2}\big\} + o(1). \tag{27}$$

Note that under the conditions of Theorem 10, the power of our test is the asymptotically same as that of the test suggested by Bai and Saranadasa (1996).

3. Simulation Studies

We conducted extensive 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) are assumed to be normal (or nonnormal) so that the W–S χ^2-approximation based on the estimators defined in (18) (or (24)) is applied. Under various simulation settings we compared L2N, L2D and some existing tests in terms of empirical size and power, aiming to see if L2N and L2D work well when the two samples (1) are actually normal or nonnormal 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 $\mathrm{ARE} = 100M^{-1}\sum_{j=1}^{M}|\hat{\alpha}_j - \alpha|/\alpha$, where α is the nominal size (e.g., 5%) and $\hat{\alpha}_j, j = 1,\ldots,M$, denote the empirical sizes under M simulation settings. A smaller value of ARE indicates better performance of a test in terms of size control.

3.1. *Simulation 1*

In this simulation study, we compare L2N and L2D with the tests proposed by Bai and Saranadasa (1996) and Chen and Qin (2010) denoted as BS and CQ, respectively. We generated the two samples (1) using $\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \Sigma^{1/2}\mathbf{z}_{ij}, j = 1,\ldots,n_i; i = 1,2$, where $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \delta\mathbf{h}$, $\Sigma = \sigma^2\left[(1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p\right]$ as in Example 1, and $\mathbf{z}_{ij} = (z_{ij1},\ldots,z_{ijp})^\top, j = 1,\ldots,n_i, i = 1,2$, are iid random variables coming from the following three models:

Model 1. $z_{ijt}, t = 1,\ldots,p, \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.
Model 2. $z_{ijt} = w_{ijt}/\sqrt{2}, t = 1,\ldots,p$ with $w_{ijt}, t = 1,\ldots,p, \overset{\text{iid}}{\sim} t_4$.
Model 3. $z_{ijt} = (w_{ijt}-1)/\sqrt{2}, t = 1,\ldots,p$ with $w_{ijt}, t = 1,\ldots,p, \overset{\text{iid}}{\sim} \chi^2_1$.

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 should increase as δ increases and the correlation increases 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 $\mathbf{u} = (1,\ldots,p)^\top$. 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 considered tests were obtained from $N = 10{,}000$ simulation runs with $\alpha = 5\%$.

Table 2 displays the empirical sizes of the four considered tests with the last row displaying their ARE values associated with the three values of ρ. We can draw several useful conclusions from Table 2. The first conclusion 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 settings. This shows that L2N and L2D are generally comparable and can be used regardless if the data are normal or nonnormal. This is consistent with the conclusions drawn from Theorems 8 and 9. The second conclusion is that BS and CQ are generally comparable, and they are generally liberal with

Table 2. Simulation 1: Empirical sizes (%).

			$\rho = 0.1$				$\rho = 0.5$				$\rho = 0.9$			
Model	p	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 (%).

				$\rho = 0.1$				$\rho = 0.5$				$\rho = 0.9$			
Model	p	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

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

			$\rho = 0.1$		$\rho = 0.5$		$\rho = 0.9$	
Model	p	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

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 approximate degrees of freedom d given in Table 1 and the estimated approximate degrees of freedom of L2N and L2D shown in Table 4 both become smaller as ρ increases, indicating that the normal approximation used by BS and CQ becomes less adequate as ρ increases. The last but not 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 better size accuracy and smaller ARE values. 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 the normal approximation used by BS and CQ is reasonable. This coincides with the conclusions drawn from Theorem 6, given in Remark 1.

Table 3 displays the empirical powers of the four tests when $\alpha = 5\%$. We can see that L2N and L2D have similar empirical powers, showing that they are comparable regardless whether the two samples (1) are normal or nonnormal. This is consistent with what we can observe from Table 2 that their empirical sizes are comparable. In addition, observe that the empirical powers of L2N and L2D are comparable with that of BS and CQ when $\rho = 0.1$ (in which case the four tests have comparable empirical sizes), and L2N and L2D have slightly lower powers than BS and CQ when $\rho = 0.5$ or 0.9 (in which case L2N and L2D have better size control than BS and CQ). Note that, when $\rho = 0.5$ or 0.9, 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. However, for the same reason that they are generally more liberal, they are less reliable in that case. Finally, we can see that under various settings the empirical powers of the four tests become smaller as ρ increases. This is reasonable because the noise level increases as ρ increases.

For each of the considered settings, Table 4 displays the estimated approximate degrees of freedom $\hat{d}$ of L2N and L2D defined in (18) and (24), respectively. First, observe that estimated approximate degrees of freedom obtained with or without normality assumption are about the same under each setting. This explains why the two approaches to implement our test give comparable results in terms of both size accuracy and power as shown in Tables 2 and 3. This is also consistent with the conclusions drawn from Theorems 8 and 9. Second, we can see that as the sample sizes increase the estimated approximate degrees of freedom of L2N and L2D become closer to the true approximate degrees of freedom given in Table 1. This is consistent with the conclusions drawn from Theorems 7 and 9. Third, notice that as the value of ρ increases the estimated approximate degrees of freedom of L2N and L2D become smaller. This shows that normal approximation becomes less adequate as ρ increases. Therefore, we expect that BS and CQ are less accurate when the correlation is larger, which can also be observed from Table 2.

3.2. Simulation 2

In this simulation study, we continue to use the setup of Simulation 1 except that we now use a new covariance matrix Σ defined as

$$\Sigma = \mathbf{DRD},\ \mathbf{R} = (\rho^{|i-j|}),$$

where $\mathbf{D} = \text{diag}(\mathbf{h})$ and $\mathbf{h}$ is as in Simulation 1. Such a covariance matrix was also used by Srivastava, Katayama, and Kano (2013) and Yamada and Himeno (2015). Note that the tuning parameter ρ here plays a somewhat different role from that 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 (or large) the p variables are less (or 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.

