Brazilian Journal of Biometrics, **44**, 1–22 (2026)
doi:10.28951/bjb.v44i2.935

BRAZILIAN JOURNAL OF BIOMETRICS
ISSN:2764-5290

ARTICLE

Proposal of multivariate multiple comparison tests with the control treatment

Lúcia Helena Costa Braz,*[,1] Miguel Carvalho Nascimento,[1] and Daniel Furtado Ferreira[2]

[1]Post-graduate Program in Statistics and Agricultural Experimentation, Institute of Exact and Technological Sciences, Federal University of Lavras, Lavras, Brazil
[2]Department of Statistics, Institute of Exact and Technological Sciences, Federal University of Lavras, Lavras, Brazil
*Corresponding author. Email: lucia.helena@ifmg.edu.br

(Received: March 07, 2025; Revised: October 16, 2025; Accepted: November 28, 2025; Published: May 12, 2026)
Section Editor: João Domingos Scalon

Abstract

Problems involving comparisons of treatment effectiveness for multivariate responses are common in various fields of knowledge. Typically, methods for comparing vectors of means use the Bonferroni inequality to construct conservative tests, avoiding the complexities of the exact distribution of the maximum test statistic $T^2_{\max}$, the maximum of Hotelling's T^2. In high-dimensional scenarios, traditional methods are not viable as they depend on the inverse of the sample covariance matrix, which becomes singular. To address this problem, Dempster's trace criterion can be used, and a second alternative is Ahmad's test statistic T_{ig}, however, in both cases, the Bonferroni inequality is used. Another issue is that both in the estimation process and in the exact distribution of statistics for multiple comparison tests, there is a need to deal with sophisticated and complex numerical methods. These facts make these approximations not readily usable. To try to overcome these challenges, this work proposes multivariate multiple comparison tests with the control treatment using the nonparametric bootstrap method. The performance of the tests was evaluated through experimentwise type I error rates (EER) and power in different scenarios using Monte Carlo simulation and the R program. The results showed that for homoscedastic scenarios, the proposed bootstrap test ATB showed more effective control of EER, in addition to having higher power, regardless of whether the distribution was normal or not, in both low and high-dimensional contexts. Thus, the ATB test proved to be the most recommended alternative in these situations. For heteroscedastic scenarios, it was not possible to identify a clearly superior test, but in several circumstances, the proposed bootstrap tests demonstrated superior performance compared to their respective asymptotic versions.

Keywords: Bootstrap; High dimensionality; Experimentwise type I error; Power; R program.

1. Introduction

In agricultural sciences, plant genetics and breeding, clinical trials, engineering, geology, soil sciences, animal science, pediatric surgery (Staffa & Zurakowski, 2020) and many other areas of human knowledge, problems often occur where comparisons of vectors of means are needed to determine the effectiveness of treatments for multivariate responses. Traditional methods for performing multiple comparisons (Dean & Voss, 1999; Dean *et al.,* 2016; Hinkelman & Kempthorne, 2008; Hochberg & Tamhane, 1987; Hsu, 1999; Machado *et al.,* 2005) are generally for univariate responses and are not adequate, as measurements are obtained in more than one response variable. Multivariate multiple comparisons, when applied, usually use the Bonferroni inequality to construct a conservative test that avoids the complexities of the exact distribution of the test statistic which is a $T^2_{\max}$ statistic, i.e., the maximum of a set of Hotelling's T^2 statistics (Kakizawa, 2008, 2009). One problem is that this set of interest does not have a distribution independent of its components. Thus, there is a great difficulty in obtaining the distribution of its maximum, which is necessary in multiple comparison problems to avoid the multiplicity problem. Generally, approximations are obtained, which in turn are conservative. One of them are procedures based on the Bonferroni inequality (Nishiyama *et al.,* 2014; Seo & Nishiyama, 2008). The main approximate procedures based on the Bonferroni inequality are the first-order and modified second-order Bonferroni procedures (Nishiyama *et al.,* 2014).

Another problem is that both in the estimation process and in the distribution of statistics for multivariate multiple comparison tests, there is a need to deal with sophisticated and complex numerical methods (Kakizawa, 2009; Nishiyama *et al.,* 2014). An alternative to overcome these numerical difficulties is the use of the bootstrap method (Westfall, 2011; Westfall & Young, 1989), and this is one of the solutions presented in this work.

In multivariate procedures, there is a problem that arises when the number of variables p is greater than the sample size n, which constitute the so-called high-dimensionality cases. In this context, it is not possible to use existing methods for low dimensionality, because they depend on the inverse of the sample covariance matrix which, in this situation, cannot be obtained since the sample covariance matrix is singular. To overcome this problem, Nishiyama *et al.* (2014) proposed the use of Dempster's trace criterion (Dempster, 1958, 1960). There are also problems when multivariate data do not come from multivariate normal distributions and when, analogous to the univariate case, treatment covariance matrices are not homogeneous. These cases under generally unfavorable conditions were addressed by Ahmad (2018, 2019) presenting asymptotic solutions based on chi-square and standard normal distributions.

The characteristic of an estimation method or hypothesis test in maintaining its optimal properties when its assumptions are violated is called robustness. In uni- and multivariate statistics, experimental or sample data (or residuals) are often contaminated by *outliers* or their distribution is not normal or there are problems with heteroscedasticity. In these cases, inference may be compromised, with unrealistic results to explain the studied phenomena or unsatisfactory solutions to real problems. This work computationally intensively addressed (bootstrap or Monte Carlo) the case called general conditions and high dimensionality presented in Ahmad (2018, 2019). This potential solution was also investigated under ideal conditions, i.e., normal, identically distributed, homoscedastic cases, in low and high-dimensional situations.

In all cases, the performance of the new or considered statistical methods was evaluated to confirm or not the expectation that they are more efficient than existing procedures. This evaluation was performed computationally through Monte Carlo simulations (Gentle, 2003; Manly, 1997). The importance of new statistical methods for the development of new varieties, technologies, machines, medicines, vaccines, among many other examples is indisputable. The proposal of a new method is justified by the expectation that it may be more powerful in detecting increasingly smaller differences between treatment effects when compared to existing methods.

Given the above, the objective of this work is to propose multivariate multiple comparison tests with the control treatment, considering the nonparametric bootstrap method. As specific objectives, we have: to evaluate the performance of the new tests through experimentwise type I error rates and power and to compare the performance of the new proposals with existing methods, using Monte Carlo simulations.

2. Methods

Without loss of generality, the p-dimensional random vectors (p is the number of variables) $\boldsymbol{X}_{ik} = [X_{ik1}, \ldots, X_{ikp}]^\top \sim \mathfrak{F}_i$, which represent the sample observations referring to the i-th treatment in the k-th sampling unit, $i = 1, 2, \ldots, g$, $k = 1, 2, \ldots, r_i$, with p-dimensional vector of means $E(\boldsymbol{X}_{ik}) = \boldsymbol{\mu}_i$ and covariance matrix $Cov(\boldsymbol{X}_{ik}) = \boldsymbol{\Sigma}_i$ positive definite $p \times p$ symmetric, with $g \geq 2$ and $\mathfrak{F}_i$ a distribution family for the i-th treatment. The very general case was considered, where $\mathfrak{F}_i$ is any p-variate distribution, not necessarily normal, with potentially unequal (heterogeneous) covariance matrices $\boldsymbol{\Sigma}_i$s and unbalanced case, where the r_is are different (r_i is the number of replications of the i-th treatment). The complete sample is the combination of the samples of the g treatments into a single sample, with n being the total size of the combined sample, where $n = \sum_{i=1}^{g} r_i$. Cases where $p < n - g$ were considered, but cases with $p \geq n - g$, of high dimensionality, were also considered, where $n - g = \nu$ are the degrees of freedom associated with the estimator of the common population covariance matrix. Other more specific cases of particular interest were also considered where the distribution family $\mathfrak{F}_i$ refers to the multivariate normal distribution.

The inference interest is in the p-dimensional vector parameters $\boldsymbol{\delta}_{ig} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_g$, for $i = 1, 2, \ldots, g-1$, i.e., the interest is in multivariate multiple comparisons with the control treatment g, resulting in $m = g - 1$ parameters or comparisons of interest between the vectors of means.

