Ann Inst Stat Math (2014) 66:33–61
DOI 10.1007/s10463-013-0404-2

A U-statistic approach for a high-dimensional two-sample mean testing problem under non-normality and Behrens–Fisher setting

M. Rauf Ahmad

Received: 11 July 2012 / Revised: 12 February 2013 / Published online: 20 April 2013

Abstract A two-sample test statistic is presented for testing the equality of mean vectors when the dimension, p, exceeds the sample sizes, $n_i, \ i = 1, 2$, and the distributions are not necessarily normal. Under mild assumptions on the traces of the covariance matrices, the statistic is shown to be asymptotically Chi-square distributed when $n_i, p \to \infty$. However, the validity of the test statistic when p is fixed but large, including $p > n_i$, and when the distributions are multivariate normal, is shown as special cases. This two-sample Chi-square approximation helps us establish the validity of Box's approximation for high-dimensional and non-normal data to a two-sample setup, valid even under Behrens–Fisher setting. The limiting Chi-square distribution of the statistic is obtained using the asymptotic theory of degenerate U-statistics, and using a result from classical asymptotic theory, it is further extended to an approximate normal distribution. Both independent and paired-sample cases are considered.

Keywords High-dimensional multivariate inference · Box's approximation · Behrens–Fisher setting · Degenerate U-statistics

M. R. Ahmad (✉)
Department of Statistics, Uppsala University, Ekonomikum, Kyrkogårdsgatan 10, 75120 Uppsala, Sweden
e-mail: rauf.ahmad@statistik.uu.se; rauf.ahmad@slu.se

M. R. Ahmad
Department of Energy and Technology, Swedish University of Agricultural Sciences, Gerda Nilssons väg, 5, 75651 Uppsala, Sweden

1 Introduction

We consider a test statistic for testing the difference of mean vectors of two independent multivariate distributions. Let

$$\mathbf{X}_{1k} = (X_{11k}, \ldots, X_{1pk})' \quad \text{and} \quad \mathbf{X}_{2l} = (X_{21l}, \ldots, X_{2pl})',$$

$k = 1, \ldots, n_1, l = 1, \ldots, n_2$, be independent, identically distributed random vectors, with

$$\mathbf{X}_{1k} \sim \mathcal{F}_1 \quad \text{and} \quad \mathbf{X}_{2l} \sim \mathcal{F}_2, \tag{1}$$

where $\mathcal{F}_1$ and $\mathcal{F}_2$ denote the distribution functions. Assume $\mathrm{E}(\mathbf{X}_{1k}) = \boldsymbol{\mu}_1$, $\mathrm{E}(\mathbf{X}_{2l}) = \boldsymbol{\mu}_2$, $\mathrm{Cov}(\mathbf{X}_{1k}) = \boldsymbol{\Sigma}_1$, and $\mathrm{Cov}(\mathbf{X}_{2l}) = \boldsymbol{\Sigma}_2$, where $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 > 0$. We are interested to test the hypothesis $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$, versus H_1 : Not H_0, when $p > n_i, i = 1, 2$.

Clearly, the classical multivariate test statistics, like Hotelling's T^2, cannot be used when $p > n$ since the estimated covariance matrix is singular and hence cannot be inverted. Dempster (1958) is perhaps the oldest reference dealing specifically with a two-sample test statistic valid for high-dimensional setup when $\mathcal{F}_1, \mathcal{F}_2$ are p-variate normal distributions, i.e., when $\mathbf{X}_{1k} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{X}_{2l} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$; see also Dempster (1969, Chapters 7, 10). Replacing normality assumption with certain assumptions on the moments of the underlying multivariate model, and additionally a few assumptions on the traces of the covariance matrices, a two-sample test is presented in Bai and Saranadasa (1996). Following the same multivariate model as introduced by Bai and Saranadasa, but under different assumptions on the traces of the covariance matrices, a modification of their test statistic has recently been considered by Chen and Qin (2010). Again under normality, a modified version of Hotelling's T^2, using the Moore–Penrose inverse of estimated covariance matrix in place of the regular inverse, is proposed by Srivastava (2007); see also Srivastava (2009). For a short literature review on similar one-sample tests, see Ahmad et al. (2012a).

Continuing with the normality assumption, Ahmad (2008, Chapter 3) gives a test statistic for general linear hypothesis of the form of H_0, but specifically designed to test the profile (interaction and time effects) hypotheses. The statistic is constructed using quadratic and bilinear forms composed of the differences (for interaction effect) and sums (for time effect) of the vectors $\mathbf{X}_{1k}$ and $\mathbf{X}_{2l}$. Assuming p fixed and $n_1, n_2 \to \infty$, including the case when $p > n_i,\ i = 1, 2$, it is shown that the test statistic, asymptotically, follows a scaled Chi-square distribution, where the distributional convergence is based on the Box's approximation (see Sect. 2 below). In simulation studies, with a variety of parameter settings, the accuracy of the test statistic is demonstrated for both size control and power, inclusive of the case when the dimension far exceeds the sample size.

In this paper, we present a modified version of this test statistic, and evaluate it to test $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$. The modification is aimed at extending the original test in a variety of directions: (1) the modified test does not require $\mathcal{F}_1, \mathcal{F}_2$ to be necessarily normal. The assumption of normality is replaced with certain mild and practically viable assumptions on the traces of the covariance matrices. (2) A serious limitation of the original test is that its derivation requires either $n_1 = n_2$ or $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. The modified version, however, does not impose any such condition on the sample sizes or covariance matrices. This indirectly helps us establish the validity of Box's approximation for the Berhens–Fisher problem. (3) The paired case is also considered. While essentially reducing to a single-sample case, the paired case verifies the results of a similar one-sample test statistic presented in Ahmad et al. (2012a). (4) Finally, the asymptotics: while the original test is developed keeping p fixed and letting $n_i \to \infty$, $i = 1, 2$, the modified statistic is so constructed that it is also valid under standard high-dimensional asymptotics, i.e., when both n_i and $p \to \infty$. The results, however, remain valid under the special cases of fixed p and normality (see Sect. 4). Further, we use the asymptotic theory of degenerate U-statistics to derive the limiting null distributions of the test statistics. With the use of U-statistics theory, our asymptotic approach strongly differs from the papers cited above, and is rather closer to the one followed by Gretton et al. (2008) for a similar two-sample problem, based on kernels, although not for high-dimensional data.

We begin with a brief review of the normal theory-based statistic in the next section, with the aim to introduce notations and set the stage for the modification. The modified test statistic is introduced in Sect. 3, along with its asymptotic distribution for both independent and paired cases. Section 4 gives a brief sketch of the validity of the results for the special cases when the distributions are multivariate normal and/or p is assumed fixed. Some special remarks are given in Sect. 5. Section 5.1 summarizes the results of Box's approximation and the validity of test statistics under different parameter settings. As the use of the theory of U-statistics, particularly the degenerate case, to tackle the problems of high-dimensional data, is relatively new, a brief motivating orientation to this issue is presented in Sect. 5.2. This motivation is primarily meant to make a point without delving deep into the rigorous mathematical aspects. A thorough mathematical treatment of the subject is postponed for another manuscript.

2 A brief review of the test under normality

Let $\mathbf{X}_{1k}$ and $\mathbf{X}_{2l}$ be as defined above, under model (1). Then

$$\mathbf{X} = \left(\mathbf{X}'_{11}, \ldots, \mathbf{X}'_{1n_1}, \mathbf{X}'_{21}, \ldots, \mathbf{X}'_{2n_2}\right)'$$

denotes the vector of all observations, with $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$, where

$$\boldsymbol{\mu} = \left(\mathbf{1}'_{n_1} \otimes \boldsymbol{\mu}'_1,\ \mathbf{1}'_{n_2} \otimes \boldsymbol{\mu}'_2\right)' \tag{2}$$

$$\boldsymbol{\Sigma} = \left(\mathbf{I}_{n_1} \otimes \boldsymbol{\Sigma}_1\right) \oplus \left(\mathbf{I}_{n_2} \otimes \boldsymbol{\Sigma}_2\right) \tag{3}$$

with the corresponding sample estimators,

$$\overline{\mathbf{X}} = \left(\mathbf{1}'_{n_1} \otimes \overline{\mathbf{X}}'_{1.},\ \mathbf{1}'_{n_2} \otimes \overline{\mathbf{X}}'_{2.}\right)' \tag{4}$$

$$\widehat{\boldsymbol{\Sigma}} = \left(\mathbf{I}_{n_1} \otimes \widehat{\boldsymbol{\Sigma}}_1\right) \oplus \left(\mathbf{I}_{n_2} \otimes \widehat{\boldsymbol{\Sigma}}_2\right). \tag{5}$$

Here, $\otimes$ denotes the Kronecker product, and $\oplus$ denotes the Kronecker sum. To test a general linear hypothesis of the form $\mathbf{T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$, where $\mathbf{T}$ is a hypothesis matrix, Ahmad (2008 Chapter 3) derives a test statistic, valid for any appropriately defined $\mathbf{T}$, including $\mathbf{T} = \mathbf{I}$ which reduces their statistic to test the hypothesis of our interest, i.e., $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$. Therefore, we only consider the reduced form of their statistic for $\mathbf{T} = \mathbf{I}$ (see also Sect. 6). Then, the statistic can be written as

$$A_N = \frac{Q}{B_0}, \tag{6}$$

where

$$Q = \tfrac{2n_1n_2}{N}\left(\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}\right)'\left(\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}\right), \tag{7}$$

$N = n_1 + n_2$, and B_0 is defined below. It is shown that

$$A_N \approx \chi^2_f / f, \tag{8}$$

as $n_1, n_2 \to \infty$, where

$$f = \frac{[\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2}{\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2}, \tag{9}$$

which is estimated as E_2/E_3. Here, B_0, E_2 and E_3 are the unbiased and consistent estimators of the traces $\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$, $[\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2$ and $\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2$, respectively, and are defined as

$$B_0 = \frac{1}{n_1n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} A_{kl} \tag{10}$$

$$E_2 = \frac{1}{n_1n_2(n_1-1)(n_2-1)}\underbrace{\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}}_{k\neq r,\ l\neq s} A_{kl}A_{rs} \tag{11}$$

$$E_3 = \frac{1}{n_1n_2(n_1-1)(n_2-1)}\underbrace{\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}}_{k\neq r,\ l\neq s} A^2_{klrs}, \tag{12}$$

where $A_{kl} = \mathbf{D}'_{kl}\mathbf{D}_{kl}$, $A_{rs} = \mathbf{D}'_{rs}\mathbf{D}_{rs}$ are quadratic forms, and $A_{klrs} = \mathbf{D}'_{kl}\mathbf{D}_{rs}$ is a symmetric bilinear form, with $\mathbf{D}_{kl} = \mathbf{X}_{1k} - \mathbf{X}_{2l}$ and $\mathbf{D}_{rs} = \mathbf{X}_{1r} - \mathbf{X}_{2s}$. Their proof of the asymptotic Chi-square distribution of A_N is based on the assumptions that $\mathcal{F}_1$ and $\mathcal{F}_2$ in (1) are multivariate normal, and p is fixed, where $N \to \infty$.

It may be worth mentioning here that when $p > n$, a test statistic for testing a hypothesis on the location parameters cannot be constructed by fully standardizing the norm of difference of means $||\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}||^2$ with an inverse of covariance matrix as a scaling factor, like in case of Hotelling's T^2 statistic, since the estimated covariance matrix is not invertible. The construction of A_N, like for the one-sample case, is based on replacing the inverse of covariance matrix in the norm with the trace of the covariance matrix in the denominator. In the above notations, B_0 is a moment estimator of such a trace, $\text{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$. In this sense, the statistic has a similar form as introduced in Dempster (1958), Bai and Saranadasa (1996).