3.3. Simulation 3

In this simulation study, we consider the moving average model studied by Chen and Qin (2010) in which the tth component of $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijp})^\top$ is generated in the following way

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

where μ_{it} is the tth component of $\boldsymbol{\mu}_i$, $i = 1, 2$, and the innovations z_{ijt}, $t = 1, \ldots, p$, $j = 1, \ldots, n_i$, $i = 1, 2$, are iid random variables coming from the following two models:

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

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

For the dependence between the components of $\mathbf{y}_{ij}, j = 1, \ldots, n_i; i = 1, 2$, 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 p variables are less (highly) correlated in the partial (full) dependence case. As in Simulations 1 and 2, the number of simulation runs was 10,000. 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.

3.4. Simulation 4

In this simulation study, we compare L2N and L2D with the tests proposed by Srivastava and Kubokawa (2013), Cai, Liu, and Xia (2014) and Gregory et al. (2015), denoted as SK, CLX, and GCBL, respectively. The data were generated as in Simulation 2, but we now focus on normal data only and the number of simulation runs was reduced to 1000 as CLX is very time-consuming to compute. Since we assume the two samples (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 Gregory et al. (2015), we chose the "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. Observe that L2N and L2D are again the winners in

Table 5. Simulation 2: Empirical sizes (%).

Model	p	n	$\rho = 0.1$ L2N	L2D	BS	CQ	$\rho = 0.5$ L2N	L2D	BS	CQ	$\rho = 0.9$ 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

Table 6. Simulation 3: Empirical sizes (%).

	Dependence		Partial				Full			
Model	p	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

Table 7. Simulation 4: Empirical sizes (%).

ρ	p	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

terms of size control as indicated by the ARE values of the six tests. SK is generally very conservative with too small empirical sizes, especially when the correlation parameter ρ is large, showing that SK may yield misleading results when the variables are moderately or highly correlated. At the same time, CLX, GCBL$_m$, and GCBL$_l$ are generally too liberal with very large empirical sizes especially when ρ is large or the sample sizes are small, showing that they may also yield misleading results in such cases.

3.5. Simulation 5

In this simulation study, we compare L2N and L2D with the tests proposed by Srivastava and Du (2008), Srivastava, Katayama, and Kano (2013), and Feng et al. (2015), denoted as SD, SKK, and FZWZ, respectively. The data were generated as in Simulation 1 but here 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 variables are moderately or highly correlated (when $\rho = 0.5, 0.9$) although they are reasonable when the variables are less correlated (when $\rho = 0.1$). FZWZ is generally more liberal than L2N and L2D and its empirical size is generally around 7%.

4. Application to the Colon Data

In this section, we apply L2N, L2D, and BS to the colon dataset which is publicly available from *http://microarray.princeton.edu/oncology/affydata/index.html*. (We do not include other competitors mentioned in Section 3 to save space.) 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 when applied to this dataset. 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 3 and the conclusions drawn from Theorems 8 and 9. Notice that, although the three considered 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.

5. Concluding Remarks

We propose and study an L^2-norm based test with the W–S 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 automatically adapts to the changing shape of the null distribution of the test statistic, and hence it 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 such as spiked eigenvalue models considered by Katayama, Kanoa, and Srivastava (2013), Aoshima and Yata (2018), and others, nor it assumes restrictive conditions on the covariance matrix such as (8)–(10) imposed by Bai and Saranadasa (1996), Chen and Qin (2010), and Srivastava and Du (2008), 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 3 demonstrate that they are comparable in terms of size and power regardless whether the distributions of the two samples are normal or nonnormal. In terms of size control, they are always the winners when compared with 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 variables are moderately or highly correlated while our tests L2N and L2D perform

Table 8. Simulation 5: Empirical sizes (%).

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

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}	–	–

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 for computational speed comparisons. Therefore, both L2N and L2D have great potentials in applications to real-world problems, in particular in the new era of big data. Although we provide different formulas for estimation of the parameters in the W–S χ^2-approximation depending on whether the data are normal or nonnormal, giving L2N and L2D, we show both theoretically and numerically they are always comparable in all the considered simulation configurations. Therefore, we suggest simply using L2N (the version for normal data) in practice for its simplicity, interpretability, and superior computational speed. 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 (Chen, Zhang, and Zhong 2010; Zhang, Peng, and Wang 2013), testing regression means, and so on. Further studies in such directions are interesting and are ongoing; see Zhang, Guo, and Zhou (2017) for some details.

Appendix. Technical Proofs

Proof of Theorem 1. Set $\mathbf{w}_{n,p} = \sqrt{n_1 n_2/n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. 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 $\mathbf{w}_{n,p} \xrightarrow{\mathcal{L}} \mathbf{w}_p$ where $\mathbf{w}_p \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$. By the continuous mapping theorem, we have $T_{n,p,0} = \|\mathbf{w}_{n,p}\|^2 \xrightarrow{\mathcal{L}} T_{p,0}$ where $T_{p,0} = \|\mathbf{w}_p\|^2 \stackrel{d}{=} \sum_{r=1}^p \lambda_{p,r} A_r$. The expressions of the first three cumulants $\mathcal{K}_1 = \mathrm{E}(T_{p,0}), \mathcal{K}_2 = \mathrm{var}(T_{p,0})$ and $\mathcal{K}_3 = \mathrm{E}[T_{p,0} - \mathrm{E}(T_{p,0})]^3$ then follow immediately from Eq. (4) of Zhang (2005). □

