A modified Bartlett test for heteroscedastic one-way MANOVA

Jin-Ting Zhang · Xuefeng Liu

Received: 18 May 2011 / Published online: 28 December 2011
© Springer-Verlag 2011

Abstract In this paper, we investigate tests of linear hypotheses in heteroscedastic one-way MANOVA via proposing a modified Bartlett (MB) test. The MB test is easy to conduct via using the usual $\chi^2$-table. It is shown to be invariant under affine transformations, different choices of the contrast matrix used to define the same hypothesis and different labeling schemes of the mean vectors. Simulation studies and real data applications demonstrate that the MB test performs well and is generally comparable to Krishnamoorthy and Lu’s (J Statist Comput Simul 80(8):873–887, 2010) parametric bootstrap test in terms of size controlling and power.

Keywords Heteroscedastic one-way MANOVA · Modified Bartlett correction · Multivariate Behrens–Fisher problem

1 Introduction

The problem of comparing the mean vectors of $k$ multivariate normal populations based on $k$ independent samples is referred to as multivariate analysis of variance (MANOVA). If the $k$ covariance matrices are assumed to be equal, Wilks’ likelihood ratio, Lawley-Hotelling’s trace, Bartlett–Nanda–Pillai’s Trace and Roy’s largest root tests (Anderson 2003, Ch. 8, Sect. 6) can be used. When $k = 2$, Hotelling’s $T^2$ test is the uniformly most powerful affine invariant test. These tests, however, may become seriously biased when the assumption of equality of covariance matrices is violated. In real data analysis, such an assumption is often violated and is hard to check.
The problem for testing the difference between two normal mean vectors without assuming equality of covariance matrices is referred to as the multivariate Behrens–Fisher (BF) problem. This problem has been well addressed in the literature. Well-known and accurate solutions include James (1954), Yao (1965), Johansen (1980), Nel and Merwe (1986), Kim (1992), Krishnamoorthy and Yu (2004), Yanagihara and Yuan (2005), and Belloni and Didier (2008) among others. When $k > 2$ and the covariance matrices are unknown and arbitrary, the problem of testing equality of the mean vectors is often referred to as the multivariate $k$-sample BF problem or heteroscedastic one-way MANOVA. This multivariate $k$-sample BF problem is more complex and is not well addressed compared with the multivariate two-sample BF problem. Existing approximate solutions include James (1954), Johansen (1980) and Gamage et al. (2004) among others. Tang and Algina (1993) compared James’s first and second-order tests, Johansen’s test, and Bartlett–Nanda–Pillai’s trace test and concluded that none of them is satisfactory for all sample sizes and parameter configurations. Overall, they recommended James (1954) second-order test and Johansen (1980) test. Krishnamoorthy and Lu (2010) claimed, based on a preliminary study, that James’s second-order test is computationally very involved, and is difficult to apply when $k = 4$ or more, and offered little improvement over Johansen’s test. They then proposed a parametric bootstrap (PB) test to the multivariate $k$-sample BF problem. They compared their PB test against the Johansen test and the generalized F-test of Gamage et al. (2004) via some intensive simulations for various sample sizes and parameter configurations and found that their PB test performs best while the Johansen test and the generalized F-test are very liberal when the number of groups compared, $k$, is large.

Since the PB test is computationally intensive, it is still worthwhile to develop some new testing procedure which is comparable to the PB test in terms of size controlling and power but with much less computational work.

In this paper, we propose such a testing procedure, namely, a modified Bartlett (MB) test to the general linear hypothesis testing (GLHT) problem in heteroscedastic one-way MANOVA. The MB test is constructed based on an application of the modified Bartlett correction of Fujikoshi (2000) to a Wald-type test statistic for the GLHT problem. It is shown that under the null hypothesis the Wald-type test statistic has an asymptotic $\chi^2$-distribution with some known degrees of freedom. This asymptotic null distribution is hardly useful for the GLHT problem when the sample sizes are small and moderate since the convergence rate of the null distribution is of order $O(n_{\min}^{-1/2})$, which is very slow where $n_{\min}$ denotes the smallest sample size among the $k$ samples. Application of the modified Bartlett correction of Fujikoshi (2000) may improve the convergence rate of the null distribution to order $O(n_{\min}^{-1})$. We show that this is true at least for the first two moments of the null distribution so that the MB test can be applied for the GLHT problem even when the sample sizes are small and moderate. The MB test admits several nice properties. First of all, it can be simply conducted since the formulas for computing the MB test statistic is very simple and the associated null distribution is the well known $\chi^2$-distribution with some known degrees of freedom. Secondly, all related tests under heteroscedastic one-way MANOVA, such as the overall, post hoc, and contrast tests, can be conducted by the MB test in a common framework. We show that the MB test is invariant under affine transformations,
Heteroscedastic one-way MANOVA 137

different choices of the contrast matrix used to define the same hypothesis and different labeling schemes of the mean vectors. Finally, the MB test works well. Simulation results reported in Sect. 3 indicate that the MB test generally works well and it is generally comparable to Krishnamoorthy and Lu’s (2010) parametric bootstrap test in terms of size controlling and power.

The MB test can be regarded as an extension of the MB test of Yanagihara and Yuan (2005) from for the multivariate two-sample BF problem to for the GLHT problem in heteroscedastic one-way MANOVA. Yanagihara and Yuan (2005) showed via some simulations that their MB test is comparable to the MNV test (Krishnamoorthy and Yu 2004) which is known to be one of the best tests for the multivariate two-sample BF problem. In view of this, it is not a surprise that the MB test for the GLHT problem has good performance in terms of size controlling and power.

The rest of the paper is organized as follows. In Sect. 2, the MB test is developed in details and its nice properties are investigated. Simulation studies are presented in Sect. 3. An application to a real data set is given in Sect. 4. Technical proofs of the main results are outlined in the “Appendix”.

2 Main results

2.1 The MB test

Given \( k \) independent normal samples \( x_{l1}, x_{l2}, \ldots, x_{lnl} \sim N_p(\mu_l, \Sigma_l) \), \( l = 1, 2, \ldots, k \), where and throughout, \( N_p(\mu, \Sigma) \) denotes a \( p \)-dimensional normal distribution with mean vector \( \mu \) and covariance matrix \( \Sigma \), we want to test whether the \( k \) mean vectors are equal:

\[
H_0: \mu_1 = \mu_2 = \cdots = \mu_k, \quad \text{versus} \quad H_1: H_0 \text{ is not true}, \tag{2.1}
\]

without assuming the equality of \( \Sigma_l, l = 1, 2, \ldots, k \). The above problem is usually referred to as the multivariate \( k \)-sample BF problem or the overall heteroscedastic one-way MANOVA test, which is a special case of the following GLHT problem in heteroscedastic one-way MANOVA:

\[
H_0: C\mu = c, \quad \text{versus} \quad H_1: C\mu \neq c, \tag{2.2}
\]

where \( \mu = (\mu_1^T, \mu_2^T, \ldots, \mu_k^T)^T \) is the mean vector obtained via stacking all the population mean vectors of the \( k \) samples into a single column vector, \( C : q \times (kp) \) is a known coefficient matrix with \( \text{rank}(C) = q \), and \( c : q \times 1 \) is a known constant vector.

In fact, the GLHT problem (2.2) reduces to the overall MANOVA test (2.1) when we set \( c = 0 \) and \( C = Q \otimes I_p \) with \( Q = (I_{k-1}, -1_{k-1}) \) where \( I_r \) and \( 1_r \) denote the identity matrix of size \( r \) and the \( r \)-dimensional vector of ones respectively, and \( \otimes \) denotes the Kronecker product operation.

The GLHT problem (2.2) is very general. It includes not only the overall MANOVA test (2.1) but also various post hoc and contrast tests as special cases since any post hoc and contrast tests can be written in the form of (2.2). For example, when the
overall MANOVA test is rejected, it is of interest to further test if \( \mu_1 = 3\mu_2 \) or if a contrast is zero, e.g., \( \mu_1 - 3\mu_2 + 2\mu_3 = 0 \). In fact, these two testing problems can be written in the form of (2.2) with \( c = 0 \) and \( C = (e_{1,k} - 3e_{2,k})^T \otimes I_p \) and \( C = (e_{1,k} - 3e_{2,k} + 2e_{3,k})^T \otimes I_p \) respectively where and throughout \( e_{r,k} \) denotes a unit vector of length \( k \) with \( r \)th entry being 1 and others 0.