The unbiased estimators of the vector of means $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{\Sigma}_i$ of the i-th treatment are given, respectively, by

$$\bar{\boldsymbol{X}}_i = \frac{1}{r_i}\sum_{k=1}^{r_i} \boldsymbol{X}_{ik} \qquad \text{and} \qquad \boldsymbol{S}_i = \frac{1}{r_i - 1}\sum_{k=1}^{r_i}(\boldsymbol{X}_{ik} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ik} - \bar{\boldsymbol{X}}_i)^\top, \qquad i = 1,\ 2,\ \ldots,\ g. \tag{1}$$

When the treatment covariance matrices are equal, i.e., the homogeneous case where $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, the estimator of the common covariance matrix $\boldsymbol{\Sigma}$ is given by $\boldsymbol{S} = 1/(n-g)\sum_{i=1}^{g}(r_i - 1)\boldsymbol{S}_i$, associated with $\nu = n - g$ degrees of freedom. The estimator of $\boldsymbol{\delta}_{ig}$ is $\hat{\boldsymbol{\delta}}_{ig} = \bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}}_g$. The null hypothesis to be tested is

$$H_0 : \boldsymbol{\delta}_{ig} = \boldsymbol{0}, \qquad i = 1,\ 2,\ \ldots,\ g-1, \tag{2}$$

against the alternative $\boldsymbol{\delta}_{ig} \neq \boldsymbol{0}$. The tests will be described next assuming the sample structures described above.

2.1 Multivariate multiple comparison tests for low dimensionality

2.1.1 Hotelling's T^2 test with first-order Bonferroni approximation (T2FO)

Under H_0 given in (2), the value of $\boldsymbol{\delta}_{ig}$ is the null vector $\boldsymbol{0}$, $\forall\ i = 1, 2, \ldots, g-1$. Thus, substituting this value, the statistic

$$T^2_{ig} = \frac{r_i r_g}{r_i + r_g}(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}}_g - \boldsymbol{\delta}_{ig})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}}_g - \boldsymbol{\delta}_{ig}), \qquad i = 1,\ 2,\ \ldots,\ g-1, \tag{3}$$

was computed for all pairs involving the control treatment g, without loss of generality. Thus, $m = g - 1$ values of the T^2 statistic were obtained in the case of comparisons with the control treatment.

The value of the statistic for each pair was compared with the critical value based on the approximate distribution of $T^2_{\max}$. Thus, when the value computed in (3) was greater than the critical value presented in Nishiyama *et al.* (2014) and given by $t^2_1 = [\nu p/(\nu - p + 1)]/F_{p,\nu-p+1}(\alpha/m)$, the pairs of vectors of means were considered significantly different at the nominal significance level α, where $F_{p,\nu-p+1}(\alpha/m)$ is the $100\alpha/m\%$ quantile of the F distribution with p and $\nu-p+1$ degrees of freedom.

2.2 Multivariate multiple comparison tests for high dimensionality

2.2.1 Hyodo, Takahashi and Nishiyama test (HTNT)

Under H_0 given in (2), we have that $\boldsymbol{\delta}_{ig} = \mathbf{0}$, $\forall\ i = 1, 2, \ldots, g-1$. Thus, substituting this value, the test statistic of Hyodo *et al.* (2014) (HTNT) was computed by

$$D_{ig} = \frac{p}{\hat{\sigma}}\left\{\frac{(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}}_g - \boldsymbol{\delta}_{ig})^\top(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}}_g - \boldsymbol{\delta}_{ig})}{r_{ig}\,tr(\boldsymbol{S})} - 1\right\}, \tag{4}$$

where $r_{ig} = 1/r_i + 1/r_g$, $\hat{\sigma} = \sqrt{2p\hat{a}_2/\hat{a}_1^2}$, $tr(\boldsymbol{S})$ corresponds to the trace of matrix $\boldsymbol{S}$, and the constants $\hat{a}_1$ and $\hat{a}_2$ are given by $\hat{a}_1 = \frac{tr(\boldsymbol{S})}{p}$ and $\hat{a}_2 = \frac{\nu^2}{(\nu+2)(\nu-1)p}\left[tr(\boldsymbol{S}^2) - \frac{tr^2(\boldsymbol{S})}{\nu}\right]$.

The p-value was obtained by (ibid.) p-value $= 1 - prob$, where

$$prob = \left\{\Phi(D_{ig}) - \phi(D_{ig})\left[\frac{1}{\sqrt{p}}\frac{\sqrt{2}\hat{a}_3}{\sqrt{\hat{a}_2^3}}h_2(D_{ig}) + \frac{1}{p}\left(\frac{\hat{a}_4}{2\hat{a}_2^2}h_3(D_{ig}) + \frac{\hat{a}_3^2}{9\hat{a}_2^3}h_5(D_{ig})\right) + \frac{1}{2\nu}h_1(D_{ig})\right]\right\},$$

$\Phi(D_{ig})$ and $\phi(D_{ig})$ are the cumulative distribution function and probability density function of the standard normal distribution, respectively, and $h_j(D_{ig})$s, $j = 1, 2, 3, 5$, are the Hermite polynomials given by $h_1(D_{ig}) = D_{ig}$, $h_2(D_{ig}) = D_{ig}^2 - 1$, $h_3(D_{ig}) = D_{ig}^3 - 3D_{ig}$, $h_5(D_{ig}) = D_{ig}^5 - 10D_{ig}^3 + 15D_{ig}$. The other necessary quantities not yet defined are given by $\hat{a}_3 = \nu^4/[(\nu+4)(\nu+2)(\nu-1)(\nu-2)p][tr(\boldsymbol{S}^3) - 3tr(\boldsymbol{S}^2)tr(\boldsymbol{S})/\nu + 2tr^3(\boldsymbol{S})/\nu^2]$ and $\hat{a}_4 = \nu^3[b_1 tr(\boldsymbol{S}^4) + b_2 tr(\boldsymbol{S}^3)tr(\boldsymbol{S}) + b_3 tr^2(\boldsymbol{S}^2) + b_4 tr(\boldsymbol{S}^2)tr^2(\boldsymbol{S}) + b_5 tr^4(\boldsymbol{S})]/[(\nu+6)(\nu+4)(\nu+2)(\nu+1)(\nu-1)(\nu-2)(\nu-3)p]$, where the b_is are $b_1 = \nu^2(\nu^2+\nu+2)$, $b_2 = -4\nu(\nu^2+\nu+2)$, $b_3 = -\nu(2\nu^2+3\nu-6)$, $b_4 = 2\nu(5\nu+6)$ and $b_5 = -(5\nu+6)$.

Therefore, in the case of comparisons with the control, the null hypothesis was rejected when the p-value was less than or equal to the adopted nominal significance level with Bonferroni protection, i.e., when p-value $\leq \alpha/m$.

2.2.2 Ahmad's test (AT)

The test statistic, T_{ig}, was computed by

$$T_{ig} = 1 + \frac{Q_{ig0}}{Q_{ig1}/p}, \tag{5}$$

where $Q_{ig1} = Q_{i1} + Q_{g1}$, with $Q_{i1} = (E_i - U_i)/r_i$, $E_i = \sum_{k=1}^{r_i}(\boldsymbol{X}_{ik}^\top \boldsymbol{X}_{ik})/r_i$ and $U_i = 1/[r_i(r_i - 1)]\sum_{k=1}^{r_i}\sum_{\substack{\ell=1,\\ k\neq\ell}}^{r_i}\boldsymbol{X}_{ik}^\top \boldsymbol{X}_{i\ell}$ and the quantity $Q_{ig0} = U_{ig0}/p$, where $Q_{ig0} = U_i + U_g - 2U_{ig}$, with $U_{ig} = 1/r_i r_g \sum_{k=1}^{r_i}\sum_{\ell=1}^{r_g}\boldsymbol{X}_{ik}^\top \boldsymbol{X}_{g\ell}$, as in Ahmad (2018).

The null hypothesis $H_0 : \boldsymbol{\delta}_{ig} = \mathbf{0}$ was rejected if $T_{ig} \geq T_{\alpha/m}$, where $T_{\alpha/m}$ is given by $T_{\alpha/m} = \chi^2_{\alpha/m;f_{ig}}/f_{ig}$, with $\chi^2_{\alpha/m;f_{ig}}$ being the upper $100\alpha/m\%$ quantile of the chi-square distribution with degrees of freedom f_{ig} given by $f_{ig} = [tr(\boldsymbol{\Omega}_{0ig})]^2/tr(\boldsymbol{\Omega}_{0ig}^2)$, where $\boldsymbol{\Omega}_{0ig} = (n/p)\hat{\boldsymbol{\Sigma}}_{0ig}$ and $\hat{\boldsymbol{\Sigma}}_{0ig} = \boldsymbol{S}_i/r_i + \boldsymbol{S}_g/r_g$.

2.2.3 Proposed bootstrap test based on Hyodo, Takahashi and Nishiyama test (HTNTB)

The statistic D_{ig} was computed in the original sample, using expression (4), for the m pairs in comparisons with the control. With the estimates $\bar{\boldsymbol{X}}_i$ from (1) of the original sample, the modified sample was obtained as $\boldsymbol{Y}_{ik} = \boldsymbol{X}_{ik} - \bar{\boldsymbol{X}}_i$, for $i = 1, 2, \ldots, g$ and $k = 1, 2, \ldots, r_i$.