Note that $\text{Cov}(\mathbf{D}_{kl}) = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$. Let $\boldsymbol{\Sigma}_0 = \text{Cov}(\overline{\mathbf{X}}_{1\cdot} - \overline{\mathbf{X}}_{2\cdot}) = \frac{1}{n_1}\boldsymbol{\Sigma}_1 + \frac{1}{n_2}\boldsymbol{\Sigma}_2$. If, for computational convenience, we write the components of the test statistic in matrix form as

$$
\begin{aligned}
Q &= \frac{2}{Nn_1n_2}\mathbf{X}'\mathbf{M}'(\mathbf{J}_{n_1} \otimes \mathbf{J}_{n_2} \otimes \mathbf{I}_p)\mathbf{M}\mathbf{X} \\
B_0 &= \frac{1}{n_1n_2}\mathbf{X}'\mathbf{M}'(\mathbf{I}_{n_1} \otimes \mathbf{I}_{n_2} \otimes \mathbf{I}_p)\mathbf{M}\mathbf{X} \; = \; \frac{1}{n_1n_2}\mathbf{X}'\mathbf{M}'\mathbf{M}\mathbf{X},
\end{aligned}
$$

where $\mathbf{M} = (\mathbf{I}_{n_1} \otimes \mathbf{1}_{n_2} \mid -\mathbf{1}_{n_1} \otimes \mathbf{I}_{n_2}) \otimes \mathbf{I}_p$, with $\mathbf{1}$ as a vector of 1s, $\mathbf{J} = \mathbf{1}\mathbf{1}'$, and $\mathbf{I}$ as identity matrix, then, under normality, the following can be immediately proved (see Ahmad 2008, Theorem 3.3).

$$
\begin{aligned}
\text{E}(Q) &= \frac{2n_1n_2}{N}\text{tr}(\boldsymbol{\Sigma}_0) && (13)\\
\text{Var}(Q) &= \frac{8n_1^2n_2^2}{N^2}\text{tr}\left(\boldsymbol{\Sigma}_0^2\right) && (14)\\
\text{E}(B_0) &= \text{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) && (15)\\
\text{Var}(B_0) &= \frac{2}{n_1n_2}\text{tr}\left(n_2\boldsymbol{\Sigma}_1^2 + 2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 + n_1\boldsymbol{\Sigma}_2^2\right) && (16)\\
\text{Cov}(Q, B_0) &= \frac{4n_1n_2}{N}\text{tr}\left(\boldsymbol{\Sigma}_0^2\right) = \frac{N}{2n_1n_2}\text{Var}(Q). && (17)
\end{aligned}
$$

Using these moments and the delta method, the first two moments of A_N in (6) are computed which, asymptotically, coincide with those of χ^2_f/f. This eventually leads to the Chi-square approximation of A_N, as given in (8), based on the following representation theorem of a quadratic form, combined with the Box's approximation Box (1954).

Theorem 1 *Let* $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ *and let* $\mathbf{G}$ *be any symmetric, positive semi-definite matrix with* r *non-zero eigenvalues,* $r \leq p$*. Then*

$$
\mathbf{X}'\mathbf{G}\mathbf{X} \sim \sum_{i=1}^{r} \lambda_i C_i
$$

where λ_i *are the eigenvalues of* $\mathbf{G}\boldsymbol{\Sigma}$ *and the* $C_i \sim \chi^2_1$ *are independent.*

As the Box's approximation calls for (1954, Theorem 3.1), we equate first two moments of $\mathbf{X}'\mathbf{G}\mathbf{X}$ with those of a scaled, $g\chi^2_f$ distribution, i.e.,

$$gf = \sum_{i=1}^{r} \lambda_i$$

$$2g^2 f = 2\sum_{i=1}^{r} \lambda_i^2 + \left(\sum_{i=1}^{r} \lambda_i\right)^2,$$

such that

$$f = \frac{[\mathrm{tr}(\mathbf{G}\boldsymbol{\Sigma})]^2}{\mathrm{tr}(\mathbf{G}\boldsymbol{\Sigma})^2} \quad \text{and} \quad g = \frac{\mathrm{tr}(\mathbf{G}\boldsymbol{\Sigma})^2}{\mathrm{tr}(\mathbf{G}\boldsymbol{\Sigma})}, \tag{18}$$

where $\sum_{i=1}^{r} \lambda_i = \mathrm{tr}(\mathbf{G}\boldsymbol{\Sigma})$ and $\sum_{i=1}^{r} \lambda_i^2 = \mathrm{tr}(\mathbf{G}\boldsymbol{\Sigma})^2$. This gives the approximation for $A_N = Q/B_0$ with f in (9), estimated as E_2/E_3. For more details, see Ahmad (2008).

We are interested to know the behavior of A_N when both p and n_i are large, but without assuming any relationship between them. Further, we want to evaluate A_N when $\mathcal{F}_1$ and $\mathcal{F}_2$ in (1) are not necessarily multivariate normal. We, however, continue to assume that $\boldsymbol{\Sigma}_i$, $i = 1, 2$, are positive definite. Results for normal distribution and fixed p case will be discussed as special cases, for reference.

In the next section, the modified version of A_N is presented and its asymptotic distribution is derived under certain assumptions.

3 The modified test statistic

3.1 The independent case

Consider the statistic A_N in (6) again. To justify the modification of A_N for high-dimensional and non-normal setup, we need the following assumptions.

Assumption 2 $\mathrm{E}(X_{1ks}^4) \le \gamma_1 < \infty$ and $\mathrm{E}(X_{2ls}^4) \le \gamma_2 < \infty$, $\forall\, s = 1, \ldots, p$, for some γ_1, γ_2.

Assumption 3 For $p \to \infty$, let $\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{p} = O(1)$, $i = 1, 2$.

Assumption 4 For $p \to \infty$, let $\frac{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}{p^2} = O(\delta)$, where $0 < \delta \le 1$, $i, j = 1, 2$.

Assumptions 2–4 are straightforward extensions of similar one-sample assumptions discussed in Ahmad et al. (2012a). For a justification of these assumptions, we, therefore, simply refer to the said paper. Additionally, to avoid any degeneracy of the asymptotic limit distribution of the test statistic due to the growing sample sizes, we also assume that $n_1/n_2 \to c \in (0, \infty)$ when $n_1, n_2 \to \infty$. Clearly, this assumption does not disturb the practical application of the test statistic, but is required as a precaution to ensure a stable limit of the test statistic under H_0.

Now, for the modification of A_N, let us begin with the denominator, B_0. For computational simplicity, we assume, under $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, that $\boldsymbol{\mu}_1 = \mathbf{0} = \boldsymbol{\mu}_2$, without any loss of generality. Then, from Eq. (10), we have

$$
\begin{aligned}
B_0 &= \frac{1}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} A_{kl} = \frac{1}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} (\mathbf{X}_{1k}-\mathbf{X}_{2l})'(\mathbf{X}_{1k}-\mathbf{X}_{2l}) \\
&= \frac{1}{n_1}\sum_{k=1}^{n_1}\mathbf{X}_{1k}'\mathbf{X}_{1k} + \frac{1}{n_2}\sum_{l=1}^{n_2}\mathbf{X}_{2l}'\mathbf{X}_{2l} - \frac{2}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbf{X}_{1k}'\mathbf{X}_{2l} \\
&= E_{11} + E_{21} - R = E_1 - R,
\end{aligned}
$$

where $R = \frac{2}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\mathbf{X}_{1k}'\mathbf{X}_{2l}$ with $\mathrm{E}(R) = 0$, under the null hypothesis, and $E_1 = E_{11} + E_{21}$. Further, $\mathrm{E}(E_1) = \mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$, which is the same as $\mathrm{E}(B_0)$. This implies that, the bilinear form, R, in B_0 does not contribute anything to estimate what B_0 is constructed to estimate, i.e., $\mathrm{tr}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)$. Further, we note that $\mathrm{E}(E_1/p) = O(1)$, under Assumption 3. It may be worthwhile here to note that, an alternative way to attain such a bound is to use E_1, and, in addition to Assumption 3, further assume that $p/n_i \to c_i \in (0, \infty)$, so that $\mathrm{E}(E_1) = (c_1 + c_2)O(1)$; see for example Ledoit and Wolf (2002), Fujikoshi et al. (2010). Since, we want to avoid such a restrictive assumption, we shall use $\frac{1}{p}E_1$ when we assume $p \to \infty$, whereas for fixed p, we may use E_1 (see Sect. 4). Moreover, as it will be clear from the main result (Theorem 5), the kernels of the U-statistics used to prove the asymptotic distribution of the modified test statistic are also normalized by p, which renders the use of $\frac{1}{p}E_1$ completely in consistency with the rest of the computations.

We also note that, in the original normality-based approximation, i.e., (8), the factor $2n_1n_2/N$ is used to replace $\mathrm{tr}\,(\boldsymbol{\Sigma}_0)$ in the denominator of A_N with $\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$, the reason being that no well-defined estimator could be given for $\mathrm{tr}\,(\boldsymbol{\Sigma}_0)$ keeping $n_1 \neq n_2$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, i.e., under Behrens–Fisher setting. The estimator B_0, as defined in Eq. (10), is therefore used to estimate $\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$, and the results are presented separately, once assuming $n_1 = n_2$, and once assuming $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ such that, for each separate case, $\mathrm{E}(Q) = \mathrm{E}(E_1)$. This helped obtain the first two moments of the proposed test statistic, using the delta method, as 1 and $2/f$, same as that of χ^2_f/f, to eventually show the approximation in (8), based on Box's approximation.

In our case, we have E_1 as an unbiased estimator of $\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$. Moreover, it will be shown in the proof of Theorem 5 that $\frac{1}{p}E_1$ is uniformly bounded, independently of p, under Assumptions 2 and 4. This proves $\frac{1}{p}E_1$ to be a well-defined, i.e., unbiased and consistent, estimator, even under high-dimensional and non-normal setup.

Now we consider the numerator of A_N in (6), i.e., Q. First, we note that

$$
\mathrm{E}\left(\frac{1}{p}Q\right) = \frac{2n_1n_2}{Np}\mathrm{tr}(\boldsymbol{\Sigma}_0) = O(1),
$$

under Assumption 3, where $\boldsymbol{\Sigma}_0 = \frac{1}{n_1}\boldsymbol{\Sigma}_1 + \frac{1}{n_2}\boldsymbol{\Sigma}_2$. To take a closer look at Q, we write it in expanded form, using Eqs. (7) and (10), as

$$
\begin{aligned}
Q &= \frac{2n_1n_2}{N}\cdot\frac{1}{n_1^2n_2^2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}A_{klrs}\\
&= \frac{2n_1n_2}{N}\cdot\frac{1}{n_1^2n_2^2}\left(\underbrace{\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}}_{k=r,l=s}A_{kl}+\underbrace{\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\sum_{r=1}^{n_1}}_{k\neq r,l=s}A_{klr}\right.\\
&\quad\left.+\underbrace{\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\sum_{s=1}^{n_2}}_{k=r,l\neq s}A_{kls}+\underbrace{\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\sum_{r=1}^{n_1}\sum_{s=1}^{n_2}}_{k\neq r,l\neq s}A_{klrs}\right),
\end{aligned}
$$

wherein terms involving the double, triple, and quadruple sums clearly are overloaded with same or similar quadratic or bilinear forms, all of which are not inevitably needed for the asymptotic distribution of the statistic. Under H_0, these terms can be written as

$$A_{kl} = \mathbf{X}_{1k}'\mathbf{X}_{1k} - \mathbf{X}_{1k}'\mathbf{X}_{2l} - \mathbf{X}_{2l}'\mathbf{X}_{1k} + \mathbf{X}_{2l}'\mathbf{X}_{2l} \tag{19}$$
$$A_{klr} = \mathbf{X}_{1k}'\mathbf{X}_{1r} - \mathbf{X}_{1k}'\mathbf{X}_{2l} - \mathbf{X}_{2l}'\mathbf{X}_{1r} + \mathbf{X}_{2l}'\mathbf{X}_{2l} \tag{20}$$
$$A_{kls} = \mathbf{X}_{1k}'\mathbf{X}_{1k} - \mathbf{X}_{1k}'\mathbf{X}_{2s} - \mathbf{X}_{2l}'\mathbf{X}_{1k} + \mathbf{X}_{2l}'\mathbf{X}_{2s} \tag{21}$$
$$A_{klrs} = \mathbf{X}_{1k}'\mathbf{X}_{1r} - \mathbf{X}_{1k}'\mathbf{X}_{2s} - \mathbf{X}_{2l}'\mathbf{X}_{1r} + \mathbf{X}_{2l}'\mathbf{X}_{2s}. \tag{22}$$