**Remark 1** From the definition of \( C \), it is seen that we have \( C = C_0 \otimes I_p \) where \( C_0 \) is a full rank matrix of size \( q_0 \times k \) so that we always have \( q = q_0p \).

To construct a proper test statistic for the GLHT problem (2.2), let \( \hat{\mu}_l = \bar{x}_l = n_l^{-1} \sum_{j=1}^{n_l} x_{lj} \) be the sample mean vector of the \( l \)th sample. Set \( \hat{\mu} = (\hat{\mu}_1^T, \hat{\mu}_2^T, \ldots, \hat{\mu}_k^T)^T \) which is an unbiased estimator of \( \mu \). Then \( \hat{\mu} \sim N_{kp}(\mu, \Sigma) \) where \( \Sigma = \text{diag} \left( \frac{\Sigma_1}{n_1}, \frac{\Sigma_2}{n_2}, \ldots, \frac{\Sigma_k}{n_k} \right) \). It follows that \( C\hat{\mu} - c \sim N_q(C\mu - c, C\Sigma C^T) \). This suggests that a Wald-type test statistic can be constructed as

\[
T = (C\hat{\mu} - c)^T \left( C\hat{\Sigma} C^T \right)^{-1} (C\hat{\mu} - c),
\]

(2.3)

where \( \hat{\Sigma} = \text{diag} \left( \hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_k \right) \) with \( \hat{\Sigma}_l = (n_l - 1)^{-1} \sum_{j=1}^{n_l} (x_{lj} - \hat{\mu}_l)(x_{lj} - \hat{\mu}_l)^T \) being the usual unbiased sample covariance matrix of the \( l \)th sample. Notice that the distribution of \( T \) is very complicated and its closed-form distribution is generally not tractable in the context of heteroscedastic one-way MANOVA.

**Remark 2** When the covariance matrix homogeneity is valid and the sample covariance matrices \( \hat{\Sigma}_l \) are replaced by their pooled sample covariance matrix \( \hat{\Sigma}_l/(N-k) \) where \( N = \sum_{l=1}^{k} n_l \) denotes the total sample size of the \( k \) samples, it is easy to show that \( T/(N-k) \) has the distribution of the well-known Lawley–Hotelling trace test statistic (Anderson 2003, Ch. 8, Sect. 6) with \( q_0 \) and \( N-k \) degrees of freedom where \( q_0 \) is defined in Remark 1.

To construct the MB test based on \( T \), following Yanagihara and Yuan (2005), we set

\[
z = (C\Sigma C^T)^{-1/2}(C\hat{\mu} - c), \quad W = H\hat{\Sigma}H^T, \quad H = (C\Sigma C^T)^{-1/2}C,
\]

(2.4)

so that we can equivalently re-express \( T \) as

\[
T = z^T W^{-1} z.
\]

(2.5)

Notice that the above re-expression theoretically helps the development of the MB test but in practice we still use (2.3) to compute the value of \( T \). We have \( z \sim N_q(\mu_z, I_q) \), where \( \mu_z = (C\Sigma C^T)^{-1/2}(C\mu - c) \). Let \( n_{\min} = \min_{l=1}^{k} n_l \) and \( n_{\max} = \max_{l=1}^{k} n_l \). To study the asymptotic distribution of \( T \), we impose the following condition:

\[
\frac{n_l}{n_{\min}} \to r_l < \infty, \quad l = 1, 2, \ldots, k, \quad \text{as } n_{\min} \to \infty.
\]

(2.6)
Remark 3 Condition (2.6) requires that all the $k$ sample sizes tend to infinity proportionally, preventing the cases where $n_{\text{max}}/n_{\text{min}}$ is too large. This guarantees that the limit of $n_{\text{min}}\Sigma$ is a full rank matrix as $n_{\text{min}} \to \infty$ so that the limit of $n_{\text{min}}C\Sigma C^T$ is invertible as $n_{\text{min}} \to \infty$.

Remark 3 indicates that when the ratio $n_{\text{max}}/n_{\text{min}}$ is too large, $(C\Sigma C^T)^{-1}$ may be instable so that the performance of the MB test proposed in this article will be affected as shown by the simulation results presented in Tables 4 and 5 in Sect. 3.2. This shows that in practice, the experimental researcher should make the one-way MANOVA design as balanced as possible. For further discussion, let $\chi^2_m$ denote a chi-squared distribution with $m$ degrees of freedom.

**Theorem 1** Under the condition (2.6) and $H_0$, as $n_{\text{min}} \to \infty$, we have that $T$ converges to $\chi^2_q$ in distribution.

Theorem 1 states that $T$ asymptotically follows the $\chi^2$-distribution with $q$ degrees of freedom. For the overall MANOVA test (2.1), this result may not be new but it is new for the GLHT problem (2.2). Based on this, one may test the GLHT problem using the usual $\chi^2$-test provided that $n_{\text{min}}$ is sufficiently large.

**Remark 4** Remark 2 implies that the convergence rate of $T$ to $\chi^2_q$ for homogeneous one-way MANOVA is of order $N^{-1/2}$ provided that $k$ is fixed. However, from the proof of Theorem 1, it is seen that the convergence rate of $T$ to $\chi^2_q$ is of order $n_{\text{min}}^{-1/2}$. Therefore, the convergence rate of $T$ to $\chi^2_q$ for heteroscedastic one-way MANOVA is much slower than that for homogeneous one-way MANOVA. This is intuitively understood since for homogeneous one-way MANOVA, the $N$ observations are pooled to estimate the common covariance matrix (see Remark 2) while for heteroscedastic one-way MANOVA, the $l$th population covariance matrix has to be estimated separately using only the observations in the $l$th sample where $l = 1, \ldots, k$.

Therefore, the $\chi^2$-test is hardly useful for the GLHT problem (2.2) with small and moderate samples. To overcome this problem, following Yanagihara and Yuan (2005), we apply the modified Bartlett correction proposed by Fujikoshi (2000) to $T$ to improve its convergence rate and hence propose the so-called MB test. Set $\Omega_l = n_l^{-1}H_l\Sigma_lH_l^T$, $l = 1, 2, \ldots, k$ where $H_l = (C\Sigma C^T)^{-1/2}C_l$, $l = 1, 2, \ldots, k$ with $C_1, C_2, \ldots, C_k$ being the $k$ matrices of size $q \times p$ such that $C = [C_1, C_2, \ldots, C_k]$. To propose the MB test, we need the following result.

**Theorem 2** Under the condition (2.6) and $H_0$, as $n_{\text{min}} \to \infty$, we have

$$E(T) = q \left(1 + \frac{\alpha_1}{n_{\text{min}}}\right) + O(n_{\text{min}}^{-2}) \quad \text{and} \quad E(T^2) = q(q + 2) \left(1 + \frac{\alpha_2}{n_{\text{min}}}\right) + O(n_{\text{min}}^{-2}),$$

(2.7)

where $\alpha_1 = n_{\text{min}}(\Delta_1 + \Delta_2)/q$, $\alpha_2 = n_{\text{min}}[(2q + 8)\Delta_1 + (2q + 6)\Delta_2]/[q(q + 2)]$, $\Delta_1 = \sum_{l=1}^k \text{tr}(\Omega_l^2)/(n_l - 1)$, and $\Delta_2 = \sum_{l=1}^k \text{tr}^2(\Omega_l)/(n_l - 1)$. Furthermore, we have
\[
\frac{q^2}{(n_{\text{max}} - 1)kp} \leq \Delta_1 \leq \frac{q}{n_{\text{min}} - 1} 
\text{and} \quad \frac{q^2}{(n_{\text{max}} - 1)k} \leq \Delta_2 \leq \frac{pq}{(n_{\text{min}} - 1)}. \tag{2.8}
\]

**Remark 5** Under the conditions of Theorem 2, it is easy to show that \( \alpha_1 \) and \( \alpha_2 \) will tend to their finite limits as \( n_{\text{min}} \to \infty \). This implies that the first two moments of \( T \) tend to the first two moments of \( \chi_q^2 \) with the rate \( O(n_{\text{min}}^{-1}) \) which is rather slow.

**Remark 6** The inequalities in (2.8) indicate that as \( n_{\text{min}} \to \infty \), both \( \Delta_1 \) and \( \Delta_2 \) will tend to 0. However, this is not always the case when \( n_{\text{max}} \to \infty \). Alternatively speaking, larger values of \( n_{\text{max}} \) alone may not push \( \Delta_1 \) and \( \Delta_2 \) closer to 0 but they do push the lower bounds of \( \Delta_1 \) and \( \Delta_2 \) closer to 0.