This sample was combined by grouping the $n = \sum_{i=1}^{g} r_i$ observations $\boldsymbol{Y}_{ik}$ into a single p-dimensional sample of size n, thus imposing the null hypothesis of equality of vectors of means. This combined sample was resampled with replacement, recreating the structure of the original sample of g treatments with r_i p-variate observations of the i-th treatment, with $i = 1, 2, \ldots, g$. This process was repeated $B = 2,000$ times.

In each generated dataset, the m statistics D_{ig} in (4) were obtained for all pairs of treatments for comparisons with the control, totaling $m = g - 1$ values. Thus, the final statistic D_ℓ, in the ℓ-th bootstrap resampling, $\ell = 1, 2, \ldots, B$, or in the original sample, when $\ell = B + 1$, derived from the $D_{\max}$ statistic, was obtained by

$$D_\ell = \max\left\{D_{1g},\ D_{2g},\ \ldots,\ D_{(g-2)g},\ D_{(g-1)g}\right\}, \tag{6}$$

in comparisons with the control.

The D_ℓ values, obtained in the B bootstrap resamplings, were stored together with the D_ℓ value obtained in the original sample, forming a vector of dimension $B + 1$, given by $\boldsymbol{D} = [D_1, D_2, \ldots, D_\ell, \ldots, D_B, D_{B+1}]^\top$.

Then, p-value $= \sum_{\ell=1}^{B+1} I\left(D_\ell \geq D_{B+1}\right)/(B+1)$ was computed, where $I\left(D_\ell \geq D_{B+1}\right)$ is the indicator function that returns 1 if $D_\ell \geq D_{B+1}$ and 0 otherwise. The null hypothesis was rejected when the obtained p-value was less than or equal to the adopted nominal significance level.

2.2.4 Proposed bootstrap test based on Ahmad's test (ATB)

In the same way as described for the D_{ig} statistic, the use of the T_{ig} statistic was considered, using expression (5). Thus, in the original sample, the statistics T_{ig} in (5) were computed. In the bootstrap resamplings, described in the previous section, the statistics were computed and the set of interest given in (6) was computed and the maximum obtained, with the values of D_{ig} replaced by T_{ig}. A set $\boldsymbol{T}$ of maxima was obtained in the B resamplings and in the original sample, $\boldsymbol{T} = [T_1, T_2, \ldots, T_\ell, \ldots, T_B, T_{B+1}]^\top$, and the p-value was computed by p-value $= \sum_{\ell=1}^{B+1} I\left(T_\ell \geq T_{B+1}\right)/B + 1$, where $I\left(T_\ell \geq T_{B+1}\right)$ is the indicator function that returns 1 if $T_\ell \geq T_{B+1}$ and 0 otherwise. The null hypothesis was rejected when the obtained p-value was less than or equal to the adopted nominal significance level.

In the next subsection, the strategies considered in this work to evaluate the performance of the tests are presented.

2.3 Simulations and evaluation of test performance

The evaluation of test performance was performed via Monte Carlo simulation and in two stages. In the first, where simulations were performed under the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$, the proportion of rejections of the null hypothesis is related to the experimentwise type I error rate and, in the second, performed under the hypothesis H_1, the proportion of rejections is related to the power.

Among the configurations used by Ahmad (2019) in the simulations, the following were considered: $g = 3$, with $(r_1, r_2, r_3) = (10, 15, 20), (50, 75, 100)$; $g = 6$, with $(r_1, r_2, r_3, r_4, r_5, r_6) = (10, 10, 10, 20, 20, 20), (30, 40, 50, 60, 70, 80)$; independent and identically distributed p-dimensional vectors generated from multivariate normal and uniform$(0, 1)$ distributions; $p \in \{50, 300, 500\}$; two covariance matrix structures: compound symmetry (CS) and first-order autoregressive ((AR(1))),

defined respectively as $\sigma^2[(1-\rho)\boldsymbol{I}+\rho\boldsymbol{J}]$ and $Cov(X_i, X_j) = \kappa\rho^{|i-j|}\ \forall\ i, j$, where $\boldsymbol{I}$ is an identity matrix and $\boldsymbol{J} = \mathbf{1}\mathbf{1}^\top$ is a matrix of ones. Values of σ^2 and κ were set to 1.

To include violation of the homoscedasticity assumption, for the case of $g = 3$, the following configuration was considered for the $\boldsymbol{\Sigma}_i$ structures, $i = 1, 2, 3$: CS(0.5), AR(1, 0.5), AR(1, 0.7), respectively, where 0.5 and 0.7 are the values of ρ. For $g = 6$, the same combinations of covariance matrix structures used for $g = 3$ were employed, repeated for the first three and the next three treatments.

In addition to the configurations presented in Ahmad (2019), $p = 5$ was also considered to include a small value of p, $p = 100$, and the multivariate t_3 distribution, in order to introduce heavy-tailed scenarios and to compare its results with those obtained under the multivariate normal distribution. To include larger values of g, $g = 60$ and $g = 100$ were considered, with $r_i = 3$ or 6, $\forall\ i = 1, 2, \ldots, g$, and for these cases, $p \in \{5, 250\}$. For both $g = 60$ and $g = 100$, the same combination of covariance matrix structures as for $g = 3$ was used, which were repeated until completing $\boldsymbol{\Sigma}_{60}$ or $\boldsymbol{\Sigma}_{100}$.

Finally, unlike Ahmad (ibid.), homoscedastic cases were also simulated and evaluated, with $g = 3$, 6, 60, 100, $p = 5$, 250, $(r_1, r_2, r_3) = (10, 10, 10)$, $(10, 15, 20)$, $(r_1, r_2, r_3, r_4, r_5, r_6) = (10, 10, 10, 10, 10, 10)$, $(10, 10, 10, 20, 20, 20)$, $r_i = 3$ or 6, $\forall\ i = 1, 2, \ldots, 60$ or 100, and the considered covariance matrix structure was CS(0.5), $\forall\ i = 1, 2, \ldots, g$.

2.3.1 Evaluation of type I error rates per experiment

In the first stage, evaluating type I error rates per experiment, data were generated under the complete null hypothesis. Therefore, rejection of the null hypothesis of equality between two mean vectors, one from a regular treatment and the other from the control treatment, was considered a type I error. In this work, the probability considered for measuring type I error was per experiment. This probability is estimated by the proportion of experiments with at least one incorrectly detected difference in the comparisons of regular treatments with the control relative to the total of N simulated experiments. Thus, the actual type I error rate per experiment was estimated by $\hat{\alpha} = \sum_{k=1}^{N} I(E_k = 1)/N$, where E_k is a binary variable that takes the value 1 if at least one type I error occurred in the k-th experiment or 0 otherwise, for $k = 1, 2, \ldots, N$, and $I(E_k = 1)$ is the indicator function that returns 1 if the equality is verified and 0 otherwise.

The type I error rates per experiment were estimated for each proposed test and for the existing asymptotic tests considered in this work, being compared among themselves and with the nominal significance level using the exact binomial test (Oliveira & Ferreira, 2010), considering Bonferroni protection, to verify whether the tests are considered liberal, conservative, or exact.

2.3.2 Exact binomial test considering Bonferroni protection

The null hypothesis considered for the exact binomial test (ibid.) with $(100 - \alpha^*)\%$ confidence was

$$H_0 : \alpha = \alpha_0 = 0.05 \ \ (\text{or}\ \ 0.01\ \ \text{or}\ \ 0.10) \tag{7}$$

against the alternative hypothesis $H_1 : \alpha \neq \alpha_0 = 0.05$, (or 0.01 or 0.10).

If the null hypothesis in (7) was rejected and the empirical type I error rate per experiment was considered significantly lower than the adopted nominal significance level, the test was considered conservative; if the null hypothesis in (7) was rejected and the empirical type I error rate per experiment was considered significantly higher than the adopted nominal significance level, the test was considered liberal; and if the null hypothesis in (7) was not rejected, the test was considered exact.

