MULTIVARIATE BEHAVIORAL RESEARCH, *41*(4), 533–548

On Selecting Tests for Equality of Two Normal Mean Vectors

K. Krishnamoorthy and Yanping Xia
Department of Mathematics
University of Louisiana at Lafayette

The conventional approach for testing the equality of two normal mean vectors is to test first the equality of covariance matrices, and if the equality assumption is tenable, then use the two-sample Hotelling T^2 test. Otherwise one can use one of the approximate tests for the multivariate Behrens–Fisher problem. In this article, we study the properties of the Hotelling T^2 test, the conventional approach, and one of the best approximate invariant tests (Krishnamoorthy & Yu, 2004) for the Behrens–Fisher problem. Our simulation studies indicated that the conventional approach often leads to inflated Type I error rates. The approximate test not only controls Type I error rates very satisfactorily when covariance matrices were arbitrary but was also comparable with the T^2 test when covariance matrices were equal.

In this article, we are concerned about the choice of a test for equality of two normal mean vectors. The main purpose of this article is to bring to the attention of applied researchers a satisfactory test that can be used for testing the equality of two normal mean vectors when the population covariance matrices are unknown and arbitrary. Because the tests and the approaches that we discuss are generalizations of the univariate results, for convenience and easy reference, we first review the results for the univariate normal case.

A normal distribution is the most common choice for modeling the data that involve univariate observations. This is especially true in situations where one is interested in estimating or testing the mean of a population. This is because the sampling distribution of the mean of a random sample is approximately normal, regardless of the actual distribution of the data, due to the central limit theorem. If the problem of interest is to test a hypothesis about the population mean based on a

Correspondence concerning this article should be sent to K. Krishnamoorthy, Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504. E-mail: krishna@louisiana.edu

random sample, then Student's t test is commonly used. The t test is also applicable to test a hypothesis about the difference between the means of two normal populations; however, it is valid only when the population variances are equal. The test may become seriously biased when the variance equality assumption is not satisfied, resulting in spurious decisions about the null hypothesis. Furthermore, the assumption of variance homogeneity is very unlikely to be satisfied in practice.

The problem of making inference about the difference between two normal means without assuming equality of population variances is referred to as the Behrens–Fisher problem, and it has been well addressed in the univariate case. Scheffé (1943) proposed an exact method, but it has some serious drawbacks; it involves random pairing of observations of two independent samples, and hence order of the observations affects the test results. For this reason, Scheffé (1970) himself recommended that this exact method should not be used. Several authors proposed approximate solutions. A popular approximate solution by Welch (1947) is based on Student's t distribution with degrees of freedom depending on not only the sample sizes but also the sample variances. Nevertheless, this approach has been well accepted and commonly used in practical applications because of its simplicity and accuracy. Some software (Minitab, Excel) and calculators (TI–83) use this approach to test and interval estimate the difference between two normal means.

A conventional approach, regarding the choice between the t and Welch's (1947) tests, is first to test the equality of variances, and if the equality is tenable then use the t test, otherwise use the Welch test. Size (Type I error rate) and power studies of this conventional approach by Moser, Stevens, and Watts (1989) indicated that the preliminary test about variance homogeneity is superfluous. These authors recommended the Welch's approximate test for all sample size and parameter configurations. Another criticism about the conventional approach is the appropriateness of the usual variance ratio test; this test is heavily dependent on the normality assumption, and nonnormality and unequal variances cannot be separated with this test. An alternative approach regarding selection of the test for the means is to use the t test if the ratio s_1/s_2, where s_1 is the larger of the sample standard deviations, is less than a specified number (say, 1.3), or else use the Welch test. In this conditional approach, no preliminary variance test is used. Even then, Zimmerman's (2004) simulation studies indicated that the choice of a test based on the examination of sample statistics makes things worse, and Welch's test is preferable to other tests. Indeed, many undergraduate-level texts in statistics (e.g., Moore, 2004, p. 454) recommend the Welch test for all parameter configurations.

We now consider the multivariate case. As far as we know, no study has been carried out regarding the choice of a test for equality of two normal mean vectors. A reason for this is that, until recently, no clear-cut winner among the approximate tests for the multivariate Behrens–Fisher problem was identified. For the past 5 decades, several authors proposed exact and approximate solutions to this problem.

Bennett (1951) derived an exact method similar to the one due to Scheffé (1943) for the univariate case. Again, for the same reasons given for the univariate case, this approach should not be recommended for practical use. Other authors proposed approximate solutions using the idea of Welch's (1947) solution for the univariate case. Some of the approximate solutions for the multivariate case simplify to Welch's solution for the univariate case, however, they are different when the dimension is two or more.

To identify the best test for the multivariate Behrens–Fisher problem, Krishnamoorthy and Yu (2004) selected three approximate affine invariant tests (James, 1954; Yao, 1965; and Johansen, 1980) based on past studies and also proposed a new invariant test. The new test is obtained by modifying Nel and van der Merwe's (1986) test so that the resulting test is invariant. The modified Nel and van der Merwe test is referred to as the MNV test. Krishnamoorthy and Yu showed via a simulation study that the MNV test is the best among available invariant tests. Furthermore, the MNV test is one of the invariant tests that simplifies to Welch's test for the univariate case. Therefore, we consider only the MNV test for the multivariate Behrens–Fisher problem in this article.