**Remark 7** Theorem 2 is derived with \( k \) fixed. That is, it is derived without taking the effect of \( k \) into account. Therefore, it is not guaranteed that Theorem 2 is still valid when \( k \to \infty \) although the simulation results presented in Tables 4 and 5 in Sect. 3.2 show that provided \( n_{\text{min}} \) is not too small, both the PB and MB tests work reasonably well when \( k \) is large. Further study about this is interesting and warranted.

Based on Theorems 1 and 2, we can now apply the modified Bartlett correction of Fujikoshi (2000) to \( T \) through the log-transformation \( T_{\text{mb}} = (n_{\text{min}}\beta_1 + \beta_2)\log(1 + \frac{T}{n_{\text{min}}\beta_1}) \), where \( \beta_1 = \frac{2}{\alpha_2 - 2\alpha_1} \) and \( \beta_2 = \frac{(q+2)\alpha_2 - 2(q+4)\alpha_1}{2(\alpha_2 - 2\alpha_1)} \).

**Remark 8** By Fujikoshi (2000), it is easy to show that \( E(T_{\text{mb}}) = q + O(n_{\text{min}}^{-2}) \) and \( E(T_{\text{mb}}^2) = q(q + 2) + O(n_{\text{min}}^{-2}) \). By Remark 5, we only have \( E(T) = q + O(n_{\text{min}}^{-1}) \) and \( E(T^2) = q(q + 2) + O(n_{\text{min}}^{-1}) \). Thus it is expected that \( T_{\text{mb}} \) converges to \( \chi_q^2 \) much faster than \( T \) does. However, the MB test is still affected by the value of \( n_{\text{min}} \). That is, when \( n_{\text{min}} \) is too small, the MB test may not perform well as showed by the simulation results presented in Tables 4 and 5 in Sect. 3.2.

In real data analysis, \( \beta_1 \) and \( \beta_2 \) have to be estimated from the data. Proper estimators can be obtained via replacing \( \hat{\Omega}_l, \ l = 1, 2, \ldots, k \) by their estimators:

\[
\hat{\Omega}_l = n_l^{-1}(C\hat{\Sigma}CT)^{-1/2}C_l\hat{\Sigma}_lC_l^T(C\hat{\Sigma}CT)^{-1/2}, \quad l = 1, 2, \ldots, k. \tag{2.9}
\]

Thus,

\[
\hat{\Delta}_1 = \sum_{l=1}^{k} \frac{\text{tr}(\hat{\Omega}_l^2)}{(n_l - 1)} \quad \text{and} \quad \hat{\Delta}_2 = \sum_{l=1}^{k} \frac{\text{tr}^2(\hat{\Omega}_l)}{(n_l - 1)}. \tag{2.10}
\]

The estimators \( \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1 \) and \( \hat{\beta}_2 \) are then obtained accordingly so that

\[
\hat{T}_{\text{mb}} = (n_{\text{min}}\hat{\beta}_1 + \hat{\beta}_2)\log\left(1 + \frac{T}{n_{\text{min}}\hat{\beta}_1}\right) \sim \chi_q^2 \text{ approximately.} \tag{2.11}
\]

Some simple algebra leads to \( n_{\text{min}}\hat{\beta}_1 = \frac{q(q+2)}{2\hat{\Delta}_1 + \hat{\Delta}_2} \) and \( n_{\text{min}}\hat{\beta}_1 + \hat{\beta}_2 = \frac{(q+2)(2q-\hat{\Delta}_2)}{4\hat{\Delta}_1 + 2\hat{\Delta}_2} \).

From the proof of Theorem 2, we can see that the ranges of \( \Delta_1 \) and \( \Delta_2 \) as given
in (2.8) are also the ranges of \( \hat{\Delta}_1 \) and \( \hat{\Delta}_2 \) respectively. Thus, provided \( n_{\text{min}} \geq p + 1 \), we always have \( n_{\text{min}}\hat{\beta}_1 > 0 \) and \( n_{\text{min}}\hat{\beta}_1 + \hat{\beta}_2 > 0 \). This guarantees that \( T_{\text{sn}} \) is a monotonically increasing nonnegative function of \( T \). The critical value of the MB test can then be specified as \( \chi^2_q(1 - \alpha) \) for any given significance level \( \alpha \). We reject the null hypothesis in (2.2) when this critical value is exceeded by \( \hat{T}_{\text{sn}} \). The MB test can also be conducted via computing the \( P \)-value based on \( \chi^2_q \). Thus, the MB test can be conducted easily via using the usual \( \chi^2 \)-table.

2.2 Desirable properties of the MB test

Notice that the matrix \( Q \) used to write the overall MANOVA test (2.1) into the GLHT problem (2.2) is a contrast matrix with all row totals being 0 and it is not unique. For example, \( \tilde{Q} = (-1_{k-1}, I_{k-1}) \) is also a valid contrast matrix for (2.1). It is known from Kshirsagar (1972, Ch. 5, Sect. 4) that for any two contrast matrices \( \tilde{Q} \) and \( Q \) used to define the same hypothesis, there is a nonsingular matrix \( P \) such that \( \tilde{Q} = PQ \). By (2.2), the C-matrix associated with \( Q \) can be expressed as \( C = Q \otimes I_p \). Let \( \tilde{C} \) be the C-matrix associated with \( \tilde{Q} \). Then we have \( \tilde{C} = \tilde{Q} \otimes I_p = (PQ) \otimes I_p = (P \otimes I_p)C \).

Theorem 3 shows that the MB test is invariant under different choices of the contrast matrix used to define the same hypothesis.

**Theorem 3** The MB test is invariant when \( C \) and \( c \) in (2.2) are replaced with

\[
\tilde{C} = (P \otimes I_p)C \quad \text{and} \quad \tilde{c} = (P \otimes I_p)c,
\]

respectively where \( P \) is any nonsingular matrix.

In practice, the observed responses \( x_{lj}, j = 1, 2, \ldots, n_l; l = 1, 2, \ldots, k \) are often re-centered or rescaled before any inference is conducted. Recentering and rescaling are two special cases of the affine transformation of \( x_{lj} \), defined as

\[
\tilde{x}_{lj} = Bx_{lj} + b, \quad l = 1, 2, \ldots, n_l; \quad l = 1, 2, \ldots, k,
\]

where \( B \) is any nonsingular matrix and \( b \) is any constant vector.

**Theorem 4** The MB test is invariant under the affine transformation (2.13).

Finally, we have the following result.

**Theorem 5** The MB test is invariant under different labeling schemes of the mean vectors \( \mu_l, l = 1, 2, \ldots, k \).

3 Simulation studies

In this section, we shall present simulation studies for comparing the MB test against the PB test under two situations: (1) when \( k \) is small or moderate, and (2) when \( k \) is large.
3.1 Simulation studies with small and moderate \( k \)

In this subsection, intensive simulations are conducted to compare the MB test against the PB test of Krishnamoorthy and Lu (2010) when the factor under consideration has a small or moderate number of levels. Krishnamoorthy and Lu (2010) demonstrated that the PB test generally outperforms the Johansen (1980) test and the generalized F-test of Gamage et al. (2004) in terms of size controlling. The Johansen test and the generalized F-test are generally very liberal and the generalized F-test is very time consuming. Therefore, we shall not include them for comparison against the MB test.

Following Krishnamoorthy and Lu (2010), for simplicity, we set \( \Sigma_1 = I_p, \Sigma_2 = \text{diag}(\lambda_1, \ldots, \lambda_p) \) and \( \Sigma_l, l = 3, 4, \ldots, k \) to be some positive definite matrices, where \( p, \lambda_1, \ldots, \lambda_k \) and other tuning parameters are specified later. Let \( n = (n_1, n_2, \ldots, n_k) \) denote the vector consisting of the \( k \) sample sizes. For given \( n \) and \( \Sigma_l, l = 1, 2, \ldots, k \), we first generated \( k \) sample mean vectors \( \hat{\mu}_l, l = 1, \ldots, k \) and \( k \) sample covariance matrices \( \hat{\Sigma}_l, l = 1, \ldots, k \) by \( \hat{\mu}_l \sim N_p(\mu_l, \Sigma_l/n_l), \hat{\Sigma}_l \sim W_p(n_l - 1, \Sigma_l/(n_l - 1)) \), \( l = 1, 2, \ldots, k \) where the population mean vectors \( \mu_l = \mu_1 + l\delta h, l = 2, \ldots, k \) with \( \mu_1 \) being the first population mean vector, \( h \) a constant unit vector specifying the direction of the population mean differences, and \( \delta \) a tuning parameter controlling the amount of the population mean differences. Without loss of generality, we specified \( \mu_1 \) as \( 0 \) and \( h \) as \( h_0/\|h_0\| \) where \( h_0 = (1, 2, \ldots, p)^T \) for any \( p \) and \( \|h_0\| \) denotes the usual \( L^2 \)-norm of \( h_0 \). We then applied the PB and MB tests to the generated sample mean vectors and the sample covariance matrices, and recorded their \( P \) values. The empirical sizes and powers of the PB and MB tests were computed based on 5,000 runs and the number of inner loops for the PB test is 1,000. In all the simulations conducted, the significance level was specified as \( 5\% \) for simplicity.