Considering γ the number of rejections of H_0 in N Monte Carlo simulations, for a significance level α_0, the exact binomial test statistic was obtained using the relationship between the F distribution and the binomial distribution, with success probability $p = \alpha_0$, and is given (ibid.)

by $F = [(\gamma + 1)/(N - \gamma)]/[(1 - \alpha_0)/(\alpha_0)]$, under H_0. This statistic follows an F distribution with $\nu_1 = 2(N - \gamma)$ and $\nu_2 = 2(\gamma + 1)$ degrees of freedom. If $F_c \leq F_{\nu_1,\nu_2}(\alpha^*/2)$ or $F_c > F_{\nu_1,\nu_2}(1 - \alpha^*/2)$, then the null hypothesis in (7) was rejected at the 1% significance level with Bonferroni protection, where $F_{\nu_1,\nu_2}(\alpha^*/2)$ and $F_{\nu_1,\nu_2}(1 - \alpha^*/2)$ are the $100\alpha^*/2\%$ and $100(1 - \alpha^*/2)\%$ quantiles, respectively, of the F distribution with ν_1 and ν_2 degrees of freedom and $\alpha^* = 1\%/s$, where s is the number of simultaneous inferences performed, which was:

1. $s = 48$, in homoscedastic cases in low dimensionality (there are 6 tests and, for each of these, 8 combinations of g, p, and $r_i s$, totaling $6 \times 8 = 48$ simultaneous inferences in this case);
2. $s = 24$, in homoscedastic cases in high dimensionality;
3. $s = 84$, in heteroscedastic cases in low dimensionality;
4. $s = 48$, in heteroscedastic cases in high dimensionality.

2.3.3 Definitions on test classifications regarding their size

After applying the exact binomial test considering Bonferroni protection, if the null hypothesis in (7) was rejected and the empirical type I error rate per experiment was considered significantly higher (or lower) than the adopted nominal significance level, the test was considered liberal (or conservative), and then a procedure was adopted to evaluate it as slightly or expressively liberal (or conservative). For this, an exact confidence interval for binomial proportions was obtained, using the nominal confidence level of 1% with Bonferroni protection, protecting the global significance level in the s simultaneous inferences performed.

Since there is no randomness, a pseudo-observation, γ, was obtained by multiplying the significance level α_0 (0.01, 0.05, or 0.10) by N and rounding to the nearest integer, where $N = 2,000$. Once this interval was obtained, i.e., the lower (LI) and upper (LS) limits, which would be the equivalent values where, below LI and above LS, the exact binomial test would reject the null hypothesis (7), the following procedure was performed:

1. if the empirical type I error rate per experiment was greater than LS and less than $(LS - \alpha_0) + LS$, the test, which had been classified as liberal by the exact binomial test, was considered slightly liberal; and if it was greater than $(LS - \alpha_0) + LS$, the test was considered expressively liberal;
2. analogously, if the empirical type I error rate per experiment was less than LI and greater than $LI - (\alpha_0 - LI)$, the test, which had been classified as conservative by the exact binomial test, was considered slightly conservative; and if it was less than $LI - (\alpha_0 - LI)$, the test was considered expressively conservative.

This evaluation of the tests is justified by the belief that slightly liberal tests may be considered plausible for use in research, provided they are within the risk acceptable to the researcher in an experiment, since, as a consequence, greater power is expected for these tests. It is worth reinforcing that the ideal is to have an exact or conservative test, provided that, in the latter case, the test has power as high as or close to those that were exact in controlling the TEE.

2.3.4 Power evaluation

In the second stage, power evaluation, under the alternative hypothesis, the simulations were performed following the procedures described for the type I error rate per experiment, except that, in this case, $\boldsymbol{\mu}_g = \boldsymbol{\mu}_0$, for some $\boldsymbol{\mu}_0$, without loss of generality, and $\boldsymbol{\mu}_i = \boldsymbol{\mu}_g + \boldsymbol{\Delta}_{ig}$, $i = 1, \ldots, g - 1$,

where $\boldsymbol{\Delta}_{ig}$ was obtained according to the simulated multivariate distribution, being:

$$\boldsymbol{\Delta}_{ig} = \begin{cases} \phi\sqrt{\frac{1}{r_i}diag(\boldsymbol{\Sigma}_i) + \frac{1}{r_g}diag(\boldsymbol{\Sigma}_g)}, & \text{for multivariate normal distribution;} \\ \phi\sqrt{\frac{1}{r_i}diag\left[\frac{\nu}{\nu-2}\boldsymbol{\Sigma}_i\right] + \frac{1}{r_g}diag\left[\frac{\nu}{\nu-2}\boldsymbol{\Sigma}_g\right]}, & \text{for multivariate } t_3 \text{ distribution;} \\ \phi\sqrt{\frac{1}{r_i}diag\left[\frac{1}{12}\boldsymbol{\Sigma}_i\right] + \frac{1}{r_g}diag\left[\frac{1}{12}\boldsymbol{\Sigma}_g\right]}, & \text{for multivariate uniform (0, 1) distribution,} \end{cases}$$

where $\boldsymbol{\Sigma}_i$ is the covariance matrix of the simulated data vector from the i-th treatment and ϕ = 1, 2, 4, and 8.

By defining $\boldsymbol{\mu}_g$ and $\boldsymbol{\mu}_i$ as described above, the objective was to create treatments whose mean vector of the i-th treatment, $i = 1, \ldots, g-1$, was different from the mean vector of the control treatment g by a difference of ϕ standard errors of the difference between two means for each univariate component of these vectors.

Power was computed by the average power method, which is given by the average proportion of correct rejections among all false hypotheses (Bretz *et al.,* 2011), and was estimated for each proposed test and for the existing asymptotic tests considered in this work, being compared among themselves.

3. Results and Discussion

From the evaluation of the results obtained in the simulations for the three adopted significance levels (0.01, 0.05, and 0.10), similar behaviors were verified when considering the same configurations for each α. Therefore, the simulation results considering α = 0.05 were presented and discussed. A similar response pattern of the tests was also observed for the multivariate normal and uniform (0, 1) distributions, and for this reason, the results obtained for the multivariate uniform (0, 1) will not be presented in the tables and graphs.

Considering the exact binomial test (Oliveira & Ferreira, 2010) and the methodology defined in subsection 2.3.3, the results obtained for the type I error rates per experiment (TEEs) were flagged with symbols to classify the tests as slightly liberal (+), expressively liberal (++), slightly conservative (–), expressively conservative (—), or exact (no symbol).

In the power graphs, the horizontal dotted lines indicate the lower and upper limits obtained in the classification of the tests regarding size, which would be the equivalent values where, below LI and above LS, the exact binomial test, considering Bonferroni protection, would reject the hypothesis (7).

3.1 Homoscedastic cases

3.1.1 Type I error rates per experiment

Table 1 presents the results obtained for TEE, under homoscedasticity, low and high dimensionality, for the multivariate normal and t_3 distributions. In the ideal situation, i.e., multivariate normal, homoscedasticity, and low dimensionality, only HTNTB and ATB controlled the TEE exactly in all configurations. The T2FO test also showed good performance, being slightly conservative only in g = 60, p = 250 and g = 100, p = 5. In the other configurations, T2FO was exact. The HTNT and AT tests were expressively liberal in the cases of g = 60 and 100, AT was also slightly liberal in g = 6, and HTNT was slightly conservative in g = 3, p = 5, balanced case. In the simulations performed by Hyodo *et al.* (2014), T2FO and HTNT also showed conservative behavior in some situations, although the authors considered only g = 3 and 6.

For the multivariate t_3 distribution and low dimensionality, it can be observed (Table 1) that T2FO and HTNTB became expressively liberal in $g = 60$ and 100. HTNT and AT maintained behaviors similar to those observed for the multivariate normal, and ATB remained exact in controlling the TEE in all configurations.

Table 1. Type I error rates per experiment of the T2FO, HTNT, TA, HTNTB, and ATB tests, considering covariance matrix structures (CS), number of treatments (g), sample sizes ($r_1, \ldots, r_g$), number of variables (p), multivariate normal and t_3 distributions, nominal significance level $\alpha = 0.05$, under H_0, homoscedastic cases, low and high dimensionality