Let $A_{1k} = \mathbf{X}_{1k}'\mathbf{X}_{1k}$ and $A_{2l} = \mathbf{X}_{2l}'\mathbf{X}_{2l}$ be the quadratic forms defined for sample 1 and sample 2, respectively. Similarly, let $A_{1kr} = \mathbf{X}_{1k}'\mathbf{X}_{1r}$, $k \neq r$, and $A_{2ls} = \mathbf{X}_{2l}'\mathbf{X}_{2s},\ l \neq s$, be the symmetric bilinear forms computed from elements of sample 1 and sample 2, respectively, and $A_{12kl} = \mathbf{X}_{1k}'\mathbf{X}_{2l}$ be another symmetric bilinear form computed from elements of both samples, $k, r = 1, \ldots, n_1, l, s = 1, \ldots, n_2$. Then, the expansion of Q, as given above, can be further simplified as following.

$$
\begin{aligned}
Q &= \frac{2n_1n_2}{N}\left(\frac{1}{n_1^2}\sum_{k=1}^{n_1}A_{1k}+\frac{1}{n_2^2}\sum_{l=1}^{n_2}A_{2l}\right.\\
&\quad\left.+\frac{1}{n_1^2}\sum_{\substack{k=1\\k\neq r}}^{n_1}\sum_{r=1}^{n_1}A_{1kr}+\frac{1}{n_2^2}\sum_{\substack{l=1\\l\neq s}}^{n_2}\sum_{s=1}^{n_2}A_{2ls}-\frac{2}{n_1n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}A_{12kl}\right)\\
&= \frac{2n_1n_2}{N}\left(\frac{1}{n_1}E_{11}+\frac{1}{n_2}E_{21}\right)+Q_0 = Q_1+Q_0,
\end{aligned} \tag{23}
$$

where $E_{11} = \frac{1}{n_1}\sum_{k=1}^{n_1} A_{1k}$, $E_{21} = \frac{1}{n_2}\sum_{l=2}^{n_2} A_{2l}$, and Q_0 is the entire expression on the second line in the expansion of Q above, along with the multiplier $\frac{2n_1n_2}{N}$. Now, since $\mathrm{E}(E_{i1}) = \mathrm{tr}(\boldsymbol{\Sigma}_i)$, $i = 1, 2$, therefore,

$$\mathrm{E}(Q_1) = \tfrac{2n_1n_2}{N}\mathrm{tr}\left(\frac{1}{n_1}\boldsymbol{\Sigma}_1 + \frac{1}{n_2}\boldsymbol{\Sigma}_2\right) = \tfrac{2n_1n_2}{N}\mathrm{tr}(\boldsymbol{\Sigma}_0) = \tfrac{2n_1n_2}{N}\mathrm{Cov}(\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}),$$

which implies that $\mathrm{E}\,(Q_1)$ approximates $\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$ when $n_1, n_2 \to \infty$, such that $\frac{1}{p}E_1$ can, asymptotically, replace $\frac{1}{p}Q_1$, under Assumptions 3 and 4. Now, consider Q_0. For convenience, assume $n_1 = n_2 = n$. Then, the terms like $\frac{2n_1n_2}{N} \cdot \frac{1}{n_1}$ converge to 1, when $n \to \infty$, which motivates us to slightly re-write Q_0 as, say E_0, where

$$E_0 = \frac{1}{n_1}\sum_{\substack{k=1 \\ k\neq r}}^{n_1}\sum_{r=1}^{n_1}\frac{1}{p}A_{1kr} + \frac{1}{n_2}\sum_{\substack{l=1 \\ l\neq s}}^{n_2}\sum_{s=1}^{n_2}\frac{1}{p}A_{2ls} - \frac{2}{\sqrt{n_1n_2}}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}\frac{1}{p}A_{12kl}. \tag{24}$$

We immediately note that $\mathrm{E}(E_0) = 0$ and

$$\mathrm{Var}(E_0) = \frac{2}{p^2}\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2 - \frac{2}{p^2}\mathrm{tr}\left(\frac{1}{n_1}\boldsymbol{\Sigma}_1^2 + \frac{1}{n_2}\boldsymbol{\Sigma}_2^2\right), \tag{25}$$

where the second term vanishes when $n_1, n_2 \to \infty$ for any fixed p, and also when $n_1, n_2, p \to \infty$ under Assumption 4. Further, $2\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2$ is exactly the same as the variance of the quadratic form $(\mathbf{X}_{1k} - \mathbf{X}_{2l})'(\mathbf{X}_{1k} - \mathbf{X}_{2l})$ under normality when $k \neq l$. Actually, it can be similarly shown that $\mathrm{Var}(Q_0)$ converges, asymptotically, to $\mathrm{Var}(Q)$ under normality; see Eq. (14). Moreover, it can be trivially shown that the second term in $\mathrm{Var}(E_0)$ is exactly $\mathrm{Var}(E_1/p)$ under normality (see Sect. 4). In other words, $\mathrm{Var}(E_0 + E_1) = \frac{2}{p^2}\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2$ if we assume normality. This clues to the fact that a test statistic of the form $(E_1 + E_0)/E_1 = 1 + E_0/E_1$ can suffice our needs to test H_0.

Now, we are ready to define the modified test statistic. From the detailed inspection of Q and B_0 above, we conclude that the term Q_1 gives the mean of Q but contributes nothing to the variance, and the term Q_0 gives the variance but contributes nothing to the mean. In other words, Q and B_0 are overloaded with extra terms which are not really needed, and which rather hamper a nice convergence of the estimators to the target limits. In fact, E_0 can replace Q and E_1 can replace B_0 in A_N, without any loss, as far as unbiasedness, consistency, and efficiency of the estimators are concerned. Relieving the statistic A_N of all what it is overburdened with, we define a modified form of it as

$$T = 1 + \frac{E_0}{\frac{1}{p}E_1}, \tag{26}$$

where $E_1 = E_{11} + E_{21}$ and E_0 is as defined in Eq. (24). The statistic T, as shown in Eq. (26), and its components E_0 and E_1, are direct two-sample extensions of a similar one-sample statistic and its components discussed in Ahmad et al. (2012a). It will be shown in the following (Theorem 5) that the asymptotic limit distribution of T is mainly determined by E_0. For this, we write E_0 as a linear combination of U-statistics as

$$E_0 = (n_1 - 1)U_{n_1} + (n_2 - 1)U_{n_2} - 2\sqrt{n_1 n_2}U_{n_1 n_2} \tag{27}$$

where

$$U_{n_1} = \frac{1}{n_1(n_1-1)} \sum_{\substack{k=1 \\ k \neq r}}^{n_1} \sum_{r=1}^{n_1} \frac{1}{p} A_{1kr} \quad \text{and} \quad U_{n_2} = \frac{1}{n_2(n_2-1)} \sum_{\substack{l=1 \\ l \neq s}}^{n_2} \sum_{s=1}^{n_2} \frac{1}{p} A_{2ls}$$

are one-sample U-statistics and

$$U_{n_1 n_2} = \frac{1}{n_1 n_2} \sum_{k=1}^{n_1} \sum_{l=1}^{n_2} \frac{1}{p} A_{12kl}$$

is a two-sample U-statistic (Hoeffding 1948; Lehmann 1999, Chapter 6). Before we move on to state and prove the main theorem on the asymptotic distribution of T, a few comments on the structure of the three U-statistics are in order. Note that, the kernels of the U-statistics are all normalized by p. This is completely in consonance with the assumptions and, as will be clear from the proof of the main theorem below, is essential to obtain a non-degenerate asymptotic limit of T. Consider U_{n_1}, with the symmetric kernel $h(\mathbf{X}_{1k}, \mathbf{X}_{1r}) = \frac{1}{p}A_{1kr} = \frac{1}{p}\mathbf{X}'_{1k}\mathbf{X}_{1r}$, $k \neq r$. For convenience, we can write $B_{1kr} = \mathbf{Y}'_{1k}\mathbf{Y}_{1r}$ with $\mathbf{Y}_{1j} = \mathbf{X}_{1j}/\sqrt{p}$, $j = k, r$, so that, $\mathrm{E}(\mathbf{Y}_j) = \mathbf{0}$, under H_0, and $\mathrm{Var}(\mathbf{Y}_j) = \frac{1}{p}\mathbf{\Sigma}_1$. This implies that, $\mathrm{E}(\mathbf{Y}'_j\mathbf{Y}_j) = \frac{1}{p}\mathrm{tr}(\mathbf{\Sigma}_1)$ and $\mathrm{E}(\mathbf{Y}'_k\mathbf{Y}_l)^2 = \frac{1}{p^2}\mathrm{tr}(\mathbf{\Sigma}_1^2)$. Exactly same arguments apply to the kernels of the other two U-statistics.

Now, if we let λ_{1j}, λ_{2j}, and λ_{3j}, $j = 1, \ldots, p$, be the eigenvalues of $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}$, respectively, then Assumptions 3 and 4 refer to the moments of $\frac{\lambda_{1j}}{p}$, $\frac{\lambda_{2j}}{p}$, $\frac{\lambda_{3j}}{p}$ which are the eigenvalues of $\frac{1}{p}\mathbf{\Sigma}_1$, $\frac{1}{p}\mathbf{\Sigma}_2$, and $\frac{1}{p}\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}$, respectively. Since, it is these p-scaled eigenvalues which, under Assumption 3, are uniformly bounded away from 0 and ∞, and which, under Assumption 4, keep the p-scaled kernel of the U-statistic square-integrable; therefore, we shall, in the sequel that follows, use these eigenvalues. For convenience, let us denote the eigenvalues of $\frac{1}{p}\mathbf{\Sigma}_1$, $\frac{1}{p}\mathbf{\Sigma}_2$, and $\frac{1}{p}\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}$ as ν_{1j}, ν_{2j}, and ν_{3j}, $j = 1, \ldots, p$, respectively. Obviously, when p grows very large, the eigenvalues of the unscaled kernels, i.e., λ_{1j}s, etc., can be very small so that the eigenvalues of the scaled kernels, i.e., ν_{1j}s, etc., can asymptotically vanish. This is exactly how the proof of our main theorem is properly justified, based on the Hilbert–Schmidt theorem (Theorem 18). This is more elaborated in Sect. 5.

Now, we prove the main theorem on the asymptotic distribution of T.

Theorem 5 *Given Assumptions 2, 3 and 4. Then, under H_0, the test statistic T, defined in Eq. (26), follows the same scaled Chi-square approximation as in (8),* *when $n_1, n_2, p \to \infty$. Further, under the same set up, and using Hájek–Šidák Lemma (see Lemma 6 below), it can be shown that,*

$$\frac{T - E(T)}{\sqrt{\widehat{\mathrm{Var}(T)}}} \xrightarrow{\mathcal{D}} N(0, 1),$$

where $E(T) = 1$ and $\widehat{\mathrm{Var}(T)}$ denotes the sample estimator of Var(T).

Proof Let ν_{1j}, ν_{2j}, and ν_{3j}, $j = 1, \ldots, p$, be the eigenvalues of $\frac{1}{p}\mathbf{\Sigma}_1$, $\frac{1}{p}\mathbf{\Sigma}_2$ and $\frac{1}{p}\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}$, respectively, as defined above. First, we show the probability convergence of the denominator of T, i.e., $\frac{1}{p}E_1$. Clearly, $\mathrm{E}(E_1) = \mathrm{tr}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)$. Now, by Cauchy–Schwarz inequality, $\mathrm{E}(\mathbf{X}_{1k}'\mathbf{X}_{1k})^2 \leq \gamma_1 p^2$ and $\mathrm{E}(\mathbf{X}_{2l}'\mathbf{X}_{2l})^2 \leq \gamma_2 p^2$, under Assumption 2. Then, by the independence of the two samples, and by Assumption 4,

$$\mathrm{Var}\left(\frac{1}{p}E_1\right) \leq 2\left(\frac{\gamma_1}{n_1} + \frac{\gamma_2}{n_2}\right) O(\delta), \tag{28}$$

so that $\frac{1}{p}E_1 \xrightarrow{\mathcal{P}} \sum_{j=1}^{\infty}(\nu_{1j} + \nu_{2j})$, which, under Assumption 3, is uniformly bounded away from 0 and ∞, where $\xrightarrow{\mathcal{P}}$ denotes the convergence in probability.