This article is organized as follows. In the next section, we briefly discuss the Hotelling T^2 test, a test for equality of covariance matrices, the conventional approach for testing the equality of two normal mean vectors when the covariance matrices are unknown and arbitrary, and the MNV test. We next study size and power properties of the following approaches: (a) application of the Hotelling T^2 test regardless of the relation between the covariance matrices, (b) the conventional approach, and (c) application of the MNV test regardless of the relation between the covariance matrices. The conventional approach is to use the Hotelling T^2 test if a test for equality of the covariance matrices produced an insignificance result, else use the MNV test. The sizes and powers of the approaches are estimated using Monte Carlo simulation. In general, the results based on our simulation studies are in agreement with those for the univariate case with the exception that the T^2 test sometimes performs poorly even when the sample sizes are equal. We also observed that a preliminary test for covariance matrices is not only superfluous but also makes things worse. The MNV test not only controls Type I error rates very satisfactorily but also performs as well as the T^2 test when the covariance matrices are equal. Our overall recommendation is that the MNV test can be used safely for all practical applications.

THE TESTS

Let $\boldsymbol{X}_{i1},\ldots\boldsymbol{X}_{iN_i}$ be a sample of vector observations from a p-variate normal population with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, $i = 1, 2$. Let $\bar{\boldsymbol{X}}_i$

and $\boldsymbol{S}_i$ denote, respectively, the mean vector and variance–covariance matrix based on $\boldsymbol{X}_{i1},\ldots\boldsymbol{X}_{iN_i}$. That is,

$$\bar{\boldsymbol{X}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{X}_{ij} \text{ and } \boldsymbol{S}_i = \frac{1}{n_i}\sum_{j=1}^{N_i}(\boldsymbol{X}_{ij}-\bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij}-\bar{\boldsymbol{X}}_i)', \tag{1}$$

where $n_i = N_i - 1$. Let us consider the problem of testing

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs. } H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{2}$$

Because $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ is equivalent to $H_0 : \mathbf{A}\boldsymbol{\mu}_1 + \mathbf{b} = \mathbf{A}\boldsymbol{\mu}_2 + \mathbf{b}$ for any nonsingular matrix $\mathbf{A}$ and $p \times 1$ vector $\mathbf{b}$, a practical solution should be nonsingular invariant; otherwise, for a given data set, the p value for testing $H_{01} : \mathbf{A}_1\boldsymbol{\mu}_1 = \mathbf{A}_1\boldsymbol{\mu}_2$ may be different from the one for testing $H_{02} : \mathbf{A}_2\boldsymbol{\mu}_1 = \mathbf{A}_2\boldsymbol{\mu}_2$ when $\mathbf{A}_1$ and $\mathbf{A}_2$ are different nonsingular matrices. As a consequence, the conclusions could be different even though H_{01} and H_{02} are equivalent. Furthermore, as shown in the following section, a nonsingular transformation reduces the parameter space considerably so that the merit of an invariant solution can be evaluated over a wide range of the parameter space.

Hotelling T^2 Test When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$

If the covariance matrices are assumed to be equal, then the usual Hotelling T^2 statistic is given by

$$T^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)'\left[\left(\frac{1}{N_1}+\frac{1}{N_2}\right)\boldsymbol{S}_p\right]^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2), \tag{3}$$

where $\boldsymbol{S}_p = \dfrac{n_1\boldsymbol{S}_1 + n_2\boldsymbol{S}_2}{n_1+n_2}$ is the pooled covariance matrix so that $E(\boldsymbol{S}_p) = \boldsymbol{\Sigma}$, where E denotes the expectation operator. Under H_0,

$$T^2 \sim \frac{(n_1+n_2)p}{n_1+n_2-p+1}F_{p,n_1+n_2-p+1}. \tag{4}$$

For an observed value T_0^2 of T^2, the null hypothesis in (2) will be rejected if

$$T_0^2 > \frac{(n_1+n_2)p}{n_1+n_2-p+1}F_{p,n_1+n_2-p+1,1-\alpha}, \tag{5}$$

where $F_{m,n,c}$ denotes the *100c*th percentile of an F distribution with numerator df m and denominator *df n*.

For the aforementioned test to be valid, covariance matrices must be equal. A test for equality of covariance matrices is outlined next.

A Test for Equality of Two Covariance Matrices

A modified likelihood ratio test (LRT) statistic (e.g., see Muirhead, 1982, p. 309) is given by

$$\Lambda = \frac{\left|\boldsymbol{S}_1\right|^{\frac{n_1}{2}} \left|\boldsymbol{S}_2\right|^{\frac{n_2}{2}}}{\left|\boldsymbol{S}_p\right|^{\frac{\eta}{2}}}, \tag{6}$$

where $n = n_1 + n_2$ and $\boldsymbol{S}_p$ is the pooled covariance matrix as defined in T^2 statistic (3). To express its asymptotic distribution, let

$$f = \frac{p(p+1)}{2}, \rho = 1 - \frac{(2p^2+3p-1)}{6(p+1)n}\left(\frac{n}{n_1} + \frac{n}{n_2} - 1\right),$$

and

$$\gamma = \frac{p(p+1)}{48}\left\{(p-1)(p+2)\left(\frac{n^2}{n_1^2} + \frac{n^2}{n_2^2} - 1\right) - 6n^2(1-\rho)^2\right\}.$$

Let $W = -2\rho\ln(\Lambda)$, and W_0 be an observed value of W. Then, the p value of the modified LRT based on an asymptotic distribution is given by

$$1.0 - P\left(\chi_f^2 \le W_0\right) - \frac{\gamma}{\rho^2 n^2}\left[P\left(\chi_{f+4}^2 \le W_0\right) - P\left(\chi_f^2 \le W_0\right)\right]. \tag{7}$$

For a given level of significance α, the null hypothesis of equality of covariance matrices will be rejected if the p value is less than α.