$\alpha = 0.05$, $\boldsymbol{\Sigma}_i$: CS (0.5), $\forall\, i = 1, \ldots, g$							
g	p	$(r_1, \ldots, r_g)$	T2FO	HTNT	AT	HTNTB	ATB
			Low dimensionality				
			Multivariate normal				
3	5	$(10, 10, 10)$	0.0465	0.0335^{-}	0.0550	0.0425	0.0420
	5	$(10, 15, 20)$	0.0445	0.0375	0.0640	0.0535	0.0530
6	5	$(10, 10, 10, 10, 10, 10)$	0.0480	0.0565	0.0790^{+}	0.0550	0.0560
	5	$(10, 10, 10, 20, 20, 20)$	0.0475	0.0510	0.0765^{+}	0.0510	0.0525
60	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.0380	0.1495^{++}	0.2030^{++}	0.0555	0.0470
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.0335^{-}	0.1355^{++}	0.1800^{++}	0.0480	0.0470
100	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.0300^{-}	0.1540^{++}	0.2070^{++}	0.0515	0.0385
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.0370	0.1700^{++}	0.2100^{++}	0.0485	0.0520
			Multivariate t_3				
3	5	$(10, 10, 10)$	0.0420	0.0345	0.0505	0.0460	0.0430
	5	$(10, 15, 20)$	0.0400	0.0235^{-}	0.0425	0.0465	0.0400
6	5	$(10, 10, 10, 10, 10, 10)$	0.0400	0.0495	0.0480	0.0520	0.0460
	5	$(10, 10, 10, 20, 20, 20)$	0.0590	0.0485	0.0440	0.0470	0.0440
60	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.3610^{++}	0.5245^{++}	0.1195^{++}	0.1355^{++}	0.0470
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.2855^{++}	0.4785^{++}	0.0740^{+}	0.0955^{++}	0.0420
100	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.5790^{++}	0.6825^{++}	0.1325^{++}	0.1810^{++}	0.0390
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.8045^{++}	0.6685^{++}	0.0825^{+}	0.1155^{++}	0.0390
			High dimensionality				
			Multivariate normal				
3	250	$(10, 10, 10)$	-	0.0450	0.0765^{+}	0.0540	0.0590
	250	$(10, 15, 20)$	-	0.0335^{-}	0.0670	0.0485	0.0505
6	250	$(10, 10, 10, 10, 10, 10)$	-	0.0580	0.0870^{+}	0.0560	0.0560
	250	$(10, 10, 10, 20, 20, 20)$	-	0.0565	0.0965^{++}	0.0540	0.0570
60	250	$r_i = 3, \forall\, i = 1, \ldots, g$	-	0.1550^{++}	0.1610^{++}	0.0580	0.0565
100	250	$r_i = 3, \forall\, i = 1, \ldots, g$	-	0.2075^{++}	0.1915^{++}	0.0645	0.0580
			Multivariate t_3				
3	250	$(10, 10, 10)$	-	0.0245^{-}	0.0400	0.0435	0.0345
	250	$(10, 15, 20)$	-	0.0300^{-}	0.0530	0.0540	0.0505
6	250	$(10, 10, 10, 10, 10, 10)$	-	0.0530	0.0555	0.0560	0.0455
	250	$(10, 10, 10, 20, 20, 20)$	-	0.0570	0.0460	0.0530	0.0415
60	250	$r_i = 3, \forall\, i = 1, \ldots, g$	-	0.5810^{++}	0.0490	0.1500^{++}	0.0305^{-}
100	250	$r_i = 3, \forall\, i = 1, \ldots, g$	-	0.7585^{++}	0.0615	0.2040^{++}	0.0350

For the high dimensionality case, in general, the behavior of the HTNT, AT, HTNTB, and ATB tests was similar to that verified in the low dimensionality cases, which was also observed by

Hyodo *et al.* (2014) for HTNT and by Ahmad (2019) for AT in the simulations performed by the authors. A noteworthy change is that AT went from slightly liberal in the low dimensionality case to exact in the multivariate t_3 distribution in g = 60 and 100, p = 250 and high dimensionality. It is also worth highlighting that, for the multivariate normal, HTNTB and ATB remained exact in all configurations, while AT was liberal (expressively or not), in most cases.

Therefore, when considering homoscedastic scenarios, both in low and high dimensionality, regardless of whether the multivariate distribution is normal or t_3 and regardless of whether the values of g are small (g = 3 or 6) or large (g = 60 or 100), the proposed ATB test stood out in controlling the TEE. In the multivariate uniform case, data not shown, the response patterns were similar to the multivariate normal. Nascimento *et al.* (2025) proposed the ATB test for comparisons between two independent mean vectors and, in their simulations, also concluded that this was the only one exact in all simulated homoscedastic scenarios, i.e., the results found here corroborate those obtained by the authors for the ATB test in the case of comparisons between two mean vectors.

Finally, it is worth noting that for larger values of g, as occurred in low dimensionality, the asymptotic tests based on the Bonferroni inequality, HTNT and AT, were not expressively conservative, as expected when using the Bonferroni criterion, due to the increase in the number of treatments and, consequently, in the number of comparisons (Nishiyama *et al.,* 2014; Seo *et al.,* 1994).

3.1.2 Power

The evaluation of power among cases with different numbers of treatments indicated similar behavior of the tests when considering small values of g (3, 6) or larger values (60, 100), which was also observed between balanced and unbalanced scenarios. This latter conclusion corroborates the findings of Takahashi *et al.* (2013) for the HTNT test and of Ahmad (2019) for AT. Thus, the discussed results focused on balanced cases, without loss of generality, and, in general, on situations with g = 6 and 60.

Figure 1 presents the results obtained considering homoscedasticity, low dimensionality, multivariate normal (N) distributions (left) and t_3 (right), g = 6 and 60. For the multivariate normal distribution, it can be observed that for g = 6, Figure 1(a), the T2FO test showed the lowest power, being well below the others, except at ϕ = 8. The HTNT, AT, HTNTB and ATB tests showed very similar power values and, from ϕ = 4, these became asymptotically equivalent, with values close to or equal to 1. For g = 60 with p = 5, Figure 1(b), among those that controlled the FWE, HTNTB showed the highest power, while the others (T2FO, ATB) showed similar powers, with the observation that at ϕ = 2, ATB was superior to T2FO and at ϕ = 4, the opposite occurred. It is worth noting that even though HTNT and AT were significantly liberal in controlling the FWE, they did not show high power as would be expected, especially when compared to HTNTB, which was exact and the most powerful. Considering now p = 250, Figure 1(c), and the tests that controlled the FWE, it can be seen that the increase in p did not affect the performance of HTNTB and ATB, which remained the most powerful. However, T2FO showed negligible power values, being highly influenced by the increase in the number of variables p.

The power values obtained in this study for T2FO, in the homoscedastic scenario of low dimensionality and multivariate normal distribution, are comparable to those reported by Santos & Ferreira (2012) for the multivariate multiple comparisons test using bootstrap, based on Hotelling's T^2 test proposed by the authors. The proximity between the results occurs when the configurations analyzed in both studies are similar.

For the multivariate t_3 distribution ((Figure 1 (d), (e), (f)), in general, the behavior of the tests was similar to that observed for the normal and uniform (0, 1) multivariate distributions, that is, T2FO remained the least powerful and the others had similar power values. It is worth noting that, unlike what was expected, the significantly liberal tests did not show high power values, especially when compared to the power of ATB, which was exact in controlling the FWE in all configurations.

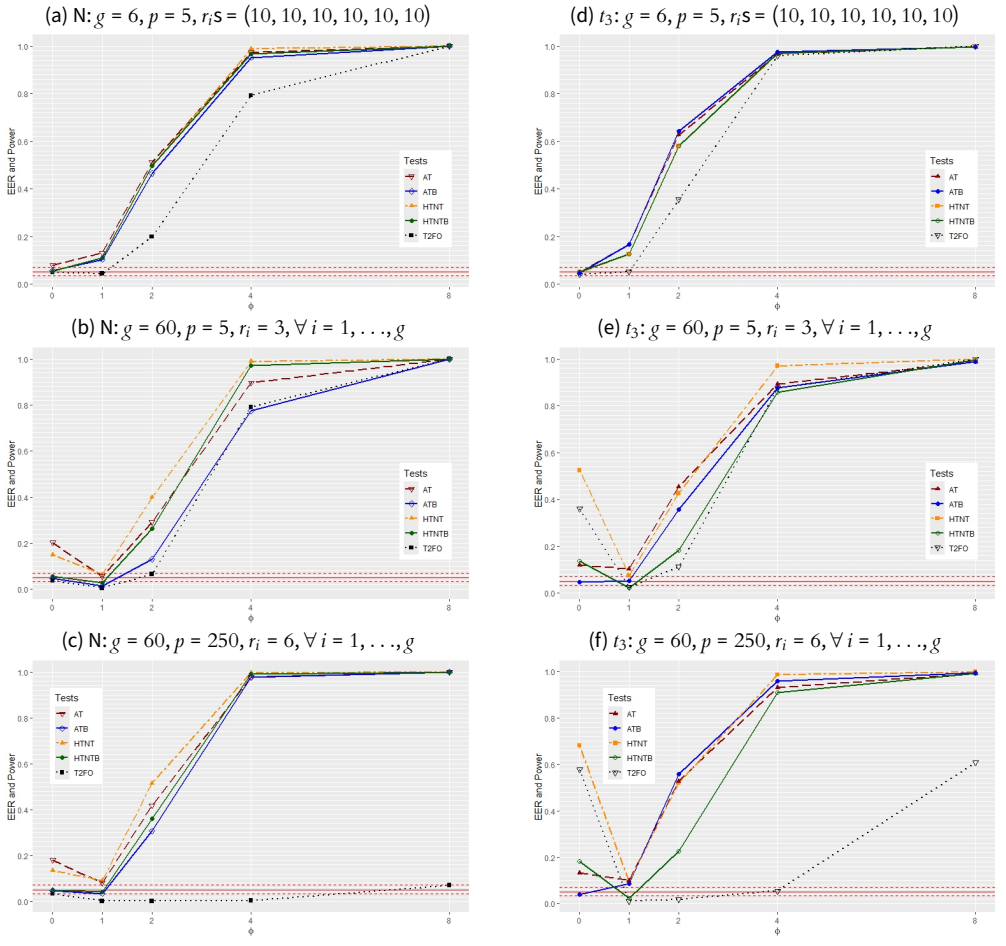

Figure 1. Power and FWE of T2FO, HTNT, AT, HTNTB and ATB tests, considering homoscedasticity, low dimensionality, multivariate normal (N) and t_3 distributions, number of treatments $g = 6$ and 60, number of variables $p \in \{5, 250\}$, sample sizes $(r_1, \ldots, r_g)$, nominal significance level $\alpha = 0.05$ and number of standard errors $\phi \in \{0, 1, 2, 4, 8\}$.