Now, we show the asymptotic limit of the numerator, E_0. We begin with U_{n_1} where

$$U_{n_1} = \frac{1}{n_1(n_1-1)}\sum_{\substack{k=1 \\ k \neq r}}^{n_1}\sum_{r=1}^{n_1}\frac{1}{p}A_{1kr}$$

with the symmetric kernel $h(\mathbf{Y}_{1k}, \mathbf{Y}_{1r}) = \frac{1}{p}A_{1kr}$, $k \neq r$. Then, following the proof of Theorem 2.6 in Ahmad et al. (2012a), the variance of U_{n_1} is given as

$$\mathrm{Var}(U_{n_1}) = \frac{2}{n_1(n_1-1)}\frac{\mathrm{tr}\left(\mathbf{\Sigma}_1^2\right)}{p^2}, \tag{29}$$

such that $\mathrm{Var}(n_1 U_{n_1})$ is bounded, under Assumption 4, as $n_1, p \to \infty$ which implies that $n_1 U_{n_1}$ has a non-degenerate limit distribution (van der Vaart 1998, p. 167) and this asymptotic limit is given as

$$n_1 U_{n_1} \xrightarrow{\mathcal{D}} \sum_{j=1}^{\infty}\nu_{1j}C_{1j} - \sum_{j=1}^{\infty}\nu_{1j}, \tag{30}$$

as $n_1, p \to \infty$, where C_{1j}s are independent χ_1^2 random variables, and, under Assumption 3, $\sum_{j=1}^{\infty}\nu_{1j}$ is uniformly bounded away from 0 and ∞ (see also Remark 1 in Sect. 5 for more details). Continuing with the same strategy for U_{n_2}, we obtain

$$\mathrm{Var}(U_{n_2}) \;=\; \frac{2}{n_2(n_2-1)}\frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_2^2\right)}{p^2}, \tag{31}$$

where the asymptotic distribution of $n_2 U_{n_2}$ is

$$n_2 U_{n_2} \;\xrightarrow{\mathcal{D}}\; \sum_{j=1}^{\infty} \nu_{2j} C_{2j} - \sum_{j=1}^{\infty} \nu_{2j}, \tag{32}$$

as $n_2, p \to \infty$, where C_{2j}s are independent χ_1^2 random variables.

For $U_{n_1 n_2}$, we note that it is a two-sample U-statistic, the general form of which is defined as (Lehmann 1999, Chapter 6)

$$\frac{1}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \sum_{P(a)} \sum_{P(b)} h\left(\mathbf{Y}_{a_1}, \ldots, \mathbf{Y}_{a_{m_1}}; \mathbf{Y}_{b_1}, \ldots, \mathbf{Y}_{b_{m_2}}\right),$$

where $P(a) = 1 \le a_1 < \cdots < a_{m_1} \le n_1$ and $P(b) = 1 \le b_1 < \cdots < b_{m_2} \le n_2$ denote the permutations over all indices, and $h(\mathbf{Y}_{a_1}, \ldots, \mathbf{Y}_{a_{m_1}}; \mathbf{Y}_{b_1}, \ldots, \mathbf{Y}_{b_{m_2}})$ is the corresponding symmetric kernel. Clearly, $U_{n_1 n_2}$ refers to the simplest two-sample U-statistic, with $m_1 = 1 = m_2$, which can be exclusively written as

$$U_{n_1 n_2} \;=\; \frac{1}{n_1 n_2} \sum_{k=1}^{n_1} \sum_{l=1}^{n_2} h(\mathbf{Y}_{1k}, \mathbf{Y}_{2l}), \tag{33}$$

with the symmetric kernel $h(\mathbf{Y}_{1k}, \mathbf{Y}_{2l}) = \frac{1}{p} A_{12kl} = \frac{1}{p} \mathbf{Y}'_{1k} \mathbf{Y}_{2l}$. Denote

$$h_{c_1 c_2}(\mathbf{Y}_{1k}, \mathbf{Y}_{2l}) = \mathrm{E}\big(h(\mathbf{Y}_{1k}, \mathbf{Y}_{2l}) \mid \mathbf{Y}_{1k} = \mathbf{Y}_{1k}, \mathbf{Y}_{2l} = \mathbf{Y}_{2l}\big),$$

where, with $c_1 = 1$ and $c_2 = 1$, we note that $h_{11} = 0$, along with $h_{01} = 0 = h_{10}$, under H_0, so that $h(\mathbf{Y}_{1k}, \mathbf{Y}_{2l})$ is a degenerate kernel. Clearly, $\mathrm{E}(U_{n_1 n_2}) = 0$, and the variance of $U_{n_1 n_2}$ can be computed as (see Lehmann 1999, p. 373)

$$\mathrm{Var}(U_{n_1 n_2}) \;=\; \frac{1}{n_1 n_2} \mathrm{Var}(h_{11}) = \frac{1}{n_1 n_2} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2)}{p^2}, \tag{34}$$

where

$$\mathrm{E}\left(h^2(\mathbf{Y}_{1k}, \mathbf{Y}_{2l})\right) = \sum_{j=1}^{p} \nu_{3j}^2 = \frac{1}{p^2} \mathrm{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2) < \infty,$$

by Assumption 4, and $\sum_{j=1}^{p} \nu_{3j}$ is uniformly bounded away from 0 and ∞, by Assumption 3. This implies that $U_{n_1 n_2}$ is first order degenerate two-sample U-statistic, and $\sqrt{n_1 n_2} U_{n_1 n_2}$ can have a finite asymptotic limit (Koroljuk and Borovskich 1994, Chapter 4). This limit, following Koroljuk and Borovskich (1994, Theorem 4.5.3, p. 156) is given as

$$\sqrt{n_1 n_2} U_{n_1 n_2} \xrightarrow{\mathcal{D}} \sum_{j=1}^{\infty} \nu_{3j} z_{1j} z_{2j},$$

as $\min(n_1, n_2)$, $p \to \infty$, where z_{1j} and z_{2j} are two independent sequences of independent standard normal variables. Combining the results for E_0 in (27), we have

$$E_0 \xrightarrow{\mathcal{D}} \sum_{j=1}^{\infty} \nu_{1j} \left(z_{1j}^2 - 1 \right) + \sum_{j=1}^{\infty} \nu_{2j} \left(z_{2j}^2 - 1 \right) - 2 \sum_{j=1}^{\infty} \nu_{3j} z_{1j} z_{2j}, \tag{35}$$

as $n_1, n_2, p \to \infty$, where $z_{1j}^2 = C_{1j}$ and $z_{3j}^2 = C_{2j}$. Combining the approximation in (35) with that of $\frac{1}{p} E_1$, and using Slutsky's lemma (van der Vaart 1998, Lemma 2.8, p. 11) the approximation of T in Eq. (26) can be written as

$$T \xrightarrow{\mathcal{D}} 1 + \frac{W}{K}, \tag{36}$$

as $n_1, n_2, p \to \infty$, where

$$W = \sum_{j=1}^{\infty} \nu_{1j} \left(z_{1j}^2 - 1 \right) + \sum_{j=1}^{\infty} \nu_{2j} \left(z_{2j}^2 - 1 \right) - 2 \sum_{j=1}^{\infty} \nu_{3j} z_{1j} z_{2j},$$

and $K = \sum_{j=1}^{\infty} (\nu_{1j} + \nu_{2j})$, where K, under Assumption 3, is uniformly bounded away from 0 and ∞. If we let $w_{1j} = \nu_{1j}/K$, and similarly, w_{2j} and w_{3j}, and write

$$W = \sum_{j=1}^{\infty} \nu_{1j} z_{1j}^2 + \sum_{j=1}^{\infty} \nu_{2j} z_{2j}^2 - 2 \sum_{j=1}^{\infty} \nu_{3j} z_{1j} z_{2j} - \left(\sum_{j=1}^{\infty} \nu_{1j} + \sum_{j=1}^{\infty} \nu_{2j} \right),$$

then (36) can be re-expressed as

$$T - 1 \xrightarrow{\mathcal{D}} \sum_{j=1}^{\infty} w_{1j} z_{1j}^2 + \sum_{j=1}^{\infty} w_{2j} z_{2j}^2 - 2 \sum_{j=1}^{\infty} w_{3j} z_{1j} z_{2j} - 1, \tag{37}$$

which represents T in the same form as the representation of a quadratic form in Theorem 1. This gives us a two-sample analog of Theorem 1 when the data are high-dimensional and non-normal, and this representation is independent of whether the sample sizes and/or covariance matrices are assumed equal or not, i.e., representation (37) is valid even under Behrens–Fisher setting.

Now, since, $\mathrm{E}(W) = 0$ and $\mathrm{Var}(W) = 2\sum_{j=1}^{\infty}(\nu_{1j} + \nu_{2j})^2$, it can be immediately verified that,

$$\mathrm{E}(T) = 1 \tag{38}$$

$$\mathrm{Var}(T) = \frac{\frac{2}{p^2}\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2}{\left[\frac{1}{p}\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\right]^2} \tag{39}$$

which are the same moments as obtained under the assumption of normality with p assumed fixed (and large), i.e., the moments coincide with those of χ^2_f/f, based on the Box's approximation, see Theorem 1 and f in (9); see also Eq. (18). This proves the main part of the theorem, that the Box's approximation, and the approximation of T to χ^2_f/f, also hold for high-dimensional and non-normal case, even under Behrens–Fisher setting.

Before we proceed for the proof of the second part of the theorem, a few comments are in order. First, a comparison of $\mathrm{Var}(W) = \frac{2}{p^2}\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2$ with $\mathrm{Var}(E_0)$ in Eq. (25) indicates that the U-statistics approximation has removed the unwanted part of the variance of E_0, i.e., variance of W captures the entire variance of $E_0 + E_1$, and hence completely justifies the modified form of the statistic as $1 + E_0/E_1$.

Second, it is clear from the computations above that the same proof of approximation of T to χ^2_f/f remains valid even if we assume p fixed (although large, including $p > n_i$) while keeping normality assumption relaxed. All the derivations go through, and even more comfortably for the fixed p case since the asymptotic limits of all three U-statistics are directly applicable for $n \to \infty$ without having to control the convergence for large p. This also implies that, we need not norm the kernels of the U-statistics by p as the condition of square-integrability of the kernels, i.e., $\mathrm{E}(h^2(\cdot)) < \infty$ remains intact when p is fixed. This shows that the original approximation of T as given in Ahmad et al. (2008) for fixed p under normality, i.e., (8), is valid without normality assumption whether p is kept fixed or is allowed to grow with n_i; see also Sect. 5.1.

Now, to prove the second part of the theorem, i.e., the convergence of T to normal distribution, we need the following special case of Lindeberg–Feller central limit theorem, usually known as Hájek–Šidák Lemma (see Jiang 2010, Example 6.6, p. 183; Hájek et al. 1999, p. 184). □

Lemma 6 *Let $X_1, X_2, \ldots$ be iid random variables with mean 0 and variance 1. Let b_{ni}, $1 \le i \le n$, be a sequence of constants such that $\max_i b^2_{ni} \to 0$ as $n \to \infty$. Then*

$$\sum_{i=1}^{n} b_{ni}X_i \xrightarrow{\mathcal{D}} N(0, 1),$$

as $n \to \infty$.