Proof of Theorem 2. We first prove the first expression of (6). Set $\mathbf{w}_{n,p} = \sqrt{n_1 n_2/n}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. Then we have $\mathrm{E}(\mathbf{w}_{n,p}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{w}_{n,p}) = \Sigma$. We shall use the characteristic function ($\psi_X(t) = \mathrm{E}(e^{itX})$ for a random variable X) method. Let $\mathbf{u}_{p,1}, \ldots, \mathbf{u}_{p,p}$ denote the eigenvectors associated with the eigenvalues $\lambda_{p,1}, \ldots, \lambda_{p,p}$ of Σ in the descending order. We have $\mathbf{w}_{n,p} = \sum_{r=1}^p \xi_{n,p,r} \mathbf{u}_{p,r}$ where $\xi_{n,p,r} = \mathbf{w}_{n,p}^\top \mathbf{u}_{p,r}$. It is known that $\xi_{n,p,r}, r = 1, \ldots, p$, are uncorrelated, and $\mathrm{E}(\xi_{n,p,r}) = 0$ and $\mathrm{var}(\xi_{n,p,r}) = \lambda_{p,r}, r = 1, 2, \ldots$. Further, by the proofs of (20) and (21) given in the supplement, we have

$$\begin{aligned}\mathrm{var}(\xi_{n,p,r}^2) = 2\lambda_{p,r}^2 + (n_2/n)^2 \,[\mathrm{E}(\mathbf{x}_{11}^\top \mathbf{u}_{p,r})^4 - 3\lambda_{p,r}^2]/n_1 + (n_1/n)^2 \\ \times [\mathrm{E}(\mathbf{x}_{21}^\top \mathbf{u}_{p,r})^4 - 3\lambda_{p,r}^2]/n_2,\end{aligned}$$

where $\mathbf{x}_{i1} = \mathbf{y}_{i1} - \boldsymbol{\mu}_i, i = 1, 2$, as defined before. Under Assumptions 1–3, by some simple algebra we have $\mathrm{E}(\mathbf{x}_{i1}^\top \mathbf{u}_{p,r})^4 \le (3 + \Delta)\lambda_{p,r}^2, i = 1, 2$. Thus, we have

$$\mathrm{var}(\xi_{n,p,r}^2) \le \left[2 + (n_2/n)^2 \,\Delta/n_1 + (n_1/n)^2 \,\Delta/n_2\right]\lambda_{p,r}^2, \quad r = 1, 2, \ldots. \tag{A.1}$$

Note that $T_{n,p,0} = \sum_{r=1}^p \xi_{n,p,r}^2$. Set

$$\begin{aligned}\tilde{T}_{n,p,0} &= [T_{n,p,0} - \mathrm{tr}(\Sigma)]/\left[2\,\mathrm{tr}(\Sigma^2)\right]^{1/2} \\ &= \sum_{r=1}^p (\xi_{n,p,r}^2 - \lambda_{p,r})/\left[2\,\mathrm{tr}(\Sigma^2)\right]^{1/2},\end{aligned}$$

and $\tilde{T}_{n,p,0}^q = \sum_{r=1}^q (\xi_{n,p,r}^2 - \lambda_{p,r})/\left[2\,\mathrm{tr}(\Sigma^2)\right]^{1/2}$. We have $|\psi_{\tilde{T}_{n,p,0}}(t) - \psi_{\tilde{T}_{n,p,0}^q}(t)| \le |t| \left[\mathrm{E}(\tilde{T}_{n,p,0} - \tilde{T}_{n,p,0}^q)^2\right]^{1/2}$. Note that

$$\begin{aligned}\mathrm{E}\left(\tilde{T}_{n,p,0} - \tilde{T}_{n,p,0}^q\right)^2 &= \mathrm{E}\left(\sum_{r=q+1}^p (\xi_{n,p,r}^2 - \lambda_{p,r})/\left[2\,\mathrm{tr}(\Sigma^2)\right]^{1/2}\right)^2 \\ &= \mathrm{var}\left(\sum_{r=q+1}^p \xi_{n,p,r}^2\right)/\left[2\,\mathrm{tr}(\Sigma^2)\right] \\ &\le \left[\sum_{r=q+1}^p \sqrt{\mathrm{var}(\xi_{n,p,r}^2)}\right]^2/\left[2\,\mathrm{tr}(\Sigma^2)\right].\end{aligned}$$

By (A.1), we have

$$\begin{aligned}&\left[\sum_{r=q+1}^p \sqrt{\mathrm{var}(\xi_{n,p,r}^2)}\right]^2/\left[2\,\mathrm{tr}(\Sigma^2)\right] \\ &\quad \le \left[1 + (n_2/n)^2 \,\Delta/(2n_1) + (n_1/n)^2 \,\Delta/(2n_2)\right]\left[\sum_{r=q+1}^p \rho_{p,r}\right]^2.\end{aligned}$$

It follows that

$$\begin{aligned}&|\psi_{\tilde{T}_{n,p,0}}(t) - \psi_{\tilde{T}_{n,p,0}^q}(t)| \\ &\quad \le |t| \left[1 + (n_2/n)^2 \,\Delta/(2n_1) + (n_1/n)^2 \,\Delta/(2n_2)\right]^{1/2}\left[\sum_{r=q+1}^p \rho_{p,r}\right].\end{aligned}$$

Let t be fixed. By Assumption 4, for any fixed q, as $p \to \infty$, we have $\sum_{r=1}^\infty \rho_{\infty,r} < \infty$ and

$$\begin{aligned}\sum_{r=q+1}^p \rho_{p,r} = \sum_{r=1}^p \rho_{p,r} - \sum_{r=1}^q \rho_{p,r} \to \sum_{r=1}^\infty \rho_{\infty,r} \\ - \sum_{r=1}^q \rho_{\infty,r} \to \sum_{r=q+1}^\infty \rho_{\infty,r}.\end{aligned}$$

By letting $q \to \infty$, we further have $\sum_{r=q+1}^\infty \rho_{\infty,r} \to 0$. Thus, for any given $\epsilon > 0$, there exist P_1, Q_1 and N_1, depending on t and ϵ, such that for any $p \ge P_1, q \ge Q_1$ and $n \ge N_1$, we have

$$|\psi_{\tilde{T}_{n,p,0}}(t) - \psi_{\tilde{T}_{n,p,0}^q}(t)| \le \epsilon. \tag{A.2}$$