The empirical sizes (associated with \( \delta = 0 \)) and powers (associated with \( \delta > 0 \)) of the PB and MB tests for the multivariate \( k \)-sample BF problem (2.1), together with the associated tuning parameters, are presented in Tables 1, 2 and 3, in the columns labeled with “PB” and “MB” respectively. As seen from the three tables, three sets of the tuning parameters for population covariance matrices are examined, with the first set specifying the homogeneous cases and seven sets of sample sizes are specified, with the first three sets specifying the balanced sample size cases. To measure the overall performance of a test in terms of maintaining the nominal size \( \alpha \), we define the average relative error as \( \text{ARE} = M^{-1} \sum_{j=1}^{M} |\hat{\alpha}_j - \alpha|/\alpha \times 100 \) where \( \hat{\alpha}_j \) denotes the \( j \)th empirical size for \( j = 1, 2, \ldots, M, \alpha = .05 \) and \( M \) is the number of empirical sizes under consideration. The smaller \( \text{ARE} \) value indicates the better overall performance of the associated test. Usually, when \( \text{ARE} \leq 10 \), the test performs very well; when \( 10 < \text{ARE} \leq 20 \), the test performs reasonably well; and when \( \text{ARE} > 20 \), the test does not perform well since its empirical sizes are either too liberal or too conservative. Notice that for a good test, the larger the sample sizes, the smaller the \( \text{ARE} \) values. Notice that for simplicity, in the specification of the covariance and sample size tuning parameters, we often use \( \alpha_r \) to denote “\( \alpha \) repeats \( r \) times”, e.g., \( (20_3) = (20, 20, 20) \) and \( (2_3, 4, 5_2) = (2, 2, 2, 4, 5, 5) \). Tables 1, 2 and 3 show the empirical sizes and powers of the two tests for a bivariate case with \( k = 2 \), a 3-variate case with \( k = 3 \) and a 5-variate case with \( k = 5 \), respectively. For the cases with balanced sample sizes, the associated empirical sizes and powers are highlighted in boldface.
\( \lambda = 2, \quad \Sigma_1 = I_2, \quad \Sigma_2 = \text{diag}(\lambda) \)

| \( \lambda \) | \( n \) | \( \delta = 0 \) | \( \delta = 0.6 \) | \( \delta = 1.2 \) | \( \delta = 1.8 \) |
|----------------|-------|----------------|----------------|----------------|----------------|
| \( \lambda_1 \) | \( n_1 \) | .045 | .042 | .120 | .117 | .382 | .380 | .726 | .725 |
| \( n_2 \) | .043 | .041 | .256 | .256 | .792 | .790 | .989 | .989 |
| \( n_3 \) | .049 | .050 | .524 | .524 | .985 | .985 | 1.00 | 1.00 |
| \( n_4 \) | .048 | .047 | .149 | .146 | .470 | .464 | .821 | .817 |
| \( n_5 \) | .047 | .047 | .345 | .343 | .906 | .903 | .999 | .999 |
| \( n_6 \) | .047 | .045 | .151 | .146 | .463 | .458 | .819 | .815 |
| \( n_7 \) | .051 | .050 | .345 | .346 | .909 | .910 | .999 | .999 |
| \( \lambda_2 \) | \( n_1 \) | .048 | .048 | .074 | .072 | .196 | .192 | .402 | .394 |
| \( n_2 \) | .049 | .050 | .135 | .135 | .446 | .444 | .816 | .818 |
| \( n_3 \) | .049 | .048 | .255 | .256 | .789 | .789 | .991 | .991 |
| \( n_4 \) | .052 | .051 | .089 | .089 | .251 | .250 | .511 | .507 |
| \( n_5 \) | .050 | .049 | .202 | .199 | .664 | .667 | .958 | .960 |
| \( n_6 \) | .056 | .053 | .086 | .081 | .230 | .225 | .442 | .432 |
| \( n_7 \) | .051 | .050 | .163 | .164 | .524 | .525 | .870 | .868 |
| \( \lambda_3 \) | \( n_1 \) | .050 | .049 | .073 | .071 | .144 | .139 | .305 | .300 |
| \( n_2 \) | .050 | .050 | .117 | .114 | .348 | .346 | .667 | .668 |
| \( n_3 \) | .058 | .059 | .200 | .202 | .645 | .646 | .951 | .952 |
| \( n_4 \) | .055 | .056 | .081 | .084 | .192 | .188 | .394 | .388 |
| \( n_5 \) | .048 | .047 | .162 | .161 | .537 | .536 | .878 | .877 |
| \( n_6 \) | .059 | .057 | .083 | .082 | .174 | .166 | .333 | .328 |
| \( n_7 \) | .052 | .052 | .135 | .135 | .399 | .397 | .768 | .767 |

\[ \text{ARE} = 5.88 \quad 6.24 \]

\( \lambda_1 = (1, 2), \lambda_2 = (1, 5), \lambda_3 = (1, 10), n_1 = (7_2), n_2 = (10_2), n_3 = (15_2), n_4 = (7, 10), n_5 = (15, 30), n_6 = (10, 7) \text{ and } n_7 = (30, 15) \)

From the three tables, it is seen that both the tests perform reasonably well with their ARE values less than 20 and their empirical sizes and powers are comparable for almost all the cases under consideration. Since the PB test is computationally intensive, the MB test is generally preferred.

3.2 Simulation studies with large \( k \)

In this subsection, we compare the MB test against the PB test when the factor under consideration has many levels. That is, when \( k \) is very large, e.g., \( k \geq 30 \). Notice that when \( k \) is large, both the MB and PB tests are time-consuming if the dimension \( p \) is large. To overcome this difficulty, we confine ourselves to the univariate case, i.e., when \( p = 1 \). In this case, both the PB and MB tests reduce to their respective counterparts in heteroscedastic one-way ANOVA discussed by Krishnamoorthy et al. (2007),
Table 2  Empirical sizes and powers of the PB and MB tests for trivariate one-way MANOVA

\[
\begin{array}{cccccc}
\lambda, \rho & n & \delta = 0 & \delta = 0.6 & \delta = 1.2 & \delta = 1.8 \\
\hline
\lambda_1, \rho_1 & n_1 & 0.47 & 0.46 & 0.074 & 0.074 & 0.199 & 0.195 & 0.439 & 0.438 \\
\lambda_2, \rho_2 & n_2 & 0.46 & 0.047 & 0.156 & 0.155 & 0.583 & 0.581 & 0.937 & 0.938 \\
\lambda_3, \rho_3 & n_3 & 0.55 & 0.55 & 0.349 & 0.348 & 0.943 & 0.945 & 0.999 & 0.999 \\
& n_4 & 0.51 & 0.51 & 0.116 & 0.118 & 0.371 & 0.369 & 0.709 & 0.711 \\
& n_5 & 0.55 & 0.057 & 0.184 & 0.189 & 0.620 & 0.628 & 0.950 & 0.953 \\
& n_6 & 0.55 & 0.057 & 0.129 & 0.130 & 0.427 & 0.429 & 0.820 & 0.821 \\
& n_7 & 0.057 & 0.058 & 0.217 & 0.220 & 0.759 & 0.767 & 0.990 & 0.992 \\
& n_1 & 0.45 & 0.47 & 0.078 & 0.078 & 0.215 & 0.220 & 0.429 & 0.434 \\
& n_2 & 0.48 & 0.49 & 0.159 & 0.158 & 0.575 & 0.578 & 0.939 & 0.939 \\
& n_3 & 0.46 & 0.45 & 0.336 & 0.337 & 0.933 & 0.936 & 1.00 & 1.00 \\
& n_4 & 0.62 & 0.66 & 0.119 & 0.122 & 0.364 & 0.375 & 0.689 & 0.703 \\
& n_5 & 0.51 & 0.53 & 0.188 & 0.196 & 0.611 & 0.623 & 0.947 & 0.950 \\
& n_6 & 0.57 & 0.059 & 0.126 & 0.133 & 0.444 & 0.458 & 0.863 & 0.872 \\
& n_7 & 0.059 & 0.061 & 0.249 & 0.259 & 0.847 & 0.855 & 0.998 & 0.998 \\
& n_1 & 0.47 & 0.48 & 0.078 & 0.077 & 0.200 & 0.203 & 0.436 & 0.441 \\
& n_2 & 0.56 & 0.54 & 0.163 & 0.165 & 0.590 & 0.593 & 0.937 & 0.937 \\
& n_3 & 0.48 & 0.49 & 0.340 & 0.341 & 0.937 & 0.937 & 1.00 & 1.00 \\
& n_4 & 0.60 & 0.62 & 0.118 & 0.123 & 0.356 & 0.366 & 0.680 & 0.694 \\
& n_5 & 0.64 & 0.66 & 0.176 & 0.186 & 0.594 & 0.609 & 0.941 & 0.947 \\
& n_6 & 0.51 & 0.052 & 0.132 & 0.137 & 0.444 & 0.460 & 0.852 & 0.864 \\
& n_7 & 0.051 & 0.053 & 0.241 & 0.252 & 0.839 & 0.851 & 0.997 & 0.997 \\
\hline
\text{ARE} & 10.26 & 11.60
\end{array}
\]