Consider U_{n_1}. We proved $n_1U_{n_1} \xrightarrow{\mathcal{D}} \sum_{j=1}^{\infty}\nu_{1j}(C_{1j} - 1)$, where C_{1j} are iid χ^2_1, with $\mathrm{E}(C_{1j}) = 1$ and $\mathrm{Var}(C_{1j}) = 2$, so that, in the notations of Lemma 6, $X_j = (C_{1j} - 1)/\sqrt{2}$. Let $a_{1j} = w_{1j}/\sqrt{\sum_{j=1}^{p} w^2_{1j}}$. Clearly, $\sum_j a^2_{1j} = 1$ and $\max_j a^2_{1j} \to 0$ as $p \to \infty$. Then, by Lemma 6, as $p \to \infty$,

$$\sum_{j=1}^{p} a_{1j} X_j = \frac{\sum_{j=1}^{p} w_{1j}(C_{1j} - 1)}{\sqrt{2\sum_{j=1}^{p} w_{1j}^2}} \xrightarrow{\mathcal{D}} N(0, 1).$$

By the same argument, we have, for U_{n_2},

$$\frac{\sum_{j=1}^{p} w_{2j}(C_{2j} - 1)}{\sqrt{2\sum_{j=1}^{p} w_{2j}^2}} \xrightarrow{\mathcal{D}} N(0, 1),$$

and, for $U_{n_1 n_2}$,

$$\frac{\sum_{j=1}^{p} w_{3j} z_{1j} z_{2j}}{\sqrt{\sum_{j=1}^{p} w_{3j}^2}} \xrightarrow{\mathcal{D}} N(0, 1),$$

as $p \to \infty$. As U_{n_1}, U_{n_2}, and $U_{n_1 n_2}$, being composed of independent random vectors, are all pairwise uncorrelated, we get, from Eq. (36),

$$T - 1 \xrightarrow{\mathcal{D}} \sqrt{2\sum_{j=1}^{\infty} \nu_{1j}^2} T_1 + \sqrt{2\sum_{j=1}^{\infty} \nu_{2j}^2} T_2 - 2\sqrt{\sum_{j=1}^{\infty} \nu_{3j}^2} T_3, \tag{40}$$

where each of T_1, T_2, T_3 is $N(0, 1)$. If we let Z_1, Z_2 and Z_3 be the standardized forms of the three U-statistics, i.e., $Z_1 = U_{n_1}/\sqrt{\mathrm{Var}(U_{n_1})}$ ($\mathrm{E}(U_{n_1}) = 0$), etc., then, as $n_1, n_2, p \to \infty$,

$$\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}\right).$$

Clearly, to make T in (40) practically workable, we need to estimate its variance, i.e., we need to estimate the quantities $\sum_{j=1}^{\infty} \nu_j^2$, and these estimates must be unbiased, and consistent under high-dimensionality. In fact, we already have such estimates. From the variances of the three U-statistics in Eqs. (29), (31), and (34), we define

$$B_1 = \frac{1}{n_1(n_1 - 1)} \sum_{\substack{k=1 \\ k \neq r}}^{n_1} \sum_{r=1}^{n_1} A_{1kr}^2 \tag{41}$$

$$B_2 = \frac{1}{n_2(n_2 - 1)} \sum_{\substack{l=1 \\ l \neq s}}^{n_1} \sum_{s=1}^{n_1} A_{2ls}^2 \tag{42}$$

$$B_{12} = \frac{1}{n_1 n_2} \sum_{k=1}^{n_1} \sum_{l=1}^{n_2} A_{12kl}^2 \tag{43}$$

as unbiased estimators of $\frac{1}{p^2}\text{tr}(\boldsymbol{\Sigma}_1^2)$, $\frac{1}{p^2}\text{tr}(\boldsymbol{\Sigma}_2^2)$, and $\frac{1}{p^2}\text{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)$, respectively, where $A_{1kr} = \mathbf{Y}_{1k}'\mathbf{Y}_{1r}$, $A_{2ls} = \mathbf{Y}_{2l}'\mathbf{Y}_{2s}$, and $A_{12kl} = \mathbf{Y}_{1k}'\mathbf{Y}_{2l}$ are all bilinear forms composed of independent vectors, as explained immediately after Eqs. (19–22). To use them in T as consistent estimators of the three traces, we need to show that the ratios of these estimators to their respective traces are uniformly bounded in p. This implies that, under unbiasedness, we only need to show that the (co)variances of the ratios like $B_1/\text{tr}(\boldsymbol{\Sigma}_1^2)$ vanish for large p and n_i, $i = 1, 2$.

We note that, the estimators B_1, B_2, B_{12}, are of the same form as E_3 defined for a similar purpose in the one-sample case in Ahmad et al. (2012a). Then the proofs for the variances of the ratios like $B_1/\text{tr}(\boldsymbol{\Sigma}_1^2)$ follow directly from the proof regarding E_3 in Ahmad et al. (2012a, p. 10). We still need to check the covariances. Because of independence of the two samples, $\text{Cov}(B_1, B_2) = 0$. The results of the other two covariances follow the same pattern, by symmetry. Consider $\text{Cov}(B_1, B_{12})$, where

$$\text{Cov}(B_1, B_{12}) = \frac{1}{n_1^2(n_1-1)n_2p^4}\sum_{\substack{k=1\\k\neq r}}^{n_1}\sum_{r=1}^{n_1}\sum_{t=1}^{n_1}\sum_{l=1}^{n_2}\text{Cov}\left(A_{1kr}^2, A_{12tl}^2\right). \tag{44}$$

The covariance in Eq. (44) involves terms of order $O\left(n_1^3 n_2\right)$, out of which the terms with non-zero covariances (when $k = t$ or $r = t$) are of order $O\left(n_1^2 n_2\right)$, which leaves the entire covariance expression of order $O\left(\frac{1}{n1}\right)$. For these non-zero covariances, we get, using Cauchy–Schwarz inequality,

$$\begin{aligned}\text{Cov}\left(A_{1kr}^2, A_{12tl}^2\right) &\le \sqrt{\text{Var}\left(A_{1kr}^2\right)\text{Var}\left(A_{12tl}^2\right)} \le \sqrt{\text{E}\left(A_{1kr}^4\right)\text{E}\left(A_{12tl}^4\right)}\\ &\le \sqrt{\text{E}\left(A_{1k}^2\right)\text{E}\left(A_{1r}^2\right)\text{E}\left(A_{1t}^2\right)\text{E}\left(A_{2l}^2\right)}\\ &\le \gamma_1^2\sqrt{\gamma_1 p^2\cdot\gamma_2 p^2} = \gamma^{3/2}\gamma_2 p^4,\end{aligned}$$

so that $\text{Cov}(B_1, B_{12}) \le \gamma^{3/2}\gamma_2 p^4 O\left(\frac{1}{n1}\right)$, which proves the required result, and the proof of $\text{Cov}(B_2, B_{12})$ follows exactly the same lines. This implies that the covariance ratios, $\frac{\text{Cov}(B_1, B_{12})}{\text{tr}(\boldsymbol{\Sigma}_1^2)\text{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)}$ (ignoring p^2s for simplicity) are also uniformly bounded in p, and vanish for any large n_i, $i = 1, 2$.

Substituting these estimators in $\text{Var}(T)$ and denoting the resulting estimated variance as $\widehat{\text{Var}(T)}$, it implies that, under Assumptions 2–4,

$$\frac{T}{\sqrt{\widehat{\text{Var}(T)}}} \xrightarrow{\mathcal{D}} N(1, 1),$$

as $n_1, n_2, p \to \infty$. This completes the proof of the theorem.

Like for the one-sample case, we summarize the moments of the components of the test statistic in the following proposition without assuming normality. The simplified results under normality will be presented and discussed in Sect. 4. Assume, for computational convenience, $H = E_0 + E_1$, without the divisor p.

Proposition 7 *For E_1 and E_0, as defined above, we have*

$$\mathrm{E}(H) = \mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) = \mathrm{E}(E_1) \tag{45}$$

$$\mathrm{Var}(H) = \mathrm{Var}(E_1) + 2\mathrm{tr}\,(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2 \tag{46}$$

$$\mathrm{Cov}(H, E_1) = \mathrm{Var}(E_1). \tag{47}$$

3.2 The paired case

We now extend the statistic to the case when the observations in $\mathbf{X}_{1k}$ and $\mathbf{X}_{2l}$ are paired. For example, the paired observations may denote the measurements before and after administering certain treatment on patients in a clinical experiment. In this context, the p elements of each vector can be the repeated measurements observed on each of n individuals, before and after the treatment. Clearly, the two samples are no longer independent, and a correlation component comes in. Further, the sample sizes have to be the same, $n_1 = n_2$. Let $\mathbf{X}_{ik} = (X_{i1k}, \ldots, X_{ipk})'$, $k = 1, \ldots, n$, $i = 1, 2$, be the sample vectors. The hypothesis to be tested reduces to $H_0 : \boldsymbol{\mu} = \mathbf{0}$, where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X}_{1k} - \mathbf{X}_{2k})$, so that the problem essentially reduces to the one-sample case.

Let $\mathbf{D}_k = \mathbf{X}_{1k} - \mathbf{X}_{2k}$, $k = 1, \ldots, n$, with $\mathbf{D}_k \sim \mathcal{F}_D$ where $\mathcal{F}_D$ denotes some multivariate distribution of the differences. Then, $\mathrm{E}(\mathbf{D}_k) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{D}_k) = \boldsymbol{\Sigma}_{\mathbf{D}}$, with

$$\boldsymbol{\Sigma}_{\mathbf{D}} = \mathrm{Cov}(\mathbf{X}_{1k} - \mathbf{X}_{2k}) = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_{12},$$

where, under H_0, $\boldsymbol{\Sigma}_1 = \mathrm{E}(\mathbf{X}_{1k}\mathbf{X}'_{1k})$, $\boldsymbol{\Sigma}_2 = \mathrm{E}(\mathbf{X}_{2k}\mathbf{X}'_{2k})$, and $\boldsymbol{\Sigma}_{12} = \mathrm{E}(\mathbf{X}_{1k}\mathbf{X}_{2k})$. If we now let $\mathbf{X} = (\mathbf{X}'_{11}, \ldots, \mathbf{X}'_{1n}, \mathbf{X}'_{21}, \ldots, \mathbf{X}'_{2n})'$ denote the vector of all observations, with $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$, then $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1,\ \boldsymbol{\mu}'_2)'$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2 \end{pmatrix} \tag{48}$$

with their respective sample estimators, $\overline{\mathbf{X}} = (\overline{\mathbf{X}}'_{1.},\ \overline{\mathbf{X}}'_{2.})'$ and

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}_1 & \widehat{\boldsymbol{\Sigma}}_{12} \\ \widehat{\boldsymbol{\Sigma}}_{21} & \widehat{\boldsymbol{\Sigma}}_2 \end{pmatrix}. \tag{49}$$

We correspondingly define $\overline{\mathbf{D}}_k = \overline{\mathbf{X}_{1k} - \mathbf{X}_{2k}} = \overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}$, so that $\mathrm{E}(\overline{\mathbf{D}}_k) = \mathbf{0}$ and $\mathrm{Cov}(\overline{\mathbf{D}}_k) = \boldsymbol{\Sigma}_{\overline{\mathbf{D}}}$, where $\boldsymbol{\Sigma}_{\overline{\mathbf{D}}} = \frac{1}{n}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_{12}) = \frac{1}{n}\boldsymbol{\Sigma}_{\mathbf{D}}$. The test statistic (6), using the same notations, is re-defined as

$$T_D = \frac{Q}{E_1}, \tag{50}$$

where, Q and E_1, now being one-sample estimators, take simpler forms as

$$Q = n\left(\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}\right)'\left(\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}\right) = n\overline{\mathbf{D}}_k'\overline{\mathbf{D}}_k \tag{51}$$

$$E_1 = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{X}_{1k} - \mathbf{X}_{2k})'(\mathbf{X}_{1k} - \mathbf{X}_{2k}). \tag{52}$$

In matrix forms, Q and E_1 can be written as

$$Q = \frac{1}{n}\mathbf{X}'\mathbf{M}'(\mathbf{J}_n \otimes \mathbf{I}_p)\mathbf{M}\mathbf{X} \tag{53}$$

$$E_1 = \frac{1}{n}\mathbf{X}'\mathbf{M}'(\mathbf{I}_n \otimes \mathbf{I}_p)\mathbf{M}\mathbf{X}, \tag{54}$$

where $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ is the vector of all $2np$ observations from both samples with $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in})'$, $i = 1, 2$, $\mathbf{J}$ is the matrix of 1s, $\mathbf{I}$ is the identity matrix, and $\mathbf{M} = (\mathbf{I}_n \otimes \mathbf{I}_p \mid -\mathbf{I}_n \otimes \mathbf{I}_p)$. With $\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$, as defined in Eq. (48), and denoting $\mathbf{U} = \mathbf{M}'\mathbf{M}$, $\mathbf{W} = \mathbf{M}'(\mathbf{J}_n \otimes \mathbf{I}_p)\mathbf{M}$, we get

$$\mathbf{U}\boldsymbol{\Sigma} = \mathbf{I}_n \otimes \begin{pmatrix} \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12} \end{pmatrix}$$

$$\mathbf{W}\boldsymbol{\Sigma} = \mathbf{J}_n \otimes \begin{pmatrix} \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_{12} \end{pmatrix}.$$