Note that we always have $p \ge q$. For any fixed $p \ge P_1, q \ge Q_1$, by the central limit theorem, it is easy to show that as $n \to \infty$, we have $\tilde{T}_{n,p,0}^q \xrightarrow{\mathcal{L}} \tilde{T}_{p,0}^q$ where $\tilde{T}_{p,0}^q = \sum_{r=1}^q \rho_{p,r}(A_r - 1)/\sqrt{2}$ since as $n \to \infty$, $\xi_{n,p,r} \xrightarrow{\mathcal{L}} N(0, \lambda_{p,r})$ and $\xi_{n,p,r}$'s are asymptotically independent for $r = 1, 2, \ldots, q$. That is, under Assumption 3, there exists N_2, depending on p, q, t and ϵ, such that for any $n \ge N_2$ we have

$$\left|\psi_{\tilde{T}_{n,p,0}^q}(t) - \psi_{\tilde{T}_{p,0}^q}(t)\right| \le \epsilon. \tag{A.3}$$

Recall that $\zeta = \sum_{r=1}^\infty \rho_{\infty,r}(A_r - 1)/\sqrt{2}$. Set $\zeta^q = \sum_{r=1}^q \rho_{\infty,r}(A_r - 1)/\sqrt{2}$. Then, under Assumption 4, for any fixed q, as $p \to \infty$, we have $\tilde{T}_{p,0}^q \xrightarrow{\mathcal{L}} \zeta^q$. That is, there exists a P_2, depending on q, t and ϵ, such that for any $p \ge P_2$ we have

$$\left|\psi_{\tilde{T}_{p,0}^q}(t) - \psi_{\zeta^q}(t)\right| \le \epsilon. \tag{A.4}$$

Furthermore, we have

$$\left|\psi_{\zeta^q}(t) - \psi_{\zeta}(t)\right| \le |t| \left[\mathrm{E}\left(\sum_{r=q+1}^\infty \rho_{\infty,r}(A_r - 1)/\sqrt{2}\right)^2\right]^{1/2}$$

$$\leq |t|\left[\operatorname{var}\left(\sum_{r=q+1}^{\infty}\rho_{\infty,r}(A_r-1)/\sqrt{2}\right)\right]^{1/2}$$
$$=|t|\left[\sum_{r=q+1}^{\infty}\rho_{\infty,r}^2\right]^{1/2}\leq |t|\left(\sum_{r=q+1}^{\infty}\rho_{\infty,r}\right),$$

which, under Assumption 4, tends to 0 as $q\to\infty$. Thus, there exists Q_2, depending on t and ϵ, such that for any $q\geq Q_2$ we have

$$\left|\psi_{\zeta^q}(t)-\psi_{\zeta}(t)\right|\leq\epsilon. \tag{A.5}$$

It follows from (A.2)–(A.5) that for any $n\geq\max(N_1,N_2)$, $p\geq\max(P_1,P_2)$ and $q\geq\max(Q_1,Q_2)$ we have

$$\begin{aligned}&\left|\psi_{\tilde{T}_{n,p,0}}(t)-\psi_{\zeta}(t)\right|\\&\leq\left|\psi_{\tilde{T}_{n,p,0}}(t)-\psi_{\tilde{T}^q_{n,p,0}}(t)\right|+\left|\psi_{\tilde{T}^q_{n,p,0}}(t)-\psi_{\tilde{T}^q_{p,0}}(t)\right|\\&+\left|\psi_{\tilde{T}^q_{p,0}}(t)-\psi_{\zeta^q}(t)\right|+\left|\psi_{\zeta^q}(t)-\psi_{\zeta}(t)\right|\leq 4\epsilon.\end{aligned}$$

The convergence in distribution of $\tilde{T}_{n,p,0}$ to ζ given in (6) follows as we can let $\epsilon\to 0$.

To show the second convergence in distribution result given in (6), set $\tilde{T}_{p,0}=[T_{p,0}-\operatorname{tr}(\Sigma)]/\left[2\operatorname{tr}(\Sigma^2)\right]^{1/2}=\sum_{r=1}^{p}\rho_{p,r}(A_r-1)/\sqrt{2}$. Using the definition of $\tilde{T}^q_{p,0}$ defined earlier, we have

$$\left|\psi_{\tilde{T}_{p,0}}(t)-\psi_{\tilde{T}^q_{p,0}}(t)\right|\leq|t|\left[\operatorname{E}(\tilde{T}_{p,0}-\tilde{T}^q_{p,0})^2\right]^{1/2}$$
$$=|t|\left[\operatorname{E}\left(\sum_{r=q+1}^{p}\rho_{p,r}(A_r-1)/\sqrt{2}\right)^2\right]^{1/2}$$
$$=|t|\left[\sum_{r=q+1}^{p}\rho_{p,r}^2\right]^{1/2}\leq|t|\sum_{r=q+1}^{p}\rho_{p,r}.$$

Then according to the proof of (A.3), under Assumption 4, for any given $\epsilon>0$, there exist P_3 and Q_3, depending on t and ϵ, such that for any $p\geq P_3$ and $q\geq Q_3$ we have

$$\left|\psi_{\tilde{T}_{p,0}}(t)-\psi_{\tilde{T}^q_{p,0}}(t)\right|\leq\epsilon. \tag{A.6}$$

It then follows from (A.4)–(A.6) that as $p\geq\max(P_2,P_3)$ and $q\geq\max(Q_2,Q_3)$, we have

$$\begin{aligned}\left|\psi_{\tilde{T}_{p,0}}(t)-\psi_{\zeta}(t)\right|&\leq\left|\psi_{\tilde{T}_{p,0}}(t)-\psi_{\tilde{T}^q_{p,0}}(t)\right|\\&+\left|\psi_{\tilde{T}^q_{p,0}}(t)-\psi_{\zeta^q}(t)\right|+\left|\psi_{\zeta^q}(t)-\psi_{\zeta}(t)\right|\leq 3\epsilon.\end{aligned}$$

The convergence in distribution of $\tilde{T}_{p,0}$ to ζ given in (6) then follows immediately.