The MNV Test

The modified Nel and van der Merwe (1986) test is based on the quadratic form $(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)'\tilde{\Omega}^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$, where $\tilde{\Omega}$ is an estimate of the $\text{Cov}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) = \frac{\boldsymbol{\Sigma}_1}{N_1} + \frac{\boldsymbol{\Sigma}_2}{N_2}$.

Using the unbiased estimator $\tilde{S}_i = \dfrac{S_i}{N_i}$ for $\dfrac{\Sigma_i}{N_i}$, we get the test statistic

$$T_u^2 = (\bar{X}_1 - \bar{X}_2)' \tilde{S}^{-1} (\bar{X}_1 - \bar{X}_2), \quad with \quad \tilde{S} = \tilde{S}_1 + \tilde{S}_2. \tag{8}$$

It is shown in Krishnamoorthy and Yu (2004) that

$$T_u^2 \sim \frac{\nu p}{\nu - p + 1} F_{p,\nu-p+1} \text{ approximately,}$$

where

$$\nu = \frac{p + p^2}{\dfrac{1}{n_1}\{\text{tr}[(\tilde{S}_1\tilde{S}^{-1})^2] + [\text{tr}(\tilde{S}_1\tilde{S}^{-1})]^2\} + \dfrac{1}{n_2}\{\text{tr}[(\tilde{S}_2\tilde{S}^{-1})^2] + [\text{tr}(\tilde{S}_2\tilde{S}^{-1})]^2\}}. \tag{9}$$

The *df* ν is different from the one given in Nel and van der Merwe (1986). As already mentioned, Krishnamoorthy and Yu (2004) modified the *df* so that the distribution of T_u^2 is affine invariant. For a given observed value T_{u0}^2 of T_u^2, the MNV test rejects the null hypothesis of equal mean vectors when

$$P\left(\frac{\nu p}{\nu - p + 1} F_{p,\nu-p+1} > T_{u0}^2\right) < \alpha. \tag{10}$$

The MNV test is affine invariant, and it simplifies to Welch's (1947) approximate *df* test for the univariate case.

The Conventional Approach

The usual practical approach for testing the difference between two mean vectors is to use the Hotelling T^2 test if the hypothesis of equality of covariance matrices is tenable, otherwise use one of the approximate tests given for Behrens–Fisher problem. Let us use the MNV test for the Behrens–Fisher problem. The approach can be described as follows: For given data sets, (i) test if $\Sigma_1 = \Sigma_2$ using the approach given in the earlier section. If the p value in (7) is greater than or equal to α_1 (α_1 is the nominal level for testing $\Sigma_1 = \Sigma_2$), then the T^2 test is used to test the equality of the mean vectors at nominal level α; otherwise, the MNV test is used to test the equality of the mean vectors at level α. Typically, α_1 and α are chosen to be equal. Using the indicator function $I[.]$, we see that the combination of the tests is based on the test statistic

$$C = T^2 I[\text{p} - value \geq \alpha] + T_u^2 I[\text{p} - value < \alpha], \tag{11}$$

where the p value is given in (7), T^2 and T_u^2 are given in (3) and (8) respectively. Let C_0 be an observed value of C in (10). Then the combined test rejects the H_0 in (2) whenever $P(C > C_0 \mid H_0) < \alpha$.

MONTE CARLO STUDIES

We first note that there exists a nonsingular matrix M such that $M\boldsymbol{\Sigma}_1 M' = I_p$ and $M\boldsymbol{\Sigma}_2 M' = \Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$, where λ_i's are the eigenvalues of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$. Using this fact and the fact that the tests are affine invariant, we can take without loss of generality that $\boldsymbol{\Sigma}_1 = I_p$, $\boldsymbol{\Sigma}_2 = \Lambda$, and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$ to compute the sizes. We estimated the sizes and powers using Monte Carlo simulation consisting of 100,000 runs. For example, to compute the sizes of the MNV test, at each parameter/sample size configuration, we generated

$$\tilde{\boldsymbol{X}}_1 \sim N_p(\boldsymbol{0}, I_p), \tilde{\boldsymbol{X}}_2 \sim N_p(\boldsymbol{0}, \Lambda), \boldsymbol{S}_1 \sim W_p(n_1, I_p/n_1), \ \boldsymbol{S}_2 \sim W_p(n_2, \Lambda/n_2),$$

and computed the T_u^2 statistic in (8), the *df* v in (9), and the p value in (10) with T_{u0}^2 replaced by T_u^2. The proportion of (out of these 100,000 runs) p values that are less than the nominal level α is a Monte Carlo estimate of the size. Because we are estimating probabilities (or proportions) at various parameter and sample size configurations, the maximum error of an estimate is $2\sqrt{.5 \times .5/100000} = 0.0032$. We used the IMSL subroutine RNMVN for generating multivariate normal random vectors and the Applied Statistics Algorithm (AS 53) due to Smith and Hocking (1972) to generate Wishart random matrices. Monte Carlo estimates of sizes and powers of the Hotelling T^2 test and the conventional approach were also computed similarly.