and Zhang and Liu (2011), respectively. Simulation 1 here aims to evaluate the performance of the PB and MB tests when \( k = 30 \), the smallest sample size increases from very small to moderately large, and the remaining sample sizes are about the same with the total sample size \( N \) unchanged. Simulation 2 here aims to evaluate the performance of the PB and MB tests when \( k = 40 \), the largest sample size is much larger than the remaining sample sizes which are the same and increase from very small to moderately large.

The data were generated in a similar way as in the previous subsection but now \( p = 1 \). For a given sample size vector \( n = (n_1, n_2, \ldots, n_k) \), a population mean vector \( \mu = (\mu_1, \mu_2, \ldots, \mu_k)^T \), and a population standard deviation vector \( \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_k)^T \), we first generated \( k \) sample means \( \hat{\mu}_l, l = 1, \ldots, k \) and \( k \) sample variances \( \hat{\sigma}_l^2, l = 1, \ldots, k \) by \( \hat{\mu}_l \sim N(\mu_l, \sigma_l^2/n_l) \), and \( \hat{\sigma}_l^2 \sim \frac{\sigma_l^2}{n_l-1} \chi^2_{n_l-1}, l = 1, 2, \ldots, k \). Springer
Table 3 Empirical sizes and powers of the PB and MB tests for 5-variate one-way MANOVA

| $k = 5$, $\Sigma_1 = I_5$, $\Sigma_2 = \text{diag}(\lambda)$, $\Sigma_3 = \text{diag}(\eta)$, $\Sigma_4 = \text{diag}(u)$, $\Sigma_5 = \text{diag}(v)$ |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| $(\lambda, \eta, u, v)$ | $n$ | $\delta = 0$ | $\delta = 0.5$ | $\delta = 1.0$ | $\delta = 1.5$ |
| $(\lambda(1), \eta(1), u(1), v(1))$ | $n(1)$ | .047 | .128 | .486 | .912 |
| | $n(2)$ | .049 | .265 | .909 | 1.00 |
| | $n(3)$ | .045 | .377 | .984 | 1.00 |
| | $n(4)$ | .049 | .159 | .666 | .982 |
| | $n(5)$ | .054 | .237 | .873 | 1.00 |
| | $n(6)$ | .054 | .175 | .726 | .993 |
| | $n(7)$ | .051 | .269 | .913 | 1.00 |
| $(\lambda(2), \eta(2), u(2), v(2))$ | $n(1)$ | .050 | .067 | .133 | .275 |
| | $n(2)$ | .049 | .095 | .297 | .658 |
| | $n(3)$ | .045 | .113 | .420 | .862 |
| | $n(4)$ | .060 | .074 | .179 | .400 |
| | $n(5)$ | .049 | .084 | .252 | .584 |
| | $n(6)$ | .043 | .082 | .183 | .446 |
| | $n(7)$ | .050 | .093 | .294 | .680 |
| $(\lambda(3), \eta(3), u(3), v(3))$ | $n(1)$ | .044 | .052 | .102 | .184 |
| | $n(2)$ | .054 | .077 | .190 | .461 |
| | $n(3)$ | .048 | .090 | .280 | .659 |
| | $n(4)$ | .051 | .064 | .125 | .284 |
| | $n(5)$ | .052 | .076 | .172 | .416 |
| | $n(6)$ | .052 | .065 | .123 | .252 |
| | $n(7)$ | .048 | .078 | .179 | .404 |

ARE 6.76 7.87

$\lambda(1) = (15), \eta(1) = (15), u(1) = (15), v(1) = (15), \lambda(2) = (122, 1, 24, 1), \eta(2) = (1, 0.1, 2, 24, 21), u(2) = (1, 3, 9, 10), v(2) = (5, 15, 5, 50), \lambda(3) = (1, 3, 9, 2), \eta(3) = (5, 15, 45, 3), u(3) = (1, 3, 9, 30), v(3) = (5, 15, 45, 100), n(1) = (15), n(2) = (255), n(3) = (50s), n(4) = (20, 25, 35, 40, 50), n(5) = (30, 35, 40, 50, 70), n(6) = (50, 40, 35, 25, 20) and n(7) = (70, 50, 40, 35, 30).

The population means $\mu_l, l = 1, 2, \ldots, k$ were specified as $\mu_l = \mu_0 + \delta u_l, l = 1, 2, \ldots, k$ where the tuning parameter $\delta$ was used to control the differences of the population means. Without loss of generality, we set $\mu_0 = 2.1, u_1 = 0.5, u_2 = -0.30, u_k = 1.1$ and $u_l = 0, l = 3, 4, \ldots, k - 1$. We then applied the PB and MB tests to the generated sample means and sample variances and recorded their $P$ values. For the PB test, 1,000 inner replicates were conducted. This process was repeated 10,000 times.

Table 4 presents the results of Simulation 1. We have $k = 30, n_{\text{max}} = 25, n_{\text{min}} = 3, 6, \ldots, 18$, and we keep the total sample size $N$ the same across various cases, which is $30 \times 24 = 720$. First of all, it is seen that the same total sample size $N$ does not imply that the PB test (and the MB test as well) has the same performance in various cases. Secondly, it is seen that the PB and MB tests are generally comparable except
Table 4  Empirical sizes and powers of the PB and MB tests for one-way ANOVA with \( k = 30 \)