Set $\tilde{x}=[x-\operatorname{tr}(\Sigma)]/\left[2\operatorname{tr}(\Sigma^2)\right]^{1/2}$ for any real number x. Since the limit ζ is a continuous random variable, by Lemma 2.11 of van der Vaart (1998), the uniform convergence result given in (7) follows directly from the convergence in distribution of both $\tilde{T}_{n,p,0}$ and $\tilde{T}_{p,0}$ to ζ and the triangular inequality

$$\begin{aligned}&\sup_x\left|\Pr(T_{n,p,0}\leq x)-\Pr(T_{p,0}\leq x)\right|\\&=\sup_x\left|\Pr(\tilde{T}_{n,p,0}\leq\tilde{x})-\Pr(\tilde{T}_{p,0}\leq\tilde{x})\right|\\&\leq\sup_x\left|\Pr(\tilde{T}_{n,p,0}\leq\tilde{x})-\Pr(\zeta\leq\tilde{x})\right|\\&+\sup_x\left|\Pr(\tilde{T}_{p,0}\leq\tilde{x})-\Pr(\zeta\leq\tilde{x})\right|\to 0\quad\text{as } n,p\to\infty.\end{aligned}\tag{A.7}$$

□

Proof of Theorem 3. Since all the eigenvalues of Σ are nonnegative, $\operatorname{tr}(\Sigma^3)=\sum_{r=1}^{p}\lambda_{p,r}^3\leq\lambda_{p,\max}\sum_{r=1}^{p}\lambda_{p,r}^2=\lambda_{p,\max}\operatorname{tr}(\Sigma^2)$ and hence we have $d^*=\operatorname{tr}^3(\Sigma^2)/\operatorname{tr}^2(\Sigma^3)\geq[\lambda_{p,\max}^2/\operatorname{tr}(\Sigma^2)]^{-1}$ for d^* defined in (15). Next, by the Cauchy–Schwarz inequality, we have $\operatorname{tr}^2\left(\Sigma^3\right)\leq\operatorname{tr}\left(\Sigma^4\right)\operatorname{tr}\left(\Sigma^2\right)$ and hence

$$d^*\geq\left[\operatorname{tr}(\Sigma^4)/\operatorname{tr}^2(\Sigma^2)\right]^{-1}. \tag{A.8}$$

Besides, we can write $d^*=p[\operatorname{tr}\left(\Sigma^2\right)/p]^3/[\operatorname{tr}\left(\Sigma^3\right)/p]^2$. Therefore, any of the conditions (8)–(10) implies $d^*\to\infty$ as $p\to\infty$. Then, it follows from Theorem 4 that $T_{p,0}$ tends to normal as $p\to\infty$ and the second convergence in distribution result in (11) follows immediately. In addition, since $d^*\to\infty$ as $p\to\infty$, we have $\lambda_{p,\max}^2/\operatorname{tr}(\Sigma^2)=(\lambda_{p,\max}^3)^{2/3}/\operatorname{tr}(\Sigma^2)\leq\left[\operatorname{tr}(\Sigma^3)\right]^{2/3}/\operatorname{tr}(\Sigma^2)=d^{*-1/3}\to 0$ and thus condition (8) holds. Together with Assumptions 1–3, implies the first convergence in distribution result given in (11) as established by Appendix A.2 of Bai and Saranadasa (1996). Now, the uniform convergence result given in (7) follows immediately from the convergence in distribution results given in (11), Lemma 2.11 of van der Vaart (1998), and the triangular inequality (A.7) with ζ replaced by a $N(0,1)$ random variable. □

Proof of Theorem 4. By definition (5), $T_{p,0}$ is a central χ^2-type mixture with all of the coefficients nonnegative. The theorem then follows immediately from Lemma S.1 of the supplement with d^* as defined in (15). □

Proof of Theorem 5. Recall that $\lambda_{p,r}, r=1,\ldots,p$ denote eigenvalues of Σ in the descending order. Since $\lambda_{p,r}, r=1,\ldots,p$, are all nonnegative,

$$\operatorname{tr}^2(\Sigma^k)=\left(\sum_{r=1}^{p}\lambda_{p,r}^k\right)^2\geq\sum_{r=1}^{p}\lambda_{p,r}^{2k}=\operatorname{tr}(\Sigma^{2k}),\quad k=1,2,\ldots. \tag{A.9}$$

It follows that $\operatorname{tr}^2(\Sigma^k)\geq\operatorname{tr}(\Sigma^{2k}), k=1,2,\ldots$. Together with (A.8), this shows that $d^*\geq\operatorname{tr}^2(\Sigma^2)/\operatorname{tr}(\Sigma^4)\geq 1$. By the Cauchy–Schwarz inequality, we have $\operatorname{tr}^2\left(\Sigma^2\right)\leq\operatorname{tr}\left(\Sigma^3\right)\operatorname{tr}\left(\Sigma\right)$. Thus,

$$d^*=\operatorname{tr}^3\left(\Sigma^2\right)/\operatorname{tr}^2\left(\Sigma^3\right)\leq\operatorname{tr}^3\left(\Sigma^2\right)/\left[\operatorname{tr}^4\left(\Sigma^2\right)/\operatorname{tr}^2\left(\Sigma\right)\right]=d. \tag{A.10}$$

To show the last inequality of (a), note that $(p^{-1}\sum_{r=1}^{p}\lambda_{p,r})^2\leq p^{-1}\sum_{r=1}^{p}\lambda_{p,r}^2$. This implies that $\left(\sum_{r=1}^{p}\lambda_{p,r}\right)^2\leq p\left(\sum_{r=1}^{p}\lambda_{p,r}^2\right)$. That is, $\operatorname{tr}^2(\Sigma)\leq p\operatorname{tr}(\Sigma^2)$. Therefore, we have $d\leq p$. Assertion (a) is now proved.