RESULTS

Size Properties

The estimated sizes of the three tests are plotted in Figure 1 for the case of $p = 2$ and nominal level $\alpha = 0.05$, and they are presented in Table 1 for the case of $p = 5$. For the conventional approach, we used $\alpha = 0.05$ for testing $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

1. From Figures 1a(i) and 1b(ii), we see that the sizes of the Hotelling T^2 test were very close to the nominal level when the sample sizes were equal. Even for

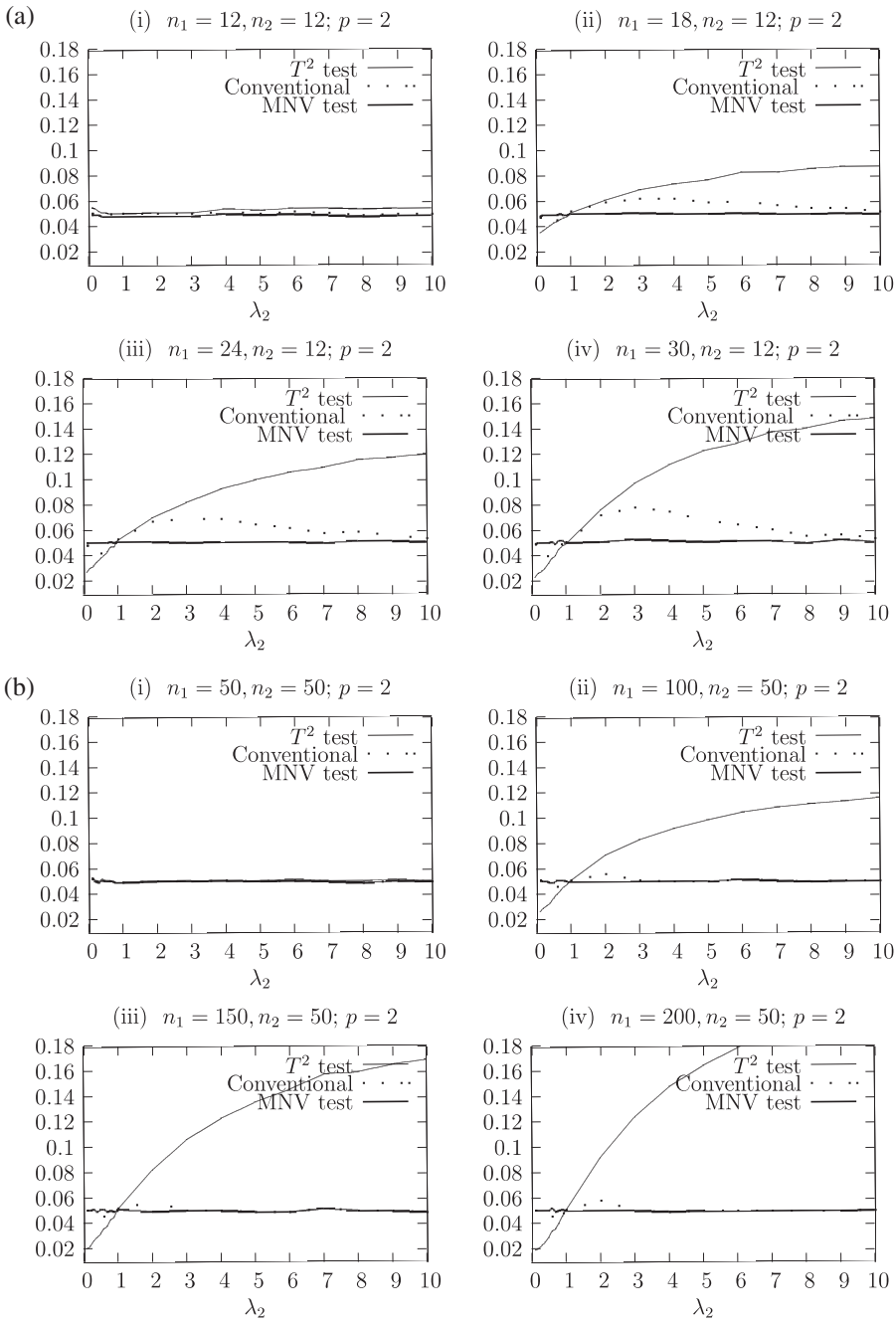

FIGURE 1 (a) The sizes of the tests as a function of λ_2; $\lambda_1 = 1$; sizes of the MNV test lie on $y = 0.05$ line; small samples. (b) The sizes of the tests as a function of λ_2; $\lambda_1 = 1$; sizes of the MNV test lie on $y = 0.05$ line; large samples.