| \( \sigma \) | \( \delta = 0 \) | \( \delta = 0.33 \) | \( \delta = 0.67 \) | \( \delta = 1 \) |
|-------------|-------------|-------------|-------------|-------------|
|             | PB | MB | PB | MB | PB | MB | PB | MB | PB | MB |
| \( n = (3, 25_{22}, 24_{17}) \) | (1, 1, 1, 127) | .060 | .088 | .086 | .138 | .301 | .433 | .728 | .844 |
| \( n_{\text{max}}/n_{\text{min}} = 8.33 \) | (2, 2, 1, 127) | .050 | .086 | .084 | .147 | .253 | .398 | .647 | .798 |
| \( N = 720 \) | (4, 3, 2, 127) | .052 | .089 | .076 | .144 | .229 | .378 | .634 | .793 |
| ARE         | (6, 6, 5, 127) | .050 | .087 | .070 | .134 | .229 | .383 | .621 | .782 |
|             | 6.65 | 74.8 |
| \( n = (6, 25_{19}, 24_{10}) \) | (1, 1, 1, 127) | .052 | .058 | .110 | .120 | .412 | .434 | .849 | .862 |
| \( n_{\text{max}}/n_{\text{min}} = 4.17 \) | (2, 2, 1, 127) | .048 | .054 | .104 | .112 | .374 | .397 | .792 | .810 |
| \( N = 720 \) | (4, 3, 2, 127) | .048 | .055 | .105 | .115 | .361 | .385 | .784 | .802 |
| ARE         | (6, 6, 5, 127) | .048 | .054 | .100 | .111 | .357 | .380 | .770 | .787 |
|             | 4.25 | 10.5 |
| \( n = (9, 25_{16}, 24_{13}) \) | (1, 1, 1, 127) | .052 | .055 | .117 | .122 | .447 | .458 | .877 | .883 |
| \( n_{\text{max}}/n_{\text{min}} = 2.78 \) | (2, 2, 1, 127) | .050 | .052 | .108 | .114 | .383 | .392 | .812 | .820 |
| \( N = 720 \) | (4, 3, 2, 127) | .051 | .054 | .102 | .106 | .371 | .383 | .793 | .801 |
| ARE         | (6, 6, 5, 127) | .050 | .053 | .102 | .108 | .362 | .373 | .793 | .804 |
|             | 1.50 | 6.90 |
| \( n = (12, 25_{13}, 24_{16}) \) | (1, 1, 1, 127) | .049 | .051 | .121 | .126 | .453 | .462 | .880 | .885 |
| \( n_{\text{max}}/n_{\text{min}} = 2.08 \) | (2, 2, 1, 127) | .052 | .054 | .110 | .115 | .399 | .407 | .814 | .821 |
| \( N = 720 \) | (4, 3, 2, 127) | .050 | .052 | .110 | .113 | .386 | .394 | .798 | .804 |
| ARE         | (6, 6, 5, 127) | .047 | .047 | .102 | .105 | .372 | .379 | .793 | .799 |
|             | 3.10 | 5.00 |
| \( n = (15, 25_{10}, 24_{19}) \) | (1, 1, 1, 127) | .052 | .052 | .125 | .128 | .472 | .478 | .894 | .897 |
| \( n_{\text{max}}/n_{\text{min}} = 1.67 \) | (2, 2, 1, 127) | .052 | .054 | .110 | .110 | .394 | .400 | .819 | .824 |
| \( N = 720 \) | (4, 3, 2, 127) | .048 | .049 | .110 | .112 | .384 | .390 | .799 | .804 |
| ARE         | (6, 6, 5, 127) | .049 | .052 | .101 | .104 | .376 | .382 | .793 | .799 |
|             | 3.50 | 4.75 |
| \( n = (18, 25_{7}, 24_{22}) \) | (1, 1, 1, 127) | .051 | .052 | .127 | .130 | .479 | .484 | .901 | .904 |
| \( n_{\text{max}}/n_{\text{min}} = 1.39 \) | (2, 2, 1, 127) | .046 | .046 | .112 | .115 | .402 | .409 | .826 | .830 |
| \( N = 720 \) | (4, 3, 2, 127) | .049 | .051 | .110 | .113 | .391 | .396 | .803 | .808 |
| ARE         | (6, 6, 5, 127) | .051 | .052 | .109 | .112 | .383 | .388 | .795 | .801 |
|             | 3.60 | 4.75 |

for those cases with \( n_{\text{min}} = 3 \). For those cases with \( n_{\text{min}} = 3 \), the MB test is rather liberal. This bad performance of the MB test may be jointly resulted from the following two reasons: (1) \( n_{\text{min}} = 3 \) is too small; and (2) the ratio \( n_{\text{max}}/n_{\text{min}} = 8.33 \) is too large. This is because when \( n_{\text{min}} \) is too small, the performance of the MB test may not be good as indicated by Remark 8; and when the ratio \( n_{\text{max}}/n_{\text{min}} \) is too large, the condition (2.6) is not well satisfied so that the MB test may not perform well as explained in the paragraph following Remark 3. Lastly, it is also seen that the PB test is less sensitive to the values of \( n_{\text{min}} \) and the ratio \( n_{\text{max}}/n_{\text{min}} \) although in terms of empirical powers, the MB test performs slightly better than the PB test.
Table 5  Empirical sizes and powers of the PB and MB tests for one-way ANOVA with $k = 40$

| $\sigma$ | $\delta = 0$ | $\delta = 0.33$ | $\delta = 0.67$ | $\delta = 1$ |
|----------|--------------|-----------------|-----------------|-----------|
|          | PB           | MB              | PB              | MB        |
| $n = (100, 339)$ | (1, 1, 1, 137) | .027            | .118            | .029      |
| $n_{\max}/n_{\min} = 30.33$ | (2, 2, 1, 137) | .024            | .118            | .024      |
| $N = 217$ | (4, 3, 2, 137) | .024            | .109            | .028      |
|          | (6, 6, 5, 137) | .024            | .108            | .029      |
| ARE     | 50.4         | 126             |                 |           |
| $n = (100, 639)$ | (1, 1, 1, 137) | .047            | .077            | .061      |
| $n_{\max}/n_{\min} = 16.67$ | (2, 2, 1, 137) | .047            | .075            | .058      |
| $N = 334$ | (4, 3, 2, 137) | .053            | .080            | .053      |
|          | (6, 6, 5, 137) | .051            | .081            | .053      |
| ARE     | 5.20         | 56.4            |                 |           |
| $n = (100, 939)$ | (1, 1, 1, 137) | .051            | .064            | .078      |
| $n_{\max}/n_{\min} = 11.11$ | (2, 2, 1, 137) | .050            | .062            | .066      |
| $N = 451$ | (4, 3, 2, 137) | .050            | .063            | .061      |
|          | (6, 6, 5, 137) | .053            | .067            | .066      |
| ARE     | 2.25         | 27.9            |                 |           |
| $n = (100, 1239)$ | (1, 1, 1, 137) | .048            | .055            | .089      |
| $n_{\max}/n_{\min} = 8.33$ | (2, 2, 1, 137) | .051            | .058            | .074      |
| $N = 568$ | (4, 3, 2, 137) | .051            | .057            | .067      |
|          | (6, 6, 5, 137) | .053            | .059            | .069      |
| ARE     | 2.95         | 13.8            |                 |           |
| $n = (100, 1539)$ | (1, 1, 1, 137) | .050            | .054            | .108      |
| $n_{\max}/n_{\min} = 6.67$ | (2, 2, 1, 137) | .050            | .054            | .078      |
| $N = 685$ | (4, 3, 2, 137) | .051            | .056            | .075      |
|          | (6, 6, 5, 137) | .049            | .052            | .072      |
| ARE     | 1.25         | 7.75            |                 |           |
| $n = (100, 1839)$ | (1, 1, 1, 137) | .046            | .047            | .116      |
| $n_{\max}/n_{\min} = 5.56$ | (2, 2, 1, 137) | .049            | .051            | .092      |
| $N = 802$ | (4, 3, 2, 137) | .048            | .052            | .090      |
|          | (6, 6, 5, 137) | .048            | .051            | .085      |
| ARE     | 4.55         | 3.20            |                 |           |

Table 5 presents the results of Simulation 2. We now have $k = 40$, $n_{\max} = 100$, $n_{\min} = 3, 6, \ldots, 18$ but the total sample size $N$ is no longer the same across various cases. Firstly, it is seen that the PB test is also affected by the values of $n_{\min}$ and $n_{\max}/n_{\min}$ as indicated by those cases with $n_{\min} = 3$ and $n_{\max}/n_{\min} = 30.33$ where the PB test is rather conservative. Nevertheless, the PB test is indeed less sensitive to the values of $n_{\min}$ and $n_{\max}/n_{\min}$ than the MB test. This observation is similar to the associated one we observed from Table 4. Secondly, it is seen that the MB test is rather liberal when $n_{\min}$ is too small and when $n_{\max}/n_{\min}$ is too large as indicated by those cases with $n_{\min} \leq 6$. This is not a surprise as indicated by Remarks 3 and 8.
Some more interesting results can also be obtained via comparing Tables 4 and 5. For example, we found that in terms of ARE, the PB and MB tests performed better in those cases with \( n_{\text{min}} = 6, n_{\text{max}} / n_{\text{min}} = 16.67, N = 334 \), \( n_{\text{min}} = 9, n_{\text{max}} / n_{\text{min}} = 11.11, N = 451 \), and \( n_{\text{min}} = 12, n_{\text{max}} / n_{\text{min}} = 8.33, N = 568 \) in Table 5 than in those cases with \( n_{\text{min}} = 3, n_{\text{max}} / n_{\text{min}} = 8.33, N = 720 \) in Table 4, implying that \( n_{\text{min}} \) plays a rather important role in determining the performances of the PB and MB tests. We also found that in terms of ARE, the MB test performed better in those cases in Table 4 than in those cases in Table 5 with the same values of \( n_{\text{min}} \), implying that \( n_{\text{max}} / n_{\text{min}} \) also plays a rather important role in determining the performance of the MB test when the value of \( n_{\text{min}} \) is the same. This is in agreement with Remark 3.