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

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

Proof of Corollary 1. Under the given conditions, as $p\to\infty$, we have $d^*=[\operatorname{tr}(\Sigma^2)/p]^3/[\operatorname{tr}(\Sigma^3)/p]^2\to b_2^3/b_3^2\in(0,\infty)$. The corollary then follows directly from Theorem 4. □

Proof of Corollary 2. By (a) of Theorem 5, when d is bounded, so is d^* so that the distribution of R does not tend to normal. By Theorem 4, the distribution of $T_{p,0}$ also does not tend to normal. On the other hand, when $d^* \to \infty$, we have $d \to \infty$ so that $[R - \operatorname{tr}(\Sigma)]/[2\operatorname{tr}(\Sigma^2)]^{1/2} = (\chi_d^2 - d)/(2d)^{1/2} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, and asymptotic normality of $T_{p,0}$ given in (11) follows directly from Theorem 4. □

Proof of Theorem 6. This theorem is a special case of Lemma S.2 given in the supplement when $\nu = 1$ and $\lambda_{p,1}, \ldots, \lambda_{p,p}$ are taken as the eigenvalues of Σ. □

Proof of Theorem 7. When the two samples (1) are normal, by the definitions of $\hat{\beta}$ and $\hat{d}$ given in (18) and Lemma S.3 of the supplement, as $n \to \infty$, we have $\hat{\beta}/\beta = [\widehat{\operatorname{tr}(\Sigma^2)}/\operatorname{tr}(\Sigma^2)]/[\operatorname{tr}(\hat{\Sigma})/\operatorname{tr}(\Sigma)] \xrightarrow{P} 1$ and $\hat{d}/d = [\widehat{\operatorname{tr}^2(\Sigma)}/\operatorname{tr}^2(\Sigma)]/[\widehat{\operatorname{tr}(\Sigma^2)}/\operatorname{tr}(\Sigma^2)] \xrightarrow{P} 1$ uniformly in p. It follows $\hat{\beta}\chi_{\hat{d}}^2(\alpha)$ is ratio-consistent for $\beta\chi_d^2(\alpha)$ uniformly in p. □

Proof of Theorem 8. Under Assumptions 1–3, by some simple algebra we have $\kappa_{i,11} = \Delta \sum_{r=1}^p \gamma_r^2, i = 1, 2$ where γ_r is the rth diagonal entry of $\Gamma^\top \Gamma$. It follows that $\kappa_{i,11} = O[\operatorname{tr}(\Sigma^2)], i = 1, 2$, since $\Delta < \infty$ and $\sum_{r=1}^p \gamma_r^2 \le \operatorname{tr}(\Sigma^2)$. Therefore, as $n \to \infty$, we have $\delta = o[\operatorname{tr}(\Sigma^2)]$, $\beta = \frac{\operatorname{tr}(\Sigma^2)+\delta/2}{\operatorname{tr}(\Sigma)} = \frac{\operatorname{tr}(\Sigma^2)}{\operatorname{tr}(\Sigma)}[1+o(1)]$, and $d = \frac{\operatorname{tr}^2(\Sigma)}{\operatorname{tr}(\Sigma^2)+\delta/2} = \frac{\operatorname{tr}^2(\Sigma)}{\operatorname{tr}(\Sigma^2)}[1+o(1)]$. □

Proof of Theorem 9. We first prove that $\operatorname{tr}(\hat{\Sigma}), \widehat{\operatorname{tr}^2(\Sigma)}$ and $\widehat{\operatorname{tr}(\Sigma^2)}$ defined in (17) are ratio-consistent for $\operatorname{tr}(\Sigma), \operatorname{tr}^2(\Sigma)$ and $\operatorname{tr}(\Sigma^2)$ based on the proof of the ratio-consistency of $\widehat{\operatorname{tr}(\Sigma^2)}$ under Assumptions 1–3 given in Appendix A.3 of Bai and Saranadasa (1996). Carefully studying Appendix A.3 of Bai and Saranadasa (1996), we have $\mathrm{E}\left[\operatorname{tr}(\hat{\Sigma})\right] = \operatorname{tr}(\Sigma)$ and $\operatorname{var}\left[\operatorname{tr}(\hat{\Sigma})\right] = 2n^{-1}\operatorname{tr}(\Sigma^2)[1 + o(1)]$. It follows that $\operatorname{var}\left[\operatorname{tr}(\hat{\Sigma})/\operatorname{tr}(\Sigma)\right] = 2n^{-1}\operatorname{tr}(\Sigma^2)/\operatorname{tr}^2(\Sigma)[1 + o(1)] \le 2n^{-1}[1 + o(1)] \to 0$ uniformly in p. Similarly, from Appendix A.3 of Bai and Saranadasa (1996), we have $\mathrm{E}\left[\widehat{\operatorname{tr}(\Sigma^2)}\right] = \operatorname{tr}(\Sigma^2)[1 + o(1)]$ and $\operatorname{var}\left[\widehat{\operatorname{tr}(\Sigma^2)}\right] = 4n^{-1}\left[2\operatorname{tr}(\Sigma^4) + \operatorname{tr}^2(\Sigma^2)/n\right][1 + o(1)]$. Thus, $\operatorname{var}\left[\widehat{\operatorname{tr}(\Sigma^2)}/\operatorname{tr}(\Sigma^2)\right] = 4n^{-1}\left[2\operatorname{tr}(\Sigma^4)/\operatorname{tr}^2(\Sigma^2) + n^{-1}\right][1 + o(1)] \le 8n^{-1}[1 + o(1)] \to 0$ uniformly in p. Therefore, $\operatorname{tr}(\hat{\Sigma})$ and $\widehat{\operatorname{tr}(\Sigma^2)}$ are ratio-consistent for $\operatorname{tr}(\Sigma)$ and $\operatorname{tr}(\Sigma^2)$ uniformly in p. We now show that $\widehat{\operatorname{tr}^2(\Sigma)}$ is also ratio-consistent for $\operatorname{tr}^2(\Sigma)$ uniformly in p. Based on the definition of $\widehat{\operatorname{tr}^2(\Sigma)}$ given in (17), we can re-express $\widehat{\operatorname{tr}^2(\Sigma)}$ in terms of $\operatorname{tr}(\hat{\Sigma})$ and $\widehat{\operatorname{tr}(\Sigma^2)}$ as $\widehat{\operatorname{tr}^2(\Sigma)} = \operatorname{tr}^2(\hat{\Sigma}) - 2(n-2)^{-1}\widehat{\operatorname{tr}(\Sigma^2)}$. It follows that, as $n \to \infty$, $\widehat{\operatorname{tr}^2(\Sigma)}/\operatorname{tr}^2(\Sigma) = \left[\operatorname{tr}(\hat{\Sigma})/\operatorname{tr}(\Sigma)\right]^2 - 2(n-2)^{-1}\left[\widehat{\operatorname{tr}(\Sigma^2)}/\operatorname{tr}^2(\Sigma)\right] = 1 + o_p(1)$ uniformly in p. That is, $\widehat{\operatorname{tr}^2(\Sigma)}$ is also ratio-consistent for $\operatorname{tr}^2(\Sigma)$ uniformly in p. The ratio-consistency results given in (19) then follow immediately. □