TABLE 1
Sizes of the Tests when $p = 5$

	$n_1 = 10, n_2 = 10$			$n_1 = 30, n_2 = 20$		
$(\lambda_1, \ldots, \lambda_5)$	(1)	(2)	(3)	(1)	(2)	(3)
(1,1,1,1,1)	.050	.047	.049	.051	.051	.050
(1,.1,.2,.3,.4)	.070	.051	.049	.026	.049	.049
(1,.1,.1,.1,.1)	.093	.054	.052	.022	.050	.050
(1,.5,.8,.9,.9)	.051	.048	.048	.043	.045	.050
(1,.1,.1,.9,.9)	.062	.049	.048	.032	.049	.049
(1,.3,.6,.1,.8)	.060	.050	.049	.032	.049	.050
(1,2,3,4,5)	.063	.051	.048	.104	.052	.051
(1,1,1,10,10)	.061	.048	.049	.126	.051	.051
(1,10,10,10,10)	.094	.054	.054	.163	.051	.051
(1,2,1,2,1)	.051	.048	.045	.063	.060	.050
(1,3,2,3,5)	.061	.051	.049	.099	.054	.051
(1,2,5,7,10)	.073	.051	.050	.125	.052	.052
	$n_1 = 15, n_2 = 30$			$n_1 = 20, n_2 = 40$		
$(\lambda_1, \ldots, \lambda_5)$	(1)	(2)	(3)	(1)	(2)	(3)
(1,1,1,1,1)	.049	.051	.051	.050	.051	.050
(1,.1,.2,.3,.4)	.180	.053	.053	.177	.052	.052
(1,.1,.1,.1,.1)	.265	.054	.054	.257	.052	.052
(1,.5,.8,.9,.9)	.068	.067	.052	.067	.065	.050
(1,.1,.1,.9,.9)	.145	.054	.053	.146	.053	.053
(1,.3,.6,.1,.8)	.134	.057	.054	.132	.052	.051
(1,2,3,4,5)	.019	.047	.049	.017	.048	.049
(1,1,1,10,10)	.026	.049	.049	.016	.050	.050
(1,10,10,10,10)	.011	.049	.049	.009	.051	.051
(1,2,1,2,1)	.037	.041	.049	.034	.039	.049
(1,3,2,3,5)	.019	.045	.049	.018	.049	.050
(1,2,5,7,10)	.015	.048	.048	.014	.049	.049

Note. (1) = T^2 test; (2) = the conventional test; (3) = the modified Nel and van der Merwe test.

small equal sample size, the sizes barely exceeded 0.05 when λ_2 was large, or equivalently $\boldsymbol{\Sigma}_2$ is much "larger" than $\boldsymbol{\Sigma}_1$. However, when $p = 5$, and $n_1 = n_2 = 10$, the sizes exceed the nominal level considerably. The Type I error rates of the T^2 test is elevated when the larger sample size is associated with smaller covariance matrix; otherwise, the T^2 test is too liberal.

2. The conventional approach also has elevated sizes when the sample sizes are unequal. See Figures 1a(ii), (iii) and (iv). This conventional test appears to have similar behavior even for large unequal sample sizes; see Figures 2a(iii) and (iv). If the covariance matrices are not very different, the preliminary covariance test often leads to the wrong choice of the test for the means, and as a result it has inflated or depressed Type I error rates (see Table 1, $n_1 = 15$, $n_2 = 30$ and $n_1 = 20$, $n_2 = 40$).