4 Application to the Egyptian skull data

The Egyptian skull data were recently analyzed by Krishnamoorthy and Lu (2010). It can be downloaded at Statlib (http://lib.stat.cmu.edu/DASL/Stories/EgyptianSkullDevelopment.html). There are five samples of 30 skulls from the early pre-dynastic period (circa 4000 BC), the late pre-dynastic period (circa 3300 BC), the 12th and 13th dynasties (circa 1850 BC), the Ptolemaic period (circa 200 BC), and the Roman period (circa AD 150). Four measurements are available on each skull, namely, \( x_1 \) = maximum breadth, \( x_2 \) = borborygmatic height, \( x_3 \) = dentoalveolar length, and \( x_4 \) = nasal height (all in mm). To compare the MB test against the PB test of Krishnamoorthy and Lu (2010) in various cases, we applied these two tests to check the significance of the mean vector differences of the first \( k \) samples, using only the first \( n = (n_1, \ldots, n_k) \) observations for \( n = (10_k), (20_k) \) and \( (30_k) \), \( k = 2, 3, 4 \) and 5. There are totally 12 cases under consideration. The number of replications in the PB test is 10,000 and hence the time spent by the PB test is about 10,000 times of that spent by the MB test. The \( P \) values of the two tests for various cases are presented in Table 6.

It is seen from Table 6 that the \( P \) values of the PB and MB tests are generally close to each other. This is in agreement with the conclusions drawn from the simulations presented in the previous section. It is also seen that the first null hypothesis in Table 6 is not significant, with the \( P \) values of the two tests increasing from about 64\% to about 81\% with increasing the sample sizes; the other three null hypotheses are significant, with the \( P \) values of the two tests decreasing to less than 5\% with increasing the sample sizes. These results suggest that the Egyptian skulls had little

| Null hypothesis | \( n = (10_k) \) | \( n = (20_k) \) | \( n = (30_k) \) |
|-----------------|-----------------|-----------------|-----------------|
|                 | PB   | MB   | PB   | MB   | PB   | MB   |
| \( H_0 : \mu_1 = \mu_2 \) | .6367 | .6415 | .7202 | .7227 | .8158 | .8140 |
| \( H_0 : \mu_1 = \mu_2 = \mu_3 \) | .6030 | .6088 | .2125 | .2071 | .0291 | .0301 |
| \( H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \) | .1025 | .0997 | .0248 | .0227 | .0002 | .0002 |
| \( H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \) | .0547 | .0448 | .0033 | .0025 | .0000 | .0000 |
change in the early and late pre-dynastic periods but experienced a significant change over the later three periods.

Acknowledgments The work was partially supported by the National University of Singapore Academic Research Grant R-155-000-108-112. The authors thank the editor and two reviewers for their invaluable comments and suggestions which helped improve the article substantially.

Appendix: Technical proofs

Proof of Theorem 1 Under the given conditions, we have

$$\hat{\Sigma}_l \sim W_q(n_l - 1, \Sigma_l/(n_l - 1)), \quad l = 1, 2, \ldots, k,$$

(A.1)

where $W_q(m, V)$ denotes a $q$-dimensional Wishart distribution with $m$ degrees of freedom and covariance matrix $\Sigma$. It follows that $(\hat{\Sigma}_l - \Sigma_l)/n_l = O_p(n_l^{-3/2}), l = 1, 2, \ldots, k$. Thus $\hat{\Sigma} - \Sigma = O_p(n_{\min}^{-3/2})$. Noticing that $\Sigma = O(n_{\min}^{-1})$, we further have

$$R = H(\hat{\Sigma} - \Sigma)H^T = O_p(n_{\min}^{-1/2}),$$

(A.2)

where $H$ is defined in (2.4) and $H = O(n_{\min}^{1/2})$. This implies that

$$W = I_q + H(\hat{\Sigma} - \Sigma)H^T = I_q + R = I_q + O_p(n_{\min}^{-1/2}).$$

(A.3)

Theorem 1 follows from Slutsky’s theorem and the fact that under $H_0, z^T z \sim \chi_q^2$. □

Proof of Theorem 2 Notice that under $H_0$, we have $z \sim N_q(0, I_q)$. Applying the conditional expectation rule, some simple algebra leads to

$$E(T) = E \text{tr}(W^{-1}) \quad \text{and} \quad E(T^2) = 2E \text{tr}(W^{-2}) + E \text{tr}^2(W^{-1}).$$

(A.4)

From the proof of Theorem 1, we have that $W = I_q + R$ with $R = O_p(n_{\min}^{-1/2})$; see (A.2). Then we have

$$W^{-1} = (I_q + R)^{-1} = I_q - R + R^2 - R^3 + O_p(n_{\min}^{-2}),$$

$$W^{-2} = (I_q + R)^{-2} = I_q - 2R + 3R^2 - 4R^3 + O_p(n_{\min}^{-2}).$$

It is easy to see from (A.2) that $E(R) = 0$ and $Etr(R) = 0$. Thus

$$Etr(W^{-1}) = q + Etr(R^2) - Etr(R^3) + O(n_{\min}^{-2}),$$

$$Etr(W^{-2}) = q + 3Etr(R^2) - 4Etr(R^3) + O(n_{\min}^{-2}),$$

$$Etr^2(W^{-1}) = q^2 + Etr^2(R) + 2qEtr(R^2) - 2qEtr(R)Etr(R^2) + O(n_{\min}^{-2}).$$

(A.5)

To find $Etr(R^2)$ and $Etr(R^3)$ among others, we need some results from Letac and Massam (2004). They showed that if $Y \sim W_q(m, V)$, then

 Springer
Etr²[Y − E(Y)] = 2mtr(V²), Etr[Y − E(Y)]² = m[tr(V²) + tr²(V)],
Etr[Y − E(Y)]³ = mtr³(V) + 3mtr(V)tr(V²) + 4mtr(V³),
Etr[Y − E(Y)]tr[Y − E(Y)]² = 4mtr(V)tr(V²) + 4mtr(V³),  (A.6)

By (A.2), R = ∑ₖ₌₁ⁿ(W_l − Ω_l) = ∑ₖ₌₁ⁿ R_l where R_l = W_l − EW_l with W_l = n_l⁻¹H_lΣ_lW_lᵀ, l = 1, 2, . . . , k. Since W₁, . . . , W_k are independent and ER_l = 0, l = 1, 2, . . . , k, we have Etr(R²) = ∑ₖ₌₁ⁿ Etr(R_l²), Etr²(R) = ∑ₖ₌₁ⁿ Etr²(R_l), Etr³(R) = ∑ₖ₌₁ⁿ Etr³(R_l), and Etr(R)tr(R²) = ∑ₖ₌₁ⁿ Etr(R_l)tr(R_l²). By (A.1) and applying (A.6), we have

Etr(R²) = ∑ₖ₌₁ⁿ[tr(Ω_l²) + tr²(Ω_l)]/(n_l − 1) = Δ₁ + Δ₂,
Etr²(R) = 2∑ₖ₌₁ⁿ tr(Ω_l²)/(n_l − 1) = 2Δ₁,
Etr³(R) = ∑ₖ₌₁ⁿ[tr³(Ω_l) + 3tr(Ω_l)tr(Ω_l²) + 4tr²(Ω_l²)]/(n_l − 1)² = O(n⁻²ₘᵟᵢₙ),
Etr(R)tr(R²) = ∑ₖ₌₁ⁿ[4tr(Ω_l)tr(Ω_l²) + 4tr(Ω_l³)]/(n_l − 1)² = O(n⁻²ₘᵟᵢₙ),  (A.7)

where Δ₁ and Δ₂ are as defined in Theorem 2 and we have used the fact that 0 ≤ tr(Ω_l) < q since ∑ₖ₌₁ⁿ tr(Ω_l) = q. Combining (A.5) and (A.7) gives that

Etr(W⁻¹) = q + Δ₁ + Δ₂ + O(n⁻²ₘᵟᵢₙ),
Etr(W⁻²) = q + 3(Δ₁ + Δ₂) + O(n⁻²ₘᵟᵢₙ),
Etr²(W⁻¹) = q² + (2q + 1)Δ₁ + 2qΔ₂ + O(n⁻²ₘᵟᵢₙ).  (A.8)

These, together with (A.4), yield that E(T) = q[1 + α₁/nₘᵟᵢₙ] + O(n⁻²ₘᵟᵢₙ) and E(T²) = q(q + 2)[1 + α₂/nₘᵟᵢₙ] + O(n⁻²ₘᵟᵢₙ) where α₁ = nₘᵟᵢₙ(Δ₁ + Δ₂)/q and α₂ = nₘᵟᵢₙ[(2q + 8)Δ₁ + (2q + 6)Δ₂]/[q(q + 2)] as desired.