This immediately gives $\mathrm{tr}(\mathbf{U}\boldsymbol{\Sigma}) = n\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{D}}) = \mathrm{tr}(\mathbf{W}\boldsymbol{\Sigma})$, $\mathrm{tr}(\mathbf{U}\boldsymbol{\Sigma})^2 = n\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{D}})^2$, and $\mathrm{tr}(\mathbf{W}\boldsymbol{\Sigma})^2 = n^2\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{D}})$, which further leads to the moments of Q and E_1 as $\mathrm{E}(Q) = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{D}}) = \mathrm{E}(E_1)$, and under normality, $\mathrm{Var}(Q) = 2\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{D}}^2)$, and $\mathrm{Var}(E_1) = \frac{2}{n}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{D}}^2)$. Then, if we continue to assume normality and keep p fixed where $n \to \infty$, then it can be immediately shown that the approximation in (8) is also valid for T_D, i.e.,

$$T_D \sim \chi_f^2/f, \tag{55}$$

where $f = [\mathrm{tr}(\boldsymbol{\Sigma}_D)]^2/\mathrm{tr}(\boldsymbol{\Sigma}_D^2)$, estimated as $E_{2(D)}/E_{3(D)}$, with $E_{2(D)}$ and $E_{3(D)}$ defined as

$$E_{2(D)} = \frac{1}{n(n-1)}\sum_{\substack{k=1 \\ k \neq l}}^{n}\sum_{l=1}^{n}A_{k(D)}A_{l(D)} \tag{56}$$

$$E_{3(D)} = \frac{1}{n(n-1)}\sum_{\substack{k=1 \\ k \neq l}}^{n}\sum_{l=1}^{n}A_{kl(D)}^2. \tag{57}$$

Here, the subscript D in parentheses indicates that the expressions are defined for the paired differences. Hence, $A_{k(D)} = \mathbf{D}_k'\mathbf{D}_k$, $A_{l(D)} = \mathbf{D}_l'\mathbf{D}_l$ are the quadratic forms, and $A_{kl(D)} = \mathbf{D}_k'\mathbf{D}_l$ is a symmetric bilinear form. For further details, see Sect. 4.

In the following, we relax normality assumption and show that the statistic T_D is still valid even under high-dimensional asymptotics, i.e., when both $n \to \infty$ and $p \to \infty$. For this, we set the following assumptions which are a special case of Assumptions 2–4 when the data are paired.

Assumption 8 $\mathrm{E}(D_{ks}^4) \leq \gamma < \infty \, \forall \, s = 1, \ldots, p$, for some γ, where $D_{ks} = X_{1ks} - X_{2ks}$.

Assumption 9 For $p \to \infty$, let $\frac{\mathrm{tr}(\boldsymbol{\Sigma}_D)}{p} = O(1)$.

Assumption 10 For $p \to \infty$, let $\frac{\mathrm{tr}(\boldsymbol{\Sigma}_D^2)}{p^2} = O(\delta)$, where $0 < \delta \leq 1$.

We prove the following theorem, as a special case of Theorem 5 when the two samples represent paired data.

Theorem 11 *Given Assumptions 8, 9, and 10. Then, under H_0, the test statistic T_D, defined in Eq. (50), follows the same scaled Chi-square approximation as in (55), when $n, p \to \infty$. Further, under the same setup, and using Lemma 6, it can be shown that,*

$$\frac{T}{\sqrt{\widehat{\mathrm{Var}(T_D)}}} \xrightarrow{\mathcal{D}} N(1, 1),$$

where $\widehat{\mathrm{Var}(T_D)}$ is the sample estimator of $\mathrm{Var}(T_D)$.

Proof The proof follows the same lines as for the independent case, see also Ahmad et al. (2012a). We, therefore, briefly sketch the main steps. We begin by re-writing Q as

$$\begin{aligned} Q = n\left(\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}\right)'\left(\overline{\mathbf{X}}_{1.} - \overline{\mathbf{X}}_{2.}\right) &= \frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{n}(\mathbf{X}_{1k} - \mathbf{X}_{2k})'(\mathbf{X}_{1l} - \mathbf{X}_{2l}) \\ &= \frac{1}{n}\sum_{k=1}^{n}(\mathbf{X}_{1k} - \mathbf{X}_{2k})'(\mathbf{X}_{1k} - \mathbf{X}_{2k}) + \frac{1}{n}\sum_{\substack{k=1 \\ k\neq l}}^{n}\sum_{l=1}^{n}(\mathbf{X}_{1k} - \mathbf{X}_{2k})'(\mathbf{X}_{1l} - \mathbf{X}_{2l}) \\ &= \frac{1}{n}\sum_{k=1}^{n}A_k + \frac{1}{n}\sum_{\substack{k=1 \\ k\neq l}}^{n}\sum_{l=1}^{n}A_{kl} = E_1 + \frac{1}{n}\sum_{\substack{k=1 \\ k\neq l}}^{n}\sum_{l=1}^{n}A_{kl}, \end{aligned} \tag{58}$$

where $A_k = (\mathbf{X}_{1k} - \mathbf{X}_{2k})'(\mathbf{X}_{1k} - \mathbf{X}_{2k})$ and $A_{kl} = (\mathbf{X}_{1k} - \mathbf{X}_{2k})'(\mathbf{X}_{1l} - \mathbf{X}_{2l})$. Clearly, $\frac{1}{p}E_1 \xrightarrow{\mathcal{P}} \frac{1}{p}\mathrm{tr}(\boldsymbol{\Sigma}_D) = O(1)$, under Assumptions 8 and 9, following the same arguments as for the independent case. The test statistic then reduces to

$$T_D = \frac{Q}{E_1} = 1 + \frac{E_0}{\frac{1}{p}E_1}. \tag{59}$$

We write

$$E_0 = \frac{1}{np}\sum_{\substack{k=1\\k\neq l}}^{n}\sum_{r=1}^{n} A_{kl}$$

$$= (n-1)\left(\frac{1}{n(n-1)}\sum_{\substack{k=1\\k\neq l}}^{n}\sum_{r=1}^{n}\frac{1}{p}A_{kl}\right) = (n-1)U_n, \tag{60}$$

where, U_n is a U-statistic with $A_{kl} = (\mathbf{X}_{1k} - \mathbf{X}_{2k})'(\mathbf{X}_{1l} - \mathbf{X}_{2l}), k \neq l$. Now, T_D and its components are the one-sample expressions similar to the ones dealt with in Ahmad et al. (2012a). Then, following their proof, we immediately obtain (see Ahmad et al. 2012a, p. 8)

$$nU_n \xrightarrow{\mathcal{D}} \sum_{j=1}^{\infty} \nu_j (C_j - 1),$$

as $n, p \to \infty$ where C_j are independent χ^2_1 random variables and ν_j are the eigenvalues of $\frac{1}{p}\mathbf{\Sigma}_D$. Further, by Lemma 6, we get (see Ahmad et al. 2012a, p. 10)

$$\frac{\sum_{j=1}^{p} \nu_j (C_j - 1)}{\sqrt{2\sum_{j=1}^{p} \nu_j^2}} \xrightarrow{\mathcal{D}} N(0, 1),$$

or that

$$nU_n \xrightarrow{\mathcal{D}} N\left(0, \frac{2\mathrm{tr}(\mathbf{\Sigma}_D^2)}{p^2}\right),$$

as $n, p \to \infty$. Finally, defining

$$B_D = \frac{1}{n(n-1)}\sum_{\substack{k=1\\k\neq l}}^{n}\sum_{r=1}^{n}\frac{1}{p^2}A_{kl}^2$$

as an unbiased and consistent estimator of $\frac{\mathrm{tr}(\mathbf{\Sigma}_D^2)}{p^2}$, it follows that

$$\frac{T_D}{\sqrt{\widehat{\mathrm{Var}(T_D)}}} \xrightarrow{\mathcal{D}} N(1, 1),$$

as $n, p \to \infty$, where $\widehat{\mathrm{Var}(T_D)}$ is an estimator of $\mathrm{Var}(T_D)$ when B_D is substituted for $\frac{\mathrm{tr}(\mathbf{\Sigma}_D^2)}{p^2}$. This completes the proof of the theorem. □

By the same arguments, Proposition 5 in Ahmad et al. (2012a) is also valid for the paired case, with the only difference that $\mathbf{\Sigma}$ in the one-sample case is now replaced with $\mathbf{\Sigma_D} = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 - 2\mathbf{\Sigma}_{12}$, so that, without normality assumption we have the following for Q and E_1 as defined in Eqs. (51) and (52) above.

Proposition 12 *For Q and E_1, as defined above, we have*

$$\mathrm{E}(Q) = \mathrm{tr}(\mathbf{\Sigma_D}) = \mathrm{E}(E_1) \tag{61}$$

$$\mathrm{Var}(Q) = \mathrm{Var}(E_1) + \frac{2(n-1)}{n}\mathrm{tr}(\mathbf{\Sigma_D^2}) \tag{62}$$

$$\mathrm{Cov}(Q, E_1) = \mathrm{Var}(E_1). \tag{63}$$

4 The normal case

If we assume $\mathcal{F}_1, \mathcal{F}_2$ to be multivariate normal, with their respective means and variances, then the results simplify to a large extent. The approximating limit distribution of T, however, remains exactly the same. Further, the moments of E_0 are also the same, with or without normality assumption. The only difference is that we can compute exact variance of E_1 under normality. For convenience, we summarize all the results in the following for further reference.

Lemma 13 *Let E_0 and E_1 be as defined in Sect. 3.1. Then*

$$\mathrm{E}(E_1) = \mathrm{tr}(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) = \mathrm{E}(H) \tag{64}$$

$$\mathrm{Var}(E_1) = 2\mathrm{tr}\left(\frac{1}{n_1}\mathbf{\Sigma}_1^2 + \frac{1}{n_2}\mathbf{\Sigma}_2^2\right) \tag{65}$$

$$\mathrm{Var}(H) = 2\mathrm{tr}\,(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^2 = \mathrm{Var}(E_1) + \mathrm{Var}(E_0) \tag{66}$$

$$\mathrm{Cov}(H, E_1) = \mathrm{Var}(E_1). \tag{67}$$

We note that, given the independence of the two samples, the moments in Lemma 13 are a direct extension of the corresponding one-sample moments; see, for example, Theorems 2.4 and 2.5 in Ahmad (2008). Further, compare these moments with those in Eqs. (15–17). Both results are derived under the assumption of normality. It can be immediately deduced that although the expectations are the same, the variances have drastically reduced with the new components of the test statistic. Additionally, the original normality-based computations in Ahmad (2008, Chapter 3) are done only under the two special cases of $n_1 = n_2$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, whereas the results in Lemma 13 hold in general. That the variances with the new components in Lemma 13 are significantly smaller for the two special cases can be witnessed when the results are compared with those reported in Ahmad (2008, Sections 3.2.3, 3.2.4).

Based on the results of Lemma 13, the probability convergence of $\frac{1}{p}E_1$ can be trivially shown, under Assumptions 3 and 4. The asymptotic limit distribution of T also remains the same as given in Theorem 5, including the normal approximation. For normal approximation, however, we can now prove the consistency of B_1, B_2, B_{12}

more precisely since the exact moments of these estimators can be computed under normality. These moments are summarized in the following lemma without proof for reference. The proof of variances can be directly followed from the proof of Theorem 2.4 in Ahmad (2008), and the proof of covariances trivially follows from the independence of the samples.