Proof of Theorem 10. Under Assumptions 1–3 and any of the conditions (8)–(10), by Theorems 3 and 9, we have (a) both $T_{n,p,0}$ and $T_{p,0}$ are asymptotically normal, (b) $\hat{\beta}$ and $\hat{d}$ are ratio-consistent for β and d uniformly in p, and (c) both d^* and d tend to infinity as $p \to \infty$. It follows that

$$
\begin{aligned}
\Pr\left[T_{n,p} \ge \hat{\beta}\chi_{\hat{d}}^2(\alpha)\right] &= \Pr\left[T_{n,p,0} \ge \beta\chi_d^2(\alpha) - \|\boldsymbol{\mu}_{n,p}\|^2\right] + o(1) \\
&= \Pr\left\{\frac{T_{n,p,0} - \operatorname{tr}(\Sigma)}{[2\operatorname{tr}(\Sigma^2)]^{1/2}} \ge \frac{\beta\chi_d^2(\alpha) - \operatorname{tr}(\Sigma)}{[2\operatorname{tr}(\Sigma^2)]^{1/2}} - \frac{n\tau(1-\tau)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{[2\operatorname{tr}(\Sigma^2)]^{1/2}}\right\} \\
&\quad + o(1) \\
&= \Phi\left\{-z_\alpha + n\tau(1-\tau)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2/[2\operatorname{tr}(\Sigma^2)]^{1/2}\right\} + o(1),
\end{aligned}
$$

where we use the fact that, as $p \to \infty$, $\left[\beta\chi_d^2(\alpha) - \operatorname{tr}(\Sigma)\right]/\left[2\operatorname{tr}(\Sigma^2)\right]^{1/2} = \left[\chi_d^2(\alpha) - d\right]/\sqrt{2d} \to z_\alpha$. □

Supplementary Materials

Histograms of simulated T_{BS}, some lemmas and their proofs, and proofs of (20) and (21).

Acknowledgments

The authors thank the co-editor, AE, and two reviewers for their constructive comments and suggestions which help us improve the article substantially.

Funding

Zhang is supported by the National University of Singapore academic research grant R-155-000-187-114. Zhou is financially supported by the First Class Discipline of Zhejiang—A(Zhejiang Gongshang University—Statistics). Cheng is supported by the Hong Kong Baptist University grants RC-ICRS17-18 and FRG2/17-18/086. Guo would like to thank Professor Wen-Lung Shiau, the Advanced Data Analysis Center (PLS-SEM of Zhejiang University of Technology), for his support on this research.

References

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis* (3rd ed.), New York: Wiley. [1]

Aoshima, M., and Yata, K. (2018), "Two-Sample Tests for High-Dimension, Strongly Spiked Eigenvalue Models," *Statistica Sinica*, 28, 43–62. [3,13]

Bai, Z., and Saranadasa, H. (1996), "Effect of High Dimension: By an Example of a Two Sample Problem," *Statistica Sinica*, 6, 311–329. [1,2,3,4,5,6,8,13,15,16]

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," *The Annals of Mathematical Statistics*, 25, 290–302. [2,4]

Cai, T. T., Liu, W., and Xia, Y. (2014), "Two-Sample Test of High Dimensional Means Under Dependence," *Journal of the Royal Statistical Society*, Series B, 76, 349–372. [11]

Chen, S., Zhang, L., and Zhong, P. (2010), "Tests for High-Dimensional Covariance Matrices," *Journal of the American Statistical Association*, 105, 810–819. [14]

Chen, S. X., and Qin, Y.-L. (2010), "A Two-Sample Test for High-Dimensional Data With Applications to Gene-Set Testing," *The Annals of Statistics*, 38, 808–835. [2,4,5,6,8,11,13]

Cheng, M.-Y., and Wu, H.-T. (2013), "Local Linear Regression on Manifolds and Its Geometric Interpretation," *Journal of the American Statistical Association*, 108, 1421–1434. [1]