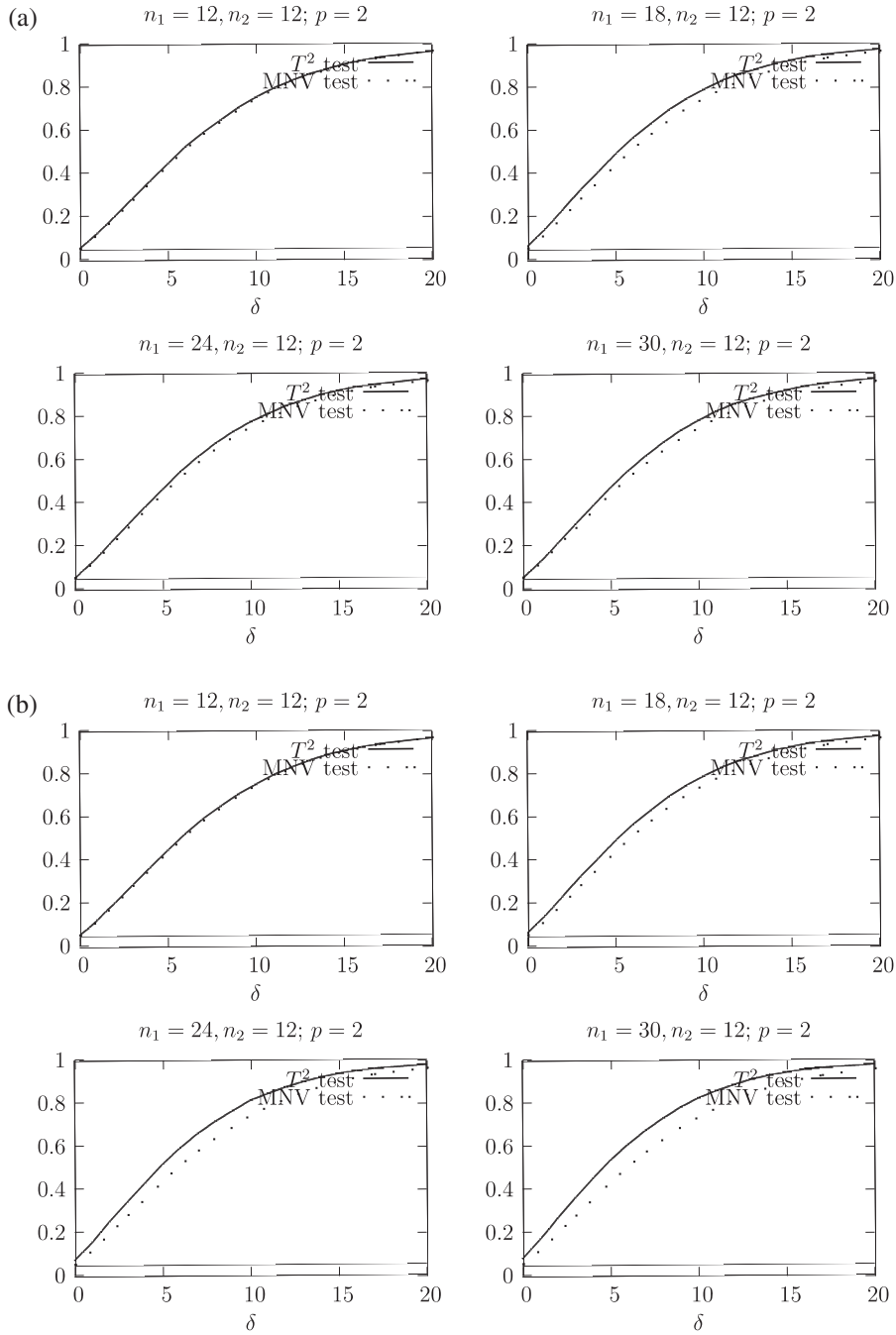

FIGURE 2 (a) Powers of the tests as a function of δ; $\Sigma_1 = \Sigma_2$ and $\alpha = 0.05$. (b) Powers of the tests as a function of δ; $\Sigma_1 \neq \Sigma_2$, $(\lambda_1, \lambda_2) = (1,2)$ and $\alpha = 0.05$.

3. The sizes of the MNV test are close to the nominal level for all the cases considered. This is true regardless of the relations between the covariance matrices and the relations between the sample sizes.

Thus, it is clear from these observations that the preliminary test of covariance equality is not useful to choose a better test, and so it will be dropped from power comparison studies. Furthermore, we observed that the MNV is the best with respect to size properties.

Power Properties

The powers were estimated for $p = 2$ and 5, and noncentrality parameter $\delta = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\left(\frac{\boldsymbol{\Sigma}_1}{N_1} + \frac{\boldsymbol{\Sigma}_2}{N_2}\right)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. The mean vectors were chosen such that $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (\sqrt{\delta/p})\mathbf{1}$, where $\mathbf{1}$ denotes the vector of ones. For $p = 2$, the estimates of the powers were plotted in Figure 2a when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ and in Figure 2b when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. For $p = 5$, similar power plots are given in Figures 3a and 3b. We observe the following from the estimated powers.

1. When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the powers of the T^2 test and the MNV test were almost identical except for the case ($n_1 = 30$, $n_2 = 12$) where the powers of the T^2 test barely exceeded those of the MNV test (see Figure 2a).

2. The power plots given in Figure 2b are for the case of unequal covariance matrices. Notice that when $n_1 = n_2 = 12$, both tests controlled the sizes and have similar power property. On the other hand, when $n_1 \neq n_2$, because of its inflated size, the T^2 test appears to be more powerful than the MNV test. See Figure 2b, $n_1 = 24$, $n_2 = 12$, $n_1 = 30$, $n_2 = 12$ and $\delta = 0$.

3. For the case of $p = 5$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the powers of the T^2 test were little more than those of the MNV test when n_1 was much larger than n_2 (see Figure 3a, $n_1 = 40$, $n_2 = 15$); otherwise, both tests seem to have similar power properties.

4. We observe again from the power plots in Figure 3b that if both the T^2 and the MNV tests have similar size properties, then they have similar power properties. The T^2 test appears to be more powerful than the MNV test in cases where it has inflated Type I error rates.

AN ILLUSTRATIVE EXAMPLE

We now illustrate the two-sample Hotelling T^2 test and the MNV test using a practical example. This example concerns the effects of delay in oral practice at the beginning of second-language learning. The data are taken from Timm (1975, Exercise 3.12). An experimental group of 28 participants was given oral practice with a