Lemma 14 *Let B_1, B_2 and B_{12} be as defined in Eqs. (41–43). Then*

$$\mathrm{Var}(B_1) = \frac{4}{n_1(n_1-1)}\left[(2n_1-1)\mathrm{tr}(\boldsymbol{\Sigma}_1^4) + \left[\mathrm{tr}\left(\boldsymbol{\Sigma}_1^2\right)\right]^2\right] \tag{68}$$

$$\mathrm{Var}(B_2) = \frac{4}{n_2(n_2-1)}\left[(2n_2-1)\mathrm{tr}(\boldsymbol{\Sigma}_2^4) + \left[\mathrm{tr}\left(\boldsymbol{\Sigma}_2^2\right)\right]^2\right] \tag{69}$$

$$\mathrm{Var}(B_{12}) = \frac{1}{n_1n_2}\left[6\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) + 2\left[\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\right]^2\right] \tag{70}$$

$$\mathrm{Cov}(B_1, B_2) = 0 \tag{71}$$

$$\mathrm{Cov}(B_1, B_{12}) = \frac{4}{n_1}\mathrm{tr}\left(\boldsymbol{\Sigma}_1^3\boldsymbol{\Sigma}_2\right) \tag{72}$$

$$\mathrm{Cov}(B_2, B_{12}) = \frac{4}{n_2}\mathrm{tr}\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^3\right). \tag{73}$$

Based on these moments, it immediately follows that $\mathrm{Var}\left(B_1/\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\right)$ and $\mathrm{Var}\left(B_2/\mathrm{tr}(\boldsymbol{\Sigma}_2^2)\right)$ are uniformly bounded by $O(1/n_1)$ and $O(1/n_2)$, and same bounds are attained by the covariance ratios $\mathrm{Cov}(B_1, B_{12})/\left[\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\right]$ and $\mathrm{Cov}(B_2, B_{12})/\left[\mathrm{tr}(\boldsymbol{\Sigma}_2^2)\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2)\right]$, respectively, where $\mathrm{Var}\,(B_{12}/\mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2))$ is bounded by $O(1/n_1n_2)$. This proves Theorem 5, assuming normality.

When p is assumed fixed but large, we can follow the same strategy as in Ahmad (2008, Chapter 3) to compute the moments of T using the delta method. We get the following moments, where $f = [\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2/\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2$.

$$\mathrm{E}(T) = 1 \tag{74}$$

$$\begin{aligned}
\mathrm{Var}(T) &= \frac{2\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2}{[\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2} + \frac{2\mathrm{tr}\left(\frac{1}{n_1}\boldsymbol{\Sigma}_1^2 + \frac{1}{n_2}\boldsymbol{\Sigma}_2^2\right)}{[\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2} \\
&\quad - \frac{4\mathrm{tr}\left(\frac{1}{n_1}\boldsymbol{\Sigma}_1^2 + \frac{1}{n_2}\boldsymbol{\Sigma}_2^2\right)}{[\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2} \\
&= \frac{2\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2}{[\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2}\left(1 - \frac{\mathrm{tr}\left(\frac{1}{n_1}\boldsymbol{\Sigma}_1^2 + \frac{1}{n_2}\boldsymbol{\Sigma}_2^2\right)}{\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2}\right) \\
&= \frac{2}{f}\left(1 - \frac{1}{n_1}\cdot\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2} - \frac{1}{n_2}\cdot\frac{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2}\right) \\
&= \frac{2}{f}\left(1 - \frac{1}{n}\cdot\frac{\mathrm{tr}(\boldsymbol{\Sigma}_1^2 + \boldsymbol{\Sigma}_2^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2}\right) \qquad (n_1 = n_2)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{f}\left(1 - \frac{N}{4n_1n_2}\right) \qquad (\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2)\\
&= \frac{2}{f}\left(1 - \frac{1}{2n}\right) \qquad (n_1 = n_2,\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2). \qquad (75)
\end{aligned}
$$

Clearly, the last two terms inside parentheses, in Eq. (75), vanish when $n_1, n_2 \to \infty$ for any fixed p. This approximates the moments of T with those of χ^2_f/f, as required, where f is given in (9); see also (18). Here, we additionally need to estimate f, i.e., estimate $[\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^2$ and $\mathrm{tr}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^2$ to make T practically applicable. The unbiased estimators of these two traces are given as E_2 and E_3, respectively, as defined in Eqs. (11) and (12), so that f is estimated as E_2/E_3. The properties of these estimators follow directly from Ahmad (2008, Chapter 3). Further, $\mathrm{tr}(\boldsymbol{\Sigma}_1^2)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_2^2)$ in $\mathrm{Var}(T)$ in Eq. (75) can be replaced with their unbiased and consistent estimators B_1 and B_2 from Eqs. (41) and (42), respectively.

Finally, the validity of the paired case of Sect. 3.2 follows exactly the same way when $\mathcal{F}_D$ is assumed normal. The moments of the components of the test statistic in Eq. (50), as reported in Proposition 12, simplifies, under normality, to the following.

Proposition 15 *For Q and E_1, as defined in Sect. 3.2, we have*

$$\mathrm{E}(Q) = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{D}}) = \mathrm{E}(E_1) \qquad (76)$$

$$\mathrm{Var}(Q) = 2\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{D}}^2\right) \qquad (77)$$

$$\mathrm{Var}(E_1) = \frac{2}{n}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\mathbf{D}}^2\right) = \mathrm{Cov}(Q, E_1). \qquad (78)$$

The results can be compared with the similar moments for the one-sample case as reported in Ahmad (2008, Theorem 2.5) or in Ahmad et al. (2008, Proposition 5) where $\boldsymbol{\Sigma}$ is now replaced with $\boldsymbol{\Sigma}_{\mathbf{D}} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_{12}$. Using these moments, the probability convergence of $\frac{1}{p}E_1$ can be immediately shown, under Assumptions 8 and 9. The asymptotic limit of T in Eq. (50) remains the same as given in Theorem 11, including the normal approximation. Additionally, under normality, we can prove the consistency of B_D more precisely, since the $\mathrm{Var}(B_D)$ can be exactly computed which, following Lemma 14, is

$$\mathrm{Var}(B_D) = \frac{4}{n(n-1)}\left[(2n-1)\mathrm{tr}\left(\boldsymbol{\Sigma}_D^4\right) + \left[\mathrm{tr}\left(\boldsymbol{\Sigma}_D^2\right)\right]^2\right]. \qquad (79)$$

Then, it immediately follows that $\mathrm{Var}(B_D/\boldsymbol{\Sigma}_D^2)$ is uniformly bounded by $O(1/n)$, independent of p, which establishes the validity of Theorem 11 under normality.

Further, if p is assumed fixed but large, we can again use the delta method approach as in Ahmad (2008, Chapter 2) to compute the following moments of T.

$$\mathrm{E}(T) = 1 \tag{80}$$

$$\begin{aligned}\mathrm{Var}(T) &= \frac{2\mathrm{tr}\left(\mathbf{\Sigma}_D^2\right)}{[\mathrm{tr}(\mathbf{\Sigma}_D)]^2} + \frac{\frac{2}{n}\mathrm{tr}\left(\mathbf{\Sigma}_D^2\right)}{[\mathrm{tr}(\mathbf{\Sigma}_D)]^2} - \frac{\frac{4}{n}\mathrm{tr}\left(\mathbf{\Sigma}_D^2\right)}{[\mathrm{tr}(\mathbf{\Sigma}_D)]^2}\\ &= \frac{2}{f}\left(1 - \frac{1}{n}\right),\end{aligned} \tag{81}$$

where $f = [\mathrm{tr}(\mathbf{\Sigma}_D)]^2 / \mathrm{tr}(\mathbf{\Sigma}_D^2)$, estimated as $E_{2(D)}/E_{3(D)}$ using Eqs. (56) and (57). The consistency of $E_{2(D)}$ and $E_{3(D)}$ follow immediately from the results in Ahmad (2008, Chapter 2); see also Ahmad et al. (2008). The detailed study of these normal cases, with an extension to cover the testing of any general linear hypothesis, is left for another manuscript which is in preparation.

5 Some remarks

5.1 Box's approximation

Following the main objective of this paper, it is shown in Sects. 3.1 and 3.2 that the test statistic T, as given in Eq. (26), follows the scaled Chi-square approximation, (8), when $n_1, n_2, p \to \infty$ and the underlying multivariate distributions are not necessarily normal. Clearly, the validity of approximation for T under the normality assumption comes as a special case, as shown in Sect. 4. Further, as mentioned in Sect. 3.1 (see the second comment immediately after Eqs. 38 and 39), T can also be used if we continue to relax normality but keep p fixed, including when $p > n_i$. Together with the original normality-based approximation for fixed p, i.e., (8), these results imply that the test statistic T, and hence the Box's approximation, are valid whether $\mathcal{F}_1$ and $\mathcal{F}_2$ in (1) are multivariate normal or not, and whether the dimension of the multivariate vector, p, is kept fixed or is allowed to grow arbitrarily with n. For further reference, we summarize these results in the following two theorems, one for the Box's approximation, and one for the statistic T.

Theorem 16 *(**Theorem 1**, **revisited**) Consider Model (1). Let $\mathbf{\Sigma}_N = \frac{2n_1n_2}{N}\mathbf{\Sigma}_0$, $N = n_1 + n_2$, and $\overline{\mathbf{Y}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{Y}_{ij}$, $i = 1, 2$, where $\mathbf{\Sigma}_0$ is defined in Sect. 2 and $\mathbf{Y}_{ij}$ are defined in Sect. 3.1. Further, let λ_j and ν_j be as defined in Sect. 3.1. Then, the following hold under the respective conditions where, in each case, C_j represents an iid χ_1^2 random variable.*

1. *Assume p fixed and $n_1, n_2 \to \infty$. If (a) $\mathcal{F}_1$ and $\mathcal{F}_2$ are multivariate normal, or (b) $\mathcal{F}_1$ and $\mathcal{F}_2$ are as in (1) for which Assumptions 2–3 are satisfied, then*

$$\frac{2n_1n_2}{N}\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right)'\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) \approx \sum_{j=1}^{p}\lambda_j C_j.$$

2. *Assume $p \to \infty$ and $n_1, n_2 \to \infty$. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are as in (1) for which Assumptions 2–4 are satisfied. Then (see Eq. 37)*

$$\frac{2n_1n_2}{N}\left(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2\right)'\left(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2\right) \approx \sum_{j=1}^{\infty} \nu_j C_j.$$

Theorem 17 (The modified test statistic) *Consider Model (1). Let λ_j and ν_j be as defined in Sect. 3.1. Then, the following hold under the respective conditions.*

1. *Assume p fixed and $n_1, n_2 \to \infty$. If (a) $\mathcal{F}_1$ and $\mathcal{F}_2$ are multivariate normal, or (b) $\mathcal{F}_1$ and $\mathcal{F}_2$ are as in (1) for which Assumptions 2–3 are satisfied, then, for the test statistic A_N, as defined in Eq. (6),*

$$f A_N \xrightarrow{\mathcal{D}} \chi^2_f,$$

where $f = [\sum_{j=1}^p \lambda_j]^2 / \sum_{j=1}^p \lambda_j^2$.

2. *Assume $p \to \infty$ and $n_1, n_2 \to \infty$. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are as in (1) for which Assumptions 2–4 are satisfied, then, for the test statistic T, as defined in Eq. (26),*

$$f T \xrightarrow{\mathcal{D}} \chi^2_f,$$

where $f = [\sum_{j=1}^\infty \nu_j]^2 / \sum_{j=1}^\infty \nu_j^2$.

Replacing $\mathcal{F}_i$, $i = 1, 2$, in Theorems 16 and 17 with $\mathcal{F}_D$, and using other corresponding notations from Sect. 3.2, similar theorems can also be stated for the paired case.

5.2 Degenerate U-statistics and high-dimensional inference

The class of U-statistics, as introduced by Hoeffding (1948), is rich and encompasses a wide variety of statistics. Some comprehensive book-length references dealing with core theory include Lee (1990), Koroljuk and Borovskich (1994), whereas Kowalski and Tu (2008) is a nice exposition for applications of U-statistics theory in a variety of settings, including, for example, linear mixed models, where a similar application specifically for repeated measures design can be found in Davis (2002).