We now find the lower and upper bounds of Δ₁ and Δ₂ as given in (2.8). For l = 1, 2, . . . , k, set B_l = n_l⁻¹/₂H_lΣ_l¹/₂, a q × p full rank matrix so that Ω_l = B_lBᵀ_l. It follows that Ω_l are nonnegative, so are their eigenvalues. Notice that Ω_l and Ω_l = B_lᵀ_lB_l : p × p have the same nonzero eigenvalues. Thus, Ω_l has at most p nonzero eigenvalues. Denote the largest p eigenvalues of Ω_l by λ_l,r, r = 1, 2, . . . , p which include all the nonzero eigenvalues of Ω_l. It is easy to verify that ∑ₖ₌₁ⁿ Ω_l = I_q.

Therefore, we have ∑ₖ₌₁ⁿ tr(Ω_l) = q and I_q − Ω_l = ∑ᵣ≠ₗ Ωᵣ. Therefore I_q − Ω_l is nonnegative, showing that the eigenvalues of Ω_l are less than 1. It follows that tr(Ω_l²) = ∑ᵣ₌₁ᵖ λ_l,r² ≤ ∑ᵣ₌₁ᵖ λ_l,r = tr(Ω_l) and tr(Ω_l) = ∑ᵣ₌₁ᵖ λ_l,r ≤ p. These,
Proof of Theorem 4. Let $\mu$ and $\Sigma$ be the means and the covariance matrices of the responses $x_{ij}$, $j = 1, \ldots, n_l$ and the affine-transformed responses $\hat{x}_{ij}$, $j = 1, \ldots, n_l$ respectively. Then we have $\mu = \hat{B} \hat{\mu} + b$ and $\Sigma = \hat{B} \Sigma \hat{B}^T$. It follows that $\Sigma = H_{\Sigma} \Sigma H_{\Sigma}^T$, where $H_{\Sigma} = I_k \otimes B$ and $H_{\Sigma}^T = I_k \otimes B^T$. It follows that the GLHT problem (2.2) can be equivalently expressed as $H_0 : \hat{C} \hat{\mu} = \hat{b}$, versus $H_1 : \hat{C} \hat{\mu} \neq \hat{b}$, where $\hat{C} = \hat{B}^{-1}$ and $\hat{b} = \hat{C} \hat{B}^{-1} \hat{b} + c$.

Since $\hat{\mu}$ and $\hat{\Sigma}$ denote the unbiased estimators of $\mu$ and $\Sigma$, respectively, then by the affine-invariance of $H_0$ and $H_1$, we have $\hat{\mu} = E \hat{\mu} = \hat{b}$ and $\hat{\Sigma} = E \hat{\Sigma} = \hat{B} \Sigma \hat{B}^T$. It follows that $\hat{\mu} = \hat{B} \hat{\mu} + b$ and $\hat{\Sigma} = \hat{B} \Sigma \hat{B}^T$. Using the above, we have $C \hat{\mu} - \hat{c} = C \hat{B}^{-1} \hat{B} \hat{\mu} + b - (C \hat{B}^{-1} \hat{b} + c) = C \mu - \hat{c}$ and $C \hat{\Sigma} \hat{C}^T = C \hat{B}^{-1} \hat{B} \Sigma \Sigma \hat{B}^T (C \hat{B}^{-1})^T = C \Sigma \Sigma C$. Thus, both $C \hat{\mu} - \hat{c}$ and $C \hat{\Sigma} \hat{C}^T$ are affine-invariant. It follows that $T(2.3)$ is affine-invariant.

We now turn to show that $\hat{\mu}$ and $\hat{\Sigma}$ are affine-invariant. It is sufficient to show that $\text{tr}(\hat{\Sigma})$ and $\text{tr}(\hat{\Sigma}^2)$ are affine-invariant. Since we have showed that $C \Sigma \Sigma C$ is affine-invariant. It follows that $\text{tr}(\hat{\Sigma})$ and $\text{tr}(\hat{\Sigma}^2)$ are affine-invariant. Since we have showed that $C \Sigma \Sigma C$ is affine-invariant.
affine-invariant, we only need to show that \( n_l^{-1} C_l \hat{\Sigma}_l C_l^T, l = 1, 2, \ldots, k \) are affine-invariant. This is obvious since \( \hat{C} = C \hat{B}^{-1} \) implies \( \hat{C}_l = C_l B^{-1}, l = 1, 2, \ldots, k \) and \( \hat{\Sigma} = B \hat{\Sigma} B^T \) implies \( \hat{\Sigma}_l = B \hat{\Sigma}_l B^T, l = 1, 2, \ldots, k \). The theorem is then proved. \( \square \)

**Proof of Theorem 5** To show this theorem, it is sufficient to show that \( T, \hat{\Delta}_1 \) and \( \hat{\Delta}_2 \) are invariant under different labeling schemes of the mean vectors \( \mu_l, l = 1, 2, \ldots, k \). Let \( l_1, l_2, \ldots, l_k \) be any permutation of 1, 2, \ldots, k. Then it is easy to see that \( \sum_{l=1}^k C_l \hat{\mu}_l = \sum_{u=1}^k C_{l_u} \hat{\mu}_{l_u} \), and \( \sum_{l=1}^k n_l^{-1} C_l \hat{\Sigma}_l C_l^T = \sum_{u=1}^k n_{l_u}^{-1} C_{l_u} \hat{\Sigma}_{l_u} C_{l_u}^T \), showing that \( C \hat{\mu} = \sum_{l=1}^k C_l \hat{\mu}_l \) and \( C \hat{\Sigma} C^T = \sum_{l=1}^k n_l^{-1} C_l \hat{\Sigma}_l C_l^T \) are invariant under different labeling schemes of the mean vectors and so is \( T \).

We now show that \( \hat{\Delta}_1 \) and \( \hat{\Delta}_2 \) are invariant under different labeling schemes of the mean vectors. Set \( S_l = n_l^{-1} C_l \hat{\Sigma}_l C_l^T, l = 1, 2, \ldots, k \) and \( S = C \hat{\Sigma} C^T \). By (2.10), we have \( \hat{\Delta}_1 = \sum_{l=1}^k \text{tr}(S_l S^{-1} S_l^2) \) and \( \hat{\Delta}_2 = \sum_{l=1}^k \text{tr}^2(S_l S^{-1}) \). Since \( S \) is previously shown to be invariant under different labeling schemes of the mean vectors, so are \( \hat{\Delta}_1 \) and \( \hat{\Delta}_2 \). This completes the proof of the theorem. \( \square \)

**References**

Anderson TW (2003) An introduction to multivariate statistical analysis. Wiley, New York

Belloni A, Didier G (2008) On the Behrens–Fisher problem: a globally convergent algorithm and a finite-sample study of the Wald, LR and LM tests. Ann Stat 36:2377–2408

Fujikoshi Y (2000) Transformations with improved chi-squared approximations. J Multivar Anal 72:249–263

Gamage J, Mathew T, Weerahandi S (2004) Generalized \( p \)-values and generalized confidence regions for the multivariate Behrens–Fisher problem and MANOVA. J Multivar Anal 88:177–189

James GS (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

Kim S (1992) A practical solution to the multivariate Behrens–Fisher problem. Biometrika 79:171–176

Krishnamoorthy K, Lu F (2010) A parametric bootstrap solution to the MANOVA under heteroscedasticity. J Statist Comput Simul 80(8):873–887

Krishnamoorthy K, Yu J (2004) Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem. Statist Prob Lett 66:161–169

Kshirsagar AM (1972) Multivariate analysis. Marcel Decker, New York

Letac G, Massam H (2004) All invariant moments of the Wishart distribution. Scand J Statist 31:295–318

Nel DG, Vander Merwe CA (1986) A solution to the multivariate Behrens–Fisher problem. Commun Statist Theory Methods 15:3719–3735

Tang KL, Algina J (1993) Performing of four multivariate tests under variance-covariance heteroscedasticity. Multivar Behav Res 28:391–405

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

Yanagihara H, Yuan KH (2005) Three approximate solutions to the multivariate Behrens–Fisher problem. Commun Statist Simul Comput 34:975–988

Zhang JT, Liu X (2011) A modified Bartlett test for linear hypotheses in heteroscedastic one-way ANOVA. Manuscript