As with FWE control, the power of HTNT, AT, HTNTB and ATB tests in high dimensionality was generally similar to that observed in low dimensionality cases, which corroborates the results obtained by Nascimento *et al.* (2025) in their simulations for these same tests considering comparisons between two mean vectors. Figure 2 presents the results obtained for the multivariate t_3 distribution, $g = 60$, $p = 250$. It can be noted that among those that controlled the FWE, ATB was slightly more powerful than AT, even though it was slightly conservative in controlling the FWE, while AT was exact, that is, in this case, it was expected that AT would show greater power, although not significantly greater, than ATB, which did not occur. This fact reinforces the good performance of ATB in homoscedastic cases.

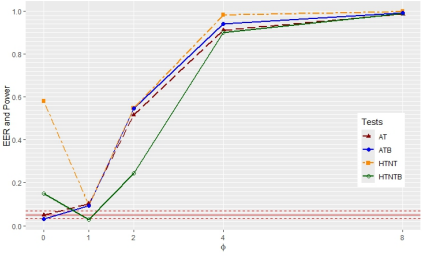

Figure 2. Power and FWE of HTNT, AT, HTNTB and ATB tests, considering homoscedasticity, high dimensionality, multivariate t_3, number of treatments $g = 60$, number of variables $p = 250$, sample sizes $r_i = 3, \forall\, i = 1, \ldots, g$, nominal significance level $\alpha = 0.05$ and number of standard errors $\phi \in \{0, 1, 2, 4, 8\}$.

3.2 Heteroscedastic cases

3.2.1 Type I error rates per experiment

Table 2 presents the results obtained for FWE, under heteroscedasticity and low dimensionality, for multivariate normal and t_3 distributions. Considering the multivariate normal distribution, it can be noted that for $g = 3$ and 6, T2FO showed the best performance, as it was the only exact one in all configurations, followed by AT, which was slightly liberal only in $g = 6$, $p = 5$, first sextuple, and exact in the other configurations, which corroborates the results obtained by Ahmad (2019) for AT. The other tests showed similar behaviors, HTNT, HTNTB and ATB went from exact to significantly liberal as p increased, and this occurred both in $g = 3$ and $g = 6$. Takahashi *et al.* (2013) had already observed, in their simulations for homoscedastic cases, that HTNT was sensitive to large values of p (in heavy-tailed distributions). It can then be noted that this behavior was maintained in the heteroscedastic cases presented in Table 2.

For $g = 60$ and 100, it can be observed in Table 2 that all tests were significantly liberal in $g = 100$, $p = 250$ and, in particular, HTNT showed this behavior in all configurations. For the others, T2FO showed the best performance, as it oscillated between exact and slightly conservative, while TA, HTNTB and ATB were significantly liberal in most configurations, with the observation that, like T2FO, HTNTB and ATB tests were exact in $g = 100$, $p = 5$ and AT was exact in $g = 60$, $p = 250$.

Considering now the multivariate t_3 distribution, it can be verified in $g = 3$ and 6 that T2FO continued to control the FWE exactly in all configurations. The TA test went from exact to conservative (significantly or not) as p increased, a behavior that was not observed by Ahmad (2019) in his simulations. Finally, the other tests showed behaviors very close to what was verified for the normal and uniform (0, 1) multivariate distributions.

In $g = 60$ and 100, T2FO, HTNT and HTNTB were significantly liberal in all configurations. The TA and ATB tests showed opposite behaviors: for a small value of p (5), AT was significantly liberal, while ATB was exact and, for large values of p (250), AT was significantly conservative ($g = 60$) and exact ($g = 100$), while ATB was significantly liberal.

Table 2. Type I error rates per experiment of T2FO, T2SO, HTNT, AT, HTNTB and ATB tests, considering covariance matrix structures (SC, AR), number of treatments (g), sample sizes ($r_1, \ldots, r_g$), number of variables (p), multivariate normal and t_3 distributions, nominal significance level $\alpha = 0.05$, under H_0, heteroscedastic and low dimensionality cases

Low dimensionality, $\boldsymbol{\Sigma}_i$: SC(0.5), AR(1, 0.5), AR(1, 0.7), $\alpha = 0.05$							
g	p	$(r_1, \ldots, r_g)$	T2FO	HTNT	AT	HTNTB	ATB
Multivariate normal							
3	5	(10, 15, 20)	0.0625	0.0385	0.0645	0.0515	0.0540
	5	(50, 75, 100)	0.0525	0.0240^{-}	0.0470	0.0480	0.0445
	50	(50, 75, 100)	0.0615	0.0540	0.0510	0.0815^{+}	0.0740^{+}
	100	(50, 75, 100)	0.0650	0.0730^{+}	0.0500	0.0935^{++}	0.0875^{+}
6	5	(10, 10, 10, 20, 20, 20)	0.0395	0.0545	0.0735^{+}	0.0580	0.0605
	5	(30, 40, 50, 60, 70, 80)	0.0415	0.0440	0.0540	0.0495	0.0500
	50	(10, 10, 10, 20, 20, 20)	0.0355	0.0580	0.0405	0.0660	0.0565
	50	(30, 40, 50, 60, 70, 80)	0.0490	0.0825^{+}	0.0700	0.0955^{++}	0.0960^{++}
	100	(30, 40, 50, 60, 70, 80)	0.0665	0.0865^{+}	0.0610	0.1005^{++}	0.1030^{++}
	300	(30, 40, 50, 60, 70, 80)	0.0410	0.1130^{++}	0.0515	0.1300^{++}	0.1240^{++}
60	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.0235^{-}	0.1530^{++}	0.2110^{++}	0.0725^{+}	0.0750^{+}
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.0385	0.2785^{++}	0.0405	0.1095^{++}	0.1605^{++}
100	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.0655	0.1655^{++}	0.2025^{++}	0.0540	0.0540
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.0990^{++}	0.5560^{++}	0.1105^{++}	0.3080^{++}	0.4325^{++}
Multivariate t_3							
3	5	(10, 15, 20)	0.0460	0.0240^{-}	0.0365	0.0395	0.0355
	5	(50, 75, 100)	0.0455	0.0240^{-}	0.0405	0.0430	0.0440
	50	(50, 75, 100)	0.0500	0.0480	0.0380	0.0785^{+}	0.0640
	100	(50, 75, 100)	0.0515	0.0510	0.0305^{-}	0.0790^{+}	0.0770^{+}
6	5	(10, 10, 10, 20, 20, 20)	0.0505	0.0580	0.0530	0.0560	0.0545
	5	(30, 40, 50, 60, 70, 80)	0.0410	0.0475	0.0410	0.0540	0.0470
	50	(10, 10, 10, 20, 20, 20)	0.0510	0.0765^{+}	0.0135^{--}	0.0715^{+}	0.0465
	50	(30, 40, 50, 60, 70, 80)	0.0595	0.0645	0.0275^{-}	0.0720^{+}	0.0630
	100	(30, 40, 50, 60, 70, 80)	0.0660	0.0850^{+}	0.0285^{-}	0.0845^{+}	0.0865^{+}
	300	(30, 40, 50, 60, 70, 80)	0.0500	0.0875^{+}	0.0230^{-}	0.0885^{+}	0.1055^{++}
60	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.3475^{++}	0.5380^{++}	0.1150^{++}	0.1445^{++}	0.0545
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.2740^{++}	0.8720^{++}	0.0095^{--}	0.1735^{++}	0.1310^{++}
100	5	$r_i = 3, \forall\, i = 1, \ldots, g$	0.5920^{++}	0.7150^{++}	0.1140^{++}	0.1740^{++}	0.0400
	250	$r_i = 6, \forall\, i = 1, \ldots, g$	0.8245^{++}	0.9560^{++}	0.0340	0.2780^{++}	0.3950^{++}

Considering now the high dimensionality cases, Table 3 presents the results obtained for FWE. Starting with the multivariate normal distribution, it can be noted that in $g = 100$, all tests were significantly liberal. In the other configurations, the test with the best performance in controlling FWE was AT, as it was the only exact one in most simulated situations, having oscillated between exact and slightly conservative in $g = 3$ and 6 and, in $g = 60$, it was significantly conservative. In this last case, it is expected that TA would show low power.