In particular, the asymptotic theory of U-statistics has attracted a lot of researchers both in theoretical and applied areas. A nice monograph dealing with asymptotic theory of both degenerate and non-degenerate U-statistics is Denker (1985), see also Denker and Keller (1983). A recent work of Denker and Gordin (2011) focuses particularly on von Mises statistics also valid for non-iid data. Of particular relevance for us in Denker and Gordin (2011)'s report is Section 8 which specifically deals with kernels of degree 2; see also an exhaustive list of references cited therein. A functional central limit theorem specifically for two-sample non-degenerate U-processes is given in Neumeyer (2004), whereas a functional version of almost sure central limit theorem for both degenerate and non-degenerate U-statistics is given in Holzmann et al. (2004).

An old, and one of the most frequently cited references for functional central limit theorems for degenerate U-statistics, is Neuhaus (1977).

Recently, people have attempted to reinvigorate the asymptotic theory of U-statistics, particularly in the degenerate case, with a focus on its application to address more complicated practical problems, for example, to deal with real inferential problems in statistics. A nice exposition, dealing with the degenerate U- and V-statistics for stationary random variables is discussed in Leucht (2012) wherein applications in hypothesis testing, including a small sample bootstrap connection, is shown. The present paper addresses the question of the application of degenerate U-statistics theory for high-dimensional problems. No mentionable bibliography can be traced on this, but it seems interesting that this issue has recently started emerging. A general discussion is given in Pinheiro et al. (2009), where a formal application of non-degenerate U-statistics in high-dimensional testing is given in Ahmad et al. (2012b).

In this sense, the present paper can be considered as a new and motivating effort to show that the theory of degenerate U-statistics is applicable to tackle the problem of high-dimensional asymptotics. A similar work for one-sample test of a multivariate mean vector under high-dimensional setup is presented in Ahmad et al. (2012a). This aspect puts the paper in a different spectrum as compared to other recent approaches to deal with high-dimensional inference, in particular testing. Our motivation of using the asymptotic theory of degenerate U-statistics for high-dimensional testing stems from two important aspects which are highlighted below.

First, the main condition for a valid asymptotic U-statistics limit, i.e., the weighted sum of iid χ^2_1 random variables, is that the kernel is square-integrable, i.e., $\mathrm{E}(h^2(\cdot))<\infty$. This, in the test statistics presented above, is ensured by norming the kernel by p, supplemented by Assumptions 4 and 10. Recall that, for the computations presented in Sect. 3, all degenerate U-statistics are of degree 2, disregarding whether they are composed of a single sample, of paired samples, or of two independent samples. Then, it is important to note that, the theory of degenerate U-statistics takes a special place for second degree kernels, as nicely elaborated in van der Vaart (1998, Section 12.3). This leads us to the second motivating aspect which we particularly want to emphasize in the context of second degree kernels. For convenience, we consider $h(\mathbf{X}, \mathbf{Y})$ as a general kernel of degree 2, which, for example, for U_{n_1} may refer to $\frac{1}{p}h(\mathbf{X}_{1k}, \mathbf{X}_{1r}) = \frac{1}{p}A_{1kr}$.

Suppose that the data vectors $\mathbf{X}$, $\mathbf{Y}$ are generated by the probability space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, and suppose that T is a bounded, compact, linear operator such that $T : (\mathcal{X}, \mathcal{A}, \mathcal{P}) \rightarrow (\mathcal{X}, \mathcal{A}, \mathcal{P})$. Note that, since we are only dealing with quadratic and bilinear forms, we can, without any loss of generality, assume $\mathcal{X}$ to be Hilbert space (which, under the given conditions, can actually be assumed as separable), and $\mathcal{A}$ be the space of all inner products generated from $\mathcal{X}$. Since $h(\cdot)$ is degenerate, symmetric, and square-integrable, T is a self-adjoint Hilbert–Schmidt operator such that

$$Tf(\mathbf{x}) = \int h(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) dP(\mathbf{y}),$$

and the kernel can be expanded as an infinite series as

$$h(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \lambda_j f_j(\mathbf{x}) f_j(\mathbf{y}) \tag{82}$$

where λ_js and f_js provide an orthonormal basis for the kernel. Note that, when we let $p \to \infty$, we let the kernel, as a bilinear form composed of independent random vectors, to expand similarly. However, for an asymptotic convergence, we need to control the behavior of such an infinite sum, and this is accomplished by norming the kernel by p, as explained in Sect. 3.1. Moreover, the convergence of this p-scaled kernel to the orthogonal expansion in (82) is in mean-square, i.e., L_2-convergence given as

$$\mathrm{E}\left(h(\mathbf{x}, \mathbf{y}) - \sum_{j=1}^{p} \lambda_j f_j(\mathbf{x}) f_j(\mathbf{y})\right)^2 = \sum_{j=p+1}^{\infty} \lambda_j^2 \to 0. \tag{83}$$

Note the final convergence, where the sum is composed of only at most countably many eigenvalues (van der Vaart 1998, p. 168). Finally, the vanishing of the eigenvalues for such a convergence to hold is guaranteed by the well-known Hilbert–Schmidt theorem which we state in the following for reference (see, for example, Reed and Simon 1980, Theorem VI.16, p. 203).

Theorem 18 (Hilbert–Schmidt Theorem) *Let h be a self-adjoint compact operator on Hilbert space $\mathcal{X}$. Then, there is a complete orthonormal basis, $\{f_p\}$, for $\mathcal{X}$ such that $hf_p = \lambda_p f_p$, and $\lambda_p \to 0$ as $p \to \infty$.*

This result, under the Assumptions 2–4 (for independent samples) or 8–10 (for paired samples), gives a proper justification for the asymptotic limit distributions of the test statistics in Sect. 3. A more mathematical treatment of this subject is part of another manuscript.

A detailed discussion of operators like T, particularly in reference to their properties for Hilbert spaces, is given in Kreyszig (1978), whereas an older, classic reference is Dunford and Schwartz (1963); see also Masujima (2009). For a very relevant use of this theory for degenerate U-statistics, see Serfling (1980, Chapter 5) Lee (1990, Chapter 3) and Koroljuk and Borovskich (1994, Chapter 4).

6 Discussion and conclusions

A two-sample test statistic is presented to test the difference of multivariate mean vectors when the dimension of the vector may exceed the sample size and the random vectors may not necessarily come from a multivariate normal distribution. The statistic is first evaluated under the standard high-dimensional aymptotics, assuming both $n_i \to \infty$, $i = 1, 2$, and $p \to \infty$. The validity of the statistic for fixed p (where $p > n_i$) is briefly discussed as a special case. Both independent and paired cases are dealt with. Finally, the statistic and its approximation are also briefly summarized when the assumption of normality holds. The details of the later two cases are, however, postponed for a separate manuscript.

The paper extends a one-sample statistic presented in Ahmad et al. (2012a) which is evaluated in similar contexts. A salient feature of the two papers is an extension of the well-known Box's approximation to high-dimensional and non-normal setup and, for the present two-sample case, also under the Behrens–Fisher setting.

The statistic and its approximation can obviously be used for any general linear hypothesis, say $H_0 : \mathbf{H}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$, by appropriately defining the hypothesis matrix $\mathbf{H}$, although in this paper, for simplicity, $\mathbf{H} = \mathbf{I}$ is used. For example, with $\mathbf{I}$ as identity matrix and $\mathbf{J}$ as matrix of ones, the matrix $\mathbf{H} = \mathbf{C}_2 \otimes \mathbf{C}_p$ leads to a test of interaction effect (parallel profiles) hypothesis, where $\mathbf{C}_p = \mathbf{I} - \frac{1}{p}\mathbf{J}_p$ (similarly $\mathbf{C}_2$ for $p = 2$). Since a complete profile analysis involves other types of hypothesis, for example, hypothesis for the time effect, these general linear hypothesis cases are not given a detailed treatment here for the sake of brevity. For details of such high-dimensional profile analysis under normality assumption, see Ahmad (2008, Chapter 4).

Acknowledgments The author is sincerely thankful to Prof. Dr. Anne Leucht, Theoretical Econometrics and Statistics, Department of Economics, University of Mannheim, Germany, for her careful perusal of the manuscript and helpful suggestions. Thanks are also due to the editor, the associate editor, and to two anonymous referees for their helpful suggestions and comments which lead to this improved version of the article.

References

Ahmad, M. R. (2008). *Analysis of high-dimensional repeated measures designs: The one- and two-sample test statistics*. Ph.D. Thesis. Göttingen: Cuvillier Verlag.

Ahmad, M. R., Werner, C., Brunner, E. (2008). Analysis of high dimensional repeated measures designs: The one sample case. *Computational Statistics & Data Analysis*, *53*, 416–427.

Ahmad, M. R., von Rosen, D., Singull, M. (2012a). A note on mean testing for high-dimensional multivariate data under non-normality. *Statistica Neerlandica*, *67*(1), 81–99.

Ahmad, M. R., Yamada, T., von Rosen, D. (2012b). Tests of covariance matrices for high-dimensional multivariate data. Test (Submitted).

Bai, Z., Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, *6*, 311–329.

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. *Annals of mathematical statistics*, *25*, 290–302.

Chen, S. X., Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, *38*(2), 808–835.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York: Springer.

Dempster, A. P. (1958). A high dimensional two sample significance test. *Annals of Mathematical Statistics*, *29*(4), 995–1010.

Dempster, A. P. (1969). *Elements of continuous multivariate analysis*. MA: Addison-Wesley.

Denker, M. (1985). *Asymptotic distribution theory in nonparametric statistics*. Vieweg, Braunschweig: Vieweg Advanced Lectures.

Denker, M., Gordin, M. (2011). Limit theorems for von Mises statistics of a measure preserving, transformation. arXiv:1109.0635v1[math.DS]. September 3, 2011.

Denker, M., Keller, G. (1983). On U-statistics and v. Mises' statistics for weakly dependent processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, *64*, 505–522.

Dunford, N., Schwartz, J. T. (1967). *Linear operators. Part II: Spectral theory-Self-adjoint operators in Hilbert space*. New York: Wiley.

Fujikoshi, Y., Ulyanov, V. V., Shimizu, R. (2010). *Multivariate statistics: High-dimensional and large-sample approximations*. New York: Wiley.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., Smola, A. (2008). A kernel method for the two-sample problem. *Journal of Machine Learning Research*, *1*, 1–43.

Hájek, J., Šidák, Z., Sen, P. K. (1999). *Theory of rank tests*. San Diego: Academic Press.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, *19*, 293–325.

Holzmann, H., Koch, S., Min, A. (2004). Almost sure limit theorems for U-statistics. *Statistics & Probability Letters*, *69*, 261–269.

Jiang, J. (2010). *Large sample techniques for statistics*. New York: Springer.

Koroljuk, V. S., Borovskich, Y. V. (1994). *Theory of U-statistics*. Dordrecht: Kluwer.

Kowalski, J., Tu, X. M. (2008). *Modern applied U-statistics*. New York: Wiley.

Kreyszig, E. (1978). *Introductory functional analysis with applications*. New York: Wiley.

Ledoit, O., Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, *30*, 1081–1102.

Lee, A. J. (1990). *U-statistics: Theory and practice*. Boca Raton: CRC Press.

Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.

Leucht, A. (2012). Degenerate U- and V-statistics under weak dependence: Asymptotic theory and bootstrap consistency. *Bernoulli*, *18*, 552–585.

Masujima, M. (2009). *Applied mathematical methods in theoretical physics* (2nd ed.). New York: Wiley.

Neuhaus, G. (1977). Functional limit theorems for U-statistics in the degenerate case. *Journal of Multivariate Analysis*, *7*, 424–439.

Neumeyer, N. (2004). A central limit theorem for two-sample U-processes. *Statistics & Probability Letters*, *67*, 73–85.

Pinheiro, A., Sen, P. K., Pinheiro, H. P. (2009). Decomposibility of high-dimensional diversity measures: Quasi-U-statistics, martigales, and nonstandard asymptotics. *Journal of Multivariate Statistics*, *100*, 1645–1656.

Reed, M., Simon, B. (1980). *Methods of modern mathematical physics, Vol. I: Functional analysis*. San Diego, CA: Academic Press.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Weinheim: Wiley.

Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of Japan Statistical Association*, *37*, 53–86.

Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *Journal of Multivariate Analysis*, *100*, 518–532.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.