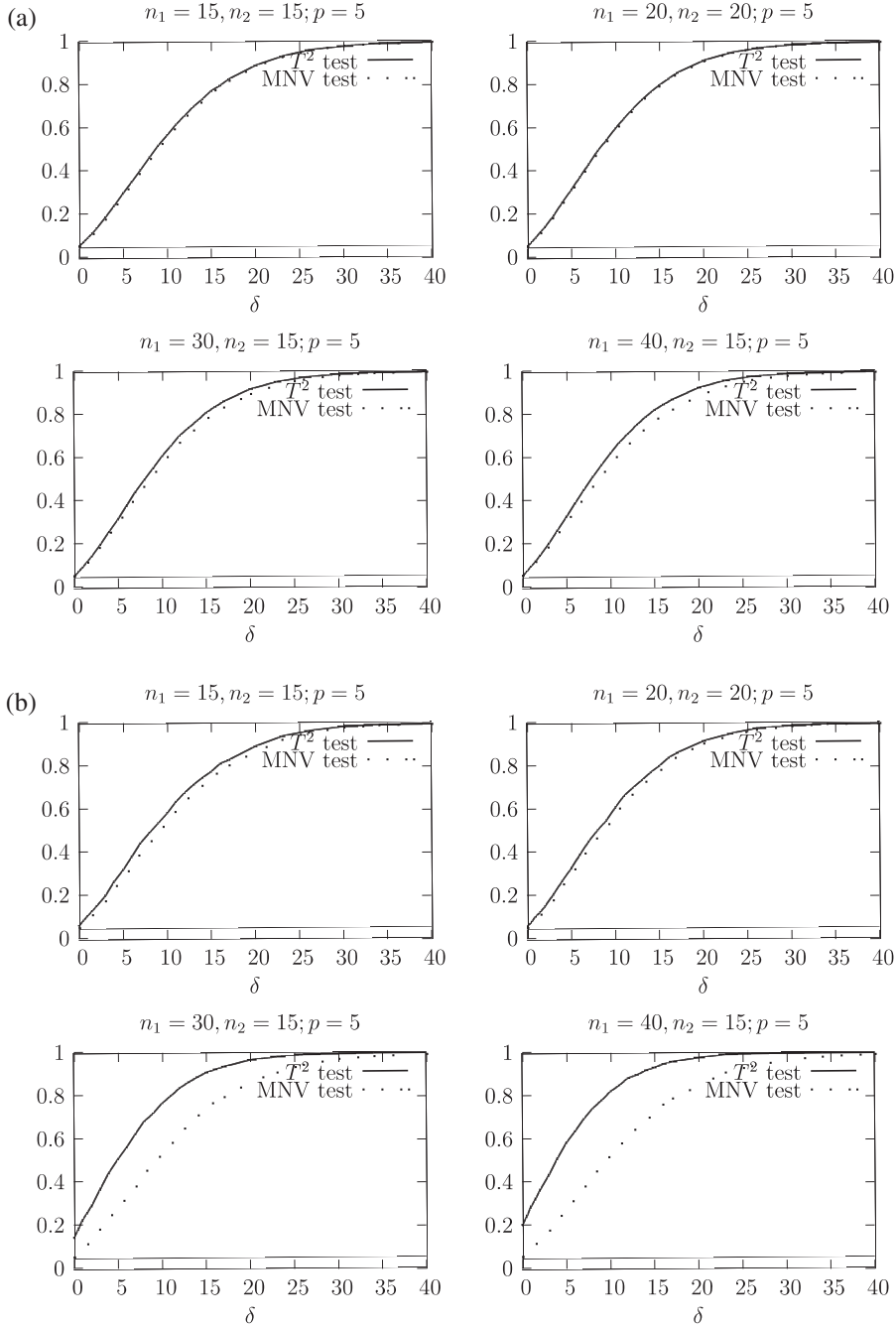

FIGURE 3 (a) Powers of the tests as a function of δ; $\Sigma_1 = \Sigma_2$. (b) Powers of the tests as a function of δ; $\Sigma_1 \neq \Sigma_2$; (λ_1, ..., λ_5) = (1, 2, 3, 4, 5)

4-week delay and a control group of 28 participants with no delay. A comprehensive examination was given to both groups at the end of the first 6 weeks, and the scores (on listening, speaking, reading, writing) were recorded. The summary statistics for the experimental group (subscript 1) and the control group (subscript 2) follow.

$$\bar{\boldsymbol{X}}_1 = \begin{pmatrix} 29.143 \\ 48.643 \\ 35.571 \\ 86.500 \end{pmatrix}, \bar{\boldsymbol{X}}_2 = \begin{pmatrix} 28.964 \\ 45.179 \\ 34.679 \\ 81.964 \end{pmatrix} \text{ and } \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 = \begin{pmatrix} 0.179 \\ 3.464 \\ 0.893 \\ 4.536 \end{pmatrix}.$$

The sample covariance matrices defined in (1) are:

$$\boldsymbol{S}_1 = \begin{pmatrix} 22.942 & 30.942 & 4.434 & 21.815 \\ 30.942 & 78.608 & 14.582 & 56.704 \\ 4.434 & 14.582 & 17.513 & 30.519 \\ 21.815 & 56.704 & 30.519 & 91.074 \end{pmatrix} \text{ and } \boldsymbol{S}_2 = \begin{pmatrix} 24.036 & 18.747 & 15.062 & 31.517 \\ 18.747 & 42.374 & 11.726 & 38.451 \\ 15.062 & 11.726 & 20.522 & 31.951 \\ 31.517 & 38.451 & 31.951 & 132.258 \end{pmatrix}.$$

Consider testing

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0} \text{ vs. } H_a : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0}.$$

Hotelling T^2 Test

The pooled sample covariance matrix is given by

$$\boldsymbol{S}_p = \begin{pmatrix} 23.489 & 24.845 & 9.748 & 26.666 \\ 24.845 & 60.491 & 13.154 & 47.577 \\ 9.748 & 13.154 & 19.018 & 31.235 \\ 26.666 & 47.577 & 31.235 & 111.666 \end{pmatrix}$$

and the Hotelling T^2 statistic defined in (3) is computed as

$$T^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \left[\left(\frac{1}{N_1} + \frac{1}{N_2} \right) \boldsymbol{S}_p \right]^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) = 5.646.$$

Taking $\alpha = .05$ and noting that $n_1 = N_1\text{-}1 = -27$ and $n_2 = N_2 - 1 = 27$, we have that

$$\frac{(n_1+n_2)p}{n_1+n_2-p+1}F_{p,n_1+n_2-p+1,0.95}=\frac{216}{51}F_{4,51,0.95}=10.814.$$

Because $T^2 = 5.646 < 10.814$, we do not reject H_0; also, the p value is .270. Thus, the data do not provide sufficient evidence to indicate the mean vectors are significantly different.