The results obtained, in general, are similar to those presented by Ahmad (ibid.) for TA in controlling the type I error rate, since, as observed in Table 3, AT controlled the FWE. It is worth noting that the author did half the number of Monte Carlo simulations done in this work to evaluate the tests and, therefore, this may be a justification for the small differences in results found between the two works.

Still considering Table 3, it is observed that the other tests were significantly liberal in most configurations. In g = 6, the HTNT test showed an increase in FWE as p increased, going from exact in p = 100 to significantly liberal in p = 500, which corroborates the conclusions of Takahashi *et al.* (2013) regarding the sensitivity of the test to large values of p, remembering that the largest value of p used by the authors was 200 and that this conclusion was for heavy-tailed distributions in homoscedastic situations.

For the case of high dimensionality and multivariate t_3 distribution, it can be observed that, in general, the tests showed better performance in controlling FWE compared to the multivariate normal distribution, which was also observed by Nascimento *et al.* (2025) in their simulations for comparisons between two mean vectors performed by the authors. In g = 3 and 6, HTNT, HTNTB and ATB were exact in most configurations, while they had been significantly liberal in the multivariate normal distribution. The TA test became significantly conservative and, therefore, low power is expected. For g = 60 and 100, HTNT and HTNTB continued to be significantly liberal, AT became significantly conservative in g = 100 and, ATB, exact in g = 60.

Table 3. Type I error rates per experiment of HTNT, AT, HTNTB and ATB tests, considering covariance matrix structures (SC, AR), number of treatments (g), sample sizes ($r_1, \ldots, r_g$), number of variables (p), multivariate normal and t_3 distributions, nominal significance level α = 0.05, under H_0, heteroscedastic and high dimensionality cases

High dimensionality, $\boldsymbol{\Sigma}_i$: SC(0.5), AR(1, 0.5), AR(1, 0.7), α = 0.05						
g	p	$(r_1, \ldots, r_g)$	HTNT	AT	HTNTB	ATB
			Multivariate normal			
3	50	(10, 15, 20)	0.0840^{+}	0.0540	0.0925^{++}	0.0695
	100	(10, 15, 20)	0.1060^{++}	0.0545	0.1165^{++}	0.0905^{+}
	300	(10, 15, 20)	0.1165^{++}	0.0325^{-}	0.1365^{++}	0.0965^{++}
	300	(50, 75, 100)	0.1010^{++}	0.0490	0.1315^{++}	0.1245^{++}
	500	(10, 15, 20)	0.1395^{++}	0.0460	0.1670^{++}	0.1155^{++}
	500	(50, 75, 100)	0.0900^{+}	0.0320^{-}	0.1170^{++}	0.1075^{++}
6	100	(10, 10, 10, 20, 20, 20)	0.0690	0.0415	0.0835^{+}	0.0710^{+}
	300	(10, 10, 10, 20, 20, 20)	0.0780^{+}	0.0440	0.0945^{++}	0.0850^{+}
	500	(10, 10, 10, 20, 20, 20)	0.0725^{+}	0.0410	0.0870^{+}	0.0765^{+}
	500	(30, 40, 50, 60, 70, 80)	0.1350^{++}	0.0555	0.1490^{++}	0.1490^{++}
60	250	r_i = 3, $\forall\, i$ = 1, . . ., g	0.3040^{++}	0.0045^{--}	0.1200^{++}	0.1325^{++}
100	250	r_i = 3, $\forall\, i$ = 1, . . ., g	0.5770^{++}	0.0940^{++}	0.2900^{++}	0.3975^{++}
			Multivariate t_3			
3	50	(10, 15, 20)	0.0410	0.0235^{-}	0.0585	0.0415
	100	(10, 15, 20)	0.0380	0.0210^{-}	0.0570	0.0425
	300	(10, 15, 20)	0.0455	0.0210^{-}	0.0660	0.0495
	300	(50, 75, 100)	0.0500	0.0215^{-}	0.0775^{+}	0.0745^{+}
	500	(10, 15, 20)	0.0430	0.0140^{--}	0.0610	0.0490
	500	(50, 75, 100)	0.0600	0.0260^{-}	0.0920^{++}	0.0845^{+}
6	100	(10, 10, 10, 20, 20, 20)	0.0560	0.0110^{--}	0.0505	0.0400
	300	(10, 10, 10, 20, 20, 20)	0.0810^{+}	0.0140^{--}	0.0775^{+}	0.0460
	500	(10, 10, 10, 20, 20, 20)	0.0675	0.0070^{--}	0.0630	0.0455
	500	(30, 40, 50, 60, 70, 80)	0.0935^{++}	0.0260^{-}	0.0925^{++}	0.0990^{++}
60	250	r_i = 3, $\forall\, i$ = 1, . . ., g	0.9110^{++}	0.0030^{--}	0.3085^{++}	0.0580
100	250	r_i = 3, $\forall\, i$ = 1, . . ., g	0.9805^{++}	0.0190^{-}	0.4755^{++}	0.2570^{++}

Finally, recalling that a large number of treatments and, consequently, the increase in the num-

ber of comparisons, should potentially make the asymptotic tests based on Bonferroni inequality significantly conservative (Nishiyama *et al.,* 2014; Seo *et al.,* 1994), it is worth noting that this effect was not observed. The exception occurred only for the TA test, in scenarios of heteroscedasticity and high dimensionality, with $g = 60$.

3.2.2 Power

Initially, power will be evaluated under low dimensionality. In cases of $g = 3$ and 6, similar behavior was observed for the tests in $p = 50$ and 100, as well as in smaller or larger sample sizes, a result that corroborates those found by Ahmad (2019) for the AT test. Figure 3 presents the results obtained considering the multivariate normal distribution, $g = 3$ with $p = 5$ and 100, $g = 6$ with $p = 5$ and 300 and sample sizes $(r_1, \ldots, r_g) = (50, 75, 100)$ or $(30, 40, 50, 60, 70, 80)$. It can be noted that:

1. although T2FO showed the best FWE control, it was the test with the lowest power values;
2. T2FO performance was negatively influenced by the increase in dimension. As p increased, the power of this test decreased. This becomes evident when comparing its performance, for example, in $g = 6$, $p = 5$, Figure 3(c), and $p = 300$, Figure 3(d);
3. the performances of the asymptotic tests HTNT and AT, as well as their respective bootstrap versions, HTNTB and ATB, proposed in this study, were quite similar;
4. TA, which remained exact in controlling FWE even with the increase in p, showed good power performance, with its values increasing as the dimension increased. It can be observed, for example, the power of AT in $\phi = 1$ at different values of p. It is worth noting that, in particular, Ahmad (ibid.) observed this same behavior for AT, in the simulations he performed, as did Nascimento *et al.* (2025).

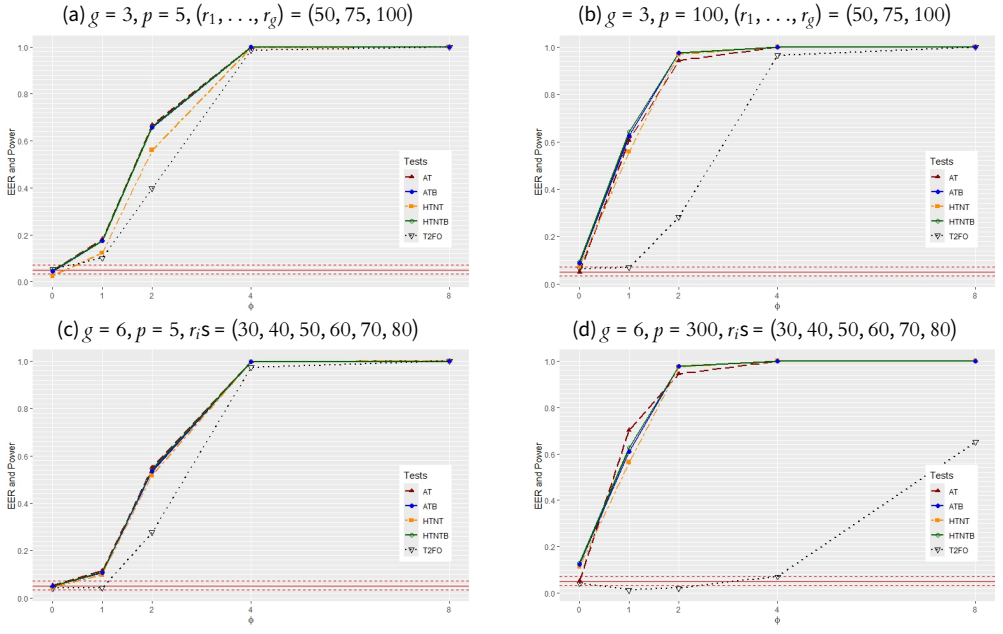

Figure 3. Power and FWE of T2FO, HTNT, AT, HTNTB and ATB tests, considering heteroscedasticity, low dimensionality, multivariate normal distribution, number of treatments $g \in \{3, 6\}$, number of variables $p \in \{5, 100, 300\}$, sample sizes $(r_1, \ldots, r_g)$, nominal significance level $\alpha = 0.05$ and number of standard errors $\phi \in \{0, 1, 2, 4, 8\}$.