Chuang, L.-L., and Shih, Y.-S. (2012), "Approximated Distributions of the Weighted Sum of Correlated Chi-Squared Random Variables," *Journal of Statistical Planning and Inference*, 142, 457–472. [6]

Cook, R. D., and Li, B. (2002), "Dimension Reduction for the Conditional Mean in Regression," *The Annals of Statistics*, 30, 455–474. [1]

Cui, H., Li, R., and Zhong, W. (2015), "Model-Free Feature Screening for Ultra-High Dimensional Discriminant Analysis," *Journal of the American Statistical Association*, 110, 630–641. [1]

Dempster, A. P. (1958), "A High Dimensional Two Sample Significance Test," *The Annals of Mathematical Statistics*, 29, 995–1010. [2]

——— (1960), "A Significance Test for the Separation of Two Highly Multivariate Small Samples," *Biometrics*, 16, 41–50. [2]

Fan, J., and Fan, Y. (2008), "High Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637. [1]

Fan, J., Feng, Y., and Tong, X. (2012), "A Road to Classification in High Dimensional Space: The Regularized Optimal Affine Discriminant," *Journal of the Royal Statistical Society*, Series B, 74, 745–771. [1]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [1]

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961. [1]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1]

Fan, Y., Kong, Y., Li, D., and Zheng, Z. (2015), "Innovated Interaction Screening for High-Dimensional Nonlinear Classification," *The Annals of Statistics*, 43, 1243–1272. [1]

Feiveson, A. H., and Delaney, F. C. (1968), *The Distribution and Properties of a Weighted Sum of Chi Squares* (Vol. 4575), Washington, DC: National Aeronautics and Space Administration. [2]

Feng, L., Zou, C., Wang, Z., and Zhu, L. (2015), "Two Sample Behrens–Fisher Problem for High-Dimensional Data," *Statistica Sinica*, 25, 1297–1312. [13]

Gregory, K. B., Carroll, R. J., Baladandayuthapani, V., and Lahiri, S. N. (2015), "A Two-Sample Test for Equality of Means in High Dimension," *Journal of the American Statistical Association*, 110, 837–849. [11]

Himeno, T., and Yamada, T. (2014), "Estimations for Some Functions of Covariance Matrix in High Dimension Under Non-Normality and Its Applications," *Journal of Multivariate Analysis*, 130, 27–44. [7]

Hotelling, H. (1931), "The Generalization of Student's Ratio," *The Annals of Mathematical Statistics*, 2, 360–378. [1]

Imhof, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 419–426. [4]

Katayama, S., Kanoa, Y., and Srivastava, M. S. (2013), "Asymptotic Distributions of Some Test Criteria for the Mean Vector With Fewer Observations Than the Dimension," *Journal of Multivariate Analysis*, 116, 410–421. [3,13]

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [1]

Lv, J., and Fan, Y. (2009), "A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares," *The Annals of Statistics*, 37, 3498–3528. [1]

Pauly, M., Ellenberger, D., and Brunner, E. (2015), "Analysis of High-Dimensional One Group Repeated Measures Designs," *Statistics*, 49, 1243–1261. [3]

Satterthwaite, F. E. (1941), "Synthesis of Variance," *Psychometrika*, 6, 309–316. [2]

——— (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114. [2]

Srivastava, M. S., and Du, M. (2008), "A Test for the Mean Vector With Fewer Observations Than the Dimension," *Journal of Multivariate Analysis*, 99, 386–402. [2,3,4,5,6,13]

Srivastava, M. S., Katayama, S., and Kano, Y. (2013), "A Two Sample Test in High Dimensional Data," *Journal of Multivariate Analysis*, 114, 349–358. [11,13]

Srivastava, M. S., and Kubokawa, T. (2013), "Tests for Multivariate Analysis of Variance in High Dimension Under Non-Normality," *Journal of Multivariate Analysis*, 115, 204–216. [11]

Srivastava, R., Li, P., and Ruppert, D. (2015), "RAPTT: An Exact Two-Sample Test in High Dimensions Using Random Projections," *Journal of Computational and Graphical Statistics*, 25, 954–970. [2]

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, UK: Cambridge University Press. [15]

Wang, L., Peng, B., and Li, R. (2015), "A High-Dimensional Nonparametric Multivariate Test for Mean Vector," *Journal of the American Statistical Association*, 110, 1658–1669. [2]

Wang, R., Peng, L., and Qi, Y. (2013), "Jackknife Empirical Likelihood Test for Equality of Two High Dimensional Means," *Statistica Sinica*, 23, 667–690. [2]

Welch, B. L. (1947), "The Generalization of 'Student's' Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28–35. [2]

Witten, D., and Tibshirani, R. (2011), "Penalized Classification Using Fisher's Linear Discriminant," *Journal of the Royal Statistical Society*, Series B, 73, 753–772. [1]

Xia, Y. (2007), "A Constructive Approach to the Estimation of Dimension Reduction Directions," *The Annals of Statistics*, 35, 2654–2690. [1]

——— (2008), "A Multiple-Index Model and Dimension Reduction," *Journal of the American Statistical Association*, 103, 1631–1640. [1]

Yamada, T., and Himeno, T. (2015), "Testing Homogeneity of Mean Vectors Under Heteroscedasticity in High-Dimension," *Journal of Multivariate Analysis*, 139, 7–27. [11]

Zhang, J.-T. (2005), "Approximate and Asymptotic Distributions of Chi-Square-Type Mixtures With Applications," *Journal of the American Statistical Association*, 100, 273–285. [2,3,6,14]

Zhang, J.-T., Guo, J., and Zhou, B. (2017), "Linear Hypothesis Testing in High-Dimensional One-Way MANOVA," *Journal of Multivariate Analysis*, 155, 200–216. [3,14]

Zhang, R., Peng, L., and Wang, R. (2013), "Tests for Covariance Matrix With Fixed or Divergent Dimension," *The Annals of Statistics*, 41, 2075–2096. [14]