The MNV Test

For this example, the values of T^2 and T_u^2 must be the same because the sample sizes are equal. Nevertheless, for the sake of illustration, we compute T_u^2. The matrix $\tilde{S}$ defined in (8) is computed as

$$\tilde{S}=\begin{pmatrix}1.678 & 1.775 & 0.696 & 1.905\\ 1.775 & 4.321 & 0.940 & 3.398\\ 0.696 & 0.940 & 1.358 & 2.231\\ 1.905 & 3.398 & 2.231 & 7.976\end{pmatrix}$$

and the MNV test statistic defined in (8) is computed as

$$T_u{}^2=(\bar{X}_1-\bar{X}_2)'\tilde{S}^{-1}(\bar{X}_1-\bar{X}_2)=5.646.$$

The approximate $df\ v = 52.113$, and the critical value

$$\frac{vp}{v-p+1}F_{p,v-p+1,0.95}=10.868.$$

The $df\ v$ was calculated using (9) and $\mathrm{tr}(\tilde{S}_1\tilde{S}^{-1})=1.875$, $\mathrm{tr}(\tilde{S}_1\tilde{S}^{-1})^2=1.040$, $\mathrm{tr}(\tilde{S}_2\tilde{S}^{-1})=2.125$, and $\mathrm{tr}(\tilde{S}_2\tilde{S}^{-1})^2=1.291$. Because $T_u{}^2=5.646<10.868$, we do not reject H_0 in (11). Also, the p value of the MNV test is 0.272.

Thus both tests yielded the same conclusions and produced similar results (see the critical values and p values).

REMARKS

Our results based on simulation studies are similar to those given in the univariate case with an exception that, for the case of unequal covariance matrices, the T^2 test can have inflated Type I error rates even when $n_1 = n_2$. This finding suggests that

the T^2 test is appropriate only when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. However, in practice, it is difficult to identify populations with the same covariance matrices. Also, we have shown that testing equality of covariance matrices, and then selecting a test for the mean vectors may lead to erroneous conclusions. Moore (2004, p. 455) mentioned that the pooled *t*-test should not be used for the univariate normal case. Our work indicates that the same conclusion holds for the multivariate normal case. Our overall recommendation, based on simulation studies in this article and by Krishnamoorthy and Yu (2004), is that one can use the MNV test for situations where the normality assumption is tenable.

Finally, we point out that the tests based on the T_u^2 statistic in (8) are in general not robust to normality assumption violation. Among others, Wilcox (1995); Lix and Keselman (2004); and Lix, Keselman, and Hinds (2005) proposed some robust test procedures based on trimmed means. The MNV test, as it is, will be sensitive to normality assumption. A modified version along the lines of Wilcox (1995) or by some other approaches may yield a robust test. We plan to modify the MNV test using trimmed means, and evaluate the robustness of the resulting test.

ACKNOWLEDGMENT

We are grateful to three reviewers and the editor for providing valuable comments and suggestions.

REFERENCES

Bennett, B. M. (1951). Note on a solution of the generalized Behrens–Fisher problem. *Annals of the Institute of Statistical Mathematics, 2,* 87–90.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika, 41,* 19–43.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85–95.

Krishnamoorthy, K., & Yu, J. (2004). Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem. *Statistics & Probability Letters, 66,* 161–169.

Lix, L. M., & Keselman, H. J. (2004). Multivariate tests of means in independent group designs: Effects of covariance heterogeneity and nonnormality. *Evaluation & the Health Professions, 27,* 45–69.

Lix, L. M., Keselman, H. J., & Hinds, A. M. (2005). Robust tests for the multivariate Behrens–Fisher problem. *Computer Methods and Programs in Biomedicine, 77,* 129–139.

Moore, D. (2004). *The basic practice of statistics.* New York: Freeman.

Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample *t* test versus the Satterthwaite's approximate *F* test. *Communications in Statistics—Theory and Methods, 18,* 3963–3975.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory.* New York: Wiley.

Nel, D. G., & van der Merwe, C. A. (1986). A solution to the Multivariate Behrens–Fisher problem. *Communications in Statistics—Theory and Methods, 15,* 3719–3735.

Scheffé, H. (1943). On the solutions of the Behrens-Fisher problem based on the *t* distribution. *The Annals of the Mathematical Statistics, 14,* 35–44.

Scheffé, H. (1970). Practical solutions of the Behrens–Fisher problem. *Journal of the American Statistical Association, 65,* 1501–1508.

Smith, W. B., & Hocking, R. R. (1972). Wishart Variates Generator (algorithm AS 53). *Applied Statistics, 21,* 341–345.

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology.* Monterey, CA: Brooks/Cole.

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

Wilcox, R. R. (1995). Simulation results on solutions to the multivariate Behrens–Fisher problem via trimmed means. *The Statistician, 44,* 213–225.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens–Fisher problem. *Biometrika, 52,* 139–147.

Zimmerman, D. W. (2004). Conditional probabilities of rejecting H_0 by pooled and separate-variances *t* tests given heterogeneity of sample variances. *Communications in Statistics—Simulation and Computation, 33,* 69–81.