Still considering the multivariate normal distribution, the results obtained for power in the case of $g = 60$ are presented in Figure 4. Recalling that all tests were significantly liberal in controlling FWE in $g = 100$, $p = 250$ (Table 3), it can be concluded that these are not satisfactory in this configuration. An alternative would be to lower the nominal significance level to a smaller level, in this case, to $\alpha = 0.01$, so that the FWE is close to 5%, and use the AT test, since it showed the lowest FWE. In particular, HTNT was significantly liberal in all configurations ($g = 60$ and 100), therefore, this is not a test that performs well in these cases and, therefore, its power will not be taken into account in the comparison with the other tests.

In $g = 60$ with $p = 5$, Figure 4(a), T2FO was the test with the lowest power, which was expected, since it was slightly conservative in controlling FWE, with the caveat that, for values of ϕ greater than or equal to 4, it had higher power than ATB, which was slightly liberal. The HTNTB test had the same behavior as ATB in controlling FWE, but showed higher power, being the most recommended in this configuration.

In $g = 60$ with $p = 250$, Figure 4(b), it can be noted how much T2FO is influenced by the increase in dimension, with its power values drastically reduced when compared to the situation where $p = 5$. Even though they were exact in controlling FWE, these did not show plausible power values. For this configuration, HTNTB and ATB were significantly liberal, while AT was exact and showed high power. The performance of the tests for $g = 100$ with $p = 5$ was quite similar to the configuration of $g = 60$ with $p = 5$. Similarly, the power of the tests under the multivariate t_3 distribution was similar to that observed under the normal and uniform (0, 1) multivariate distributions.

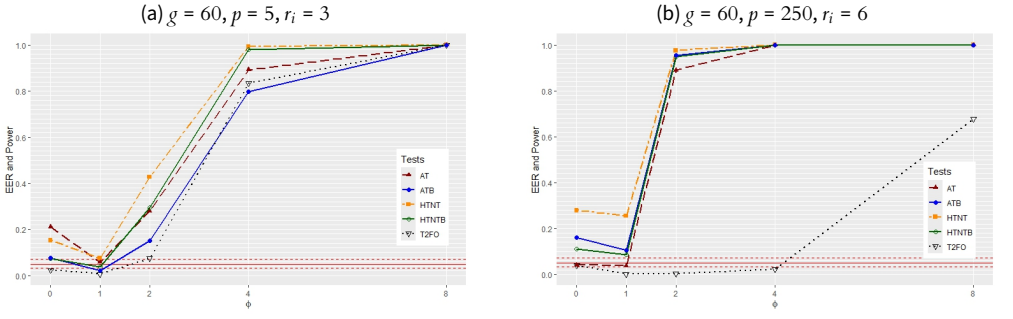

Figure 4. Power and FWE of T2FO, HTNT, AT, HTNTB and ATB tests, considering heteroscedasticity, low dimensionality, multivariate normal distribution, number of treatments $g = 60$, number of variables $p \in \{5, 250\}$, sample sizes r_is, $\forall\, i = 1, \ldots, g$, nominal significance level $\alpha = 0.05$ and number of standard errors $\phi \in \{0, 1, 2, 4, 8\}$.

When analyzing the power of the tests in high dimensionality scenarios for the multivariate normal distribution, it was possible to observe that the behavior of HTNT, AT, HTNTB and ATB remained similar to that observed in low dimensionality contexts. In addition, the tests showed comparable performances in terms of power, without any of them standing out significantly in relation to the others. In this context, the choice of the most appropriate test can be guided by FWE control, given that the differences in power between the evaluated tests were of little relevance. It is only worth noting that, in $g = 60$, it was expected that TA would show low power, for having been significantly conservative in controlling FWE, but this did not occur, which corroborates the good performance of TA, already observed by Ahmad (2019) for smaller values of g.

Considering now the multivariate t_3 distribution and high dimensionality, Figure 5 presents the power results obtained for $g = 3$, $(r_1, \ldots, r_g) = (10, 15, 20)$, $g = 6$, with $(r_1, \ldots, r_g) = (10, 10, 10, 20, 20, 20)$ and, in both cases, $p \in \{100, 300\}$. It is worth recalling that the performance of the tests in controlling FWE was better in the multivariate t_3 than in the multivariate normal distribution. It can be noted, in Figure 5, that, for $g = 3$ and regardless of the value of p, HTNTB and ATB were

the tests with the highest power, with ATB being slightly superior to HTNTB for values of ϕ less than 4 and, from ϕ = 4, these, together with HTNT and AT, became asymptotically equivalent, with power values close to 1. In $g = 6$, ATB was superior to the others, which showed very similar power values.

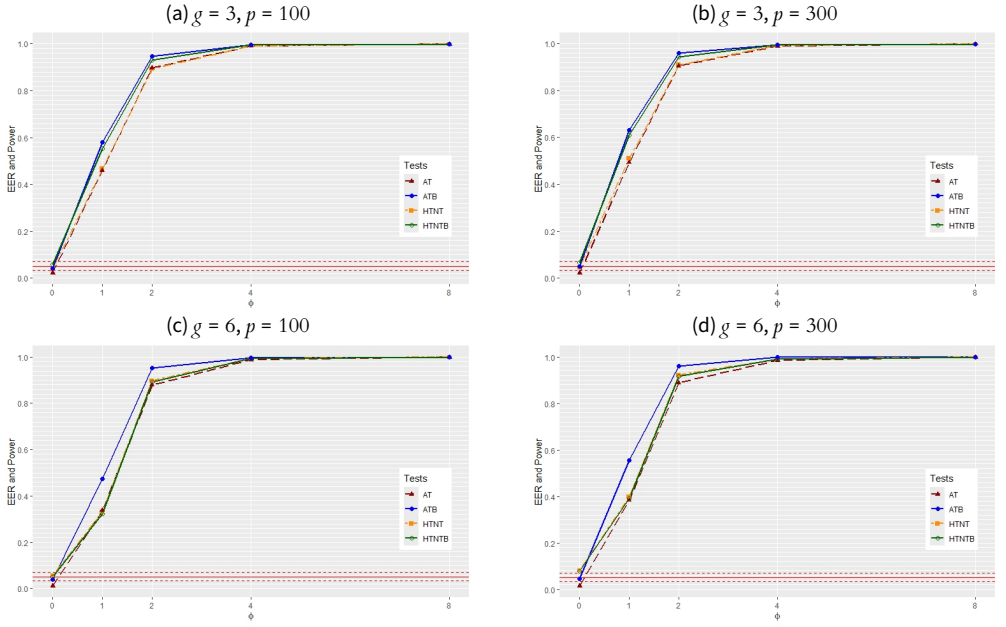

Figure 5. Power and FWE of HTNT, AT, HTNTB and ATB tests, considering heteroscedasticity, high dimensionality, multivariate t_3, number of treatments $g \in \{3, 6\}$, number of variables $p \in \{100, 300\}$, sample sizes $(r_1, \ldots, r_g) = (10, 15, 20)$, nominal significance level $\alpha = 0.05$ and number of standard errors $\phi \in \{0, 1, 2, 4, 8\}$.

For $g = 60$, ATB was the test with the best performance, as can be seen in Figure 6(a), because it controlled FWE exactly and had high power values. TA was significantly conservative and, as expected, showed lower power than ATB. In $g = 100$, Figure 6(b), only AT can be considered for application, since it controlled FWE in a significantly conservative manner and, even so, with the exception of when ϕ = 1, it showed good power performance, which was not expected. HTNT and HTNTB were significantly liberal in $g = 60$ and 100.

3.2.3 General considerations

For the homoscedastic cases, the proposed bootstrap tests, HTNTB and ATB, showed the best performances. The ATB test can be considered superior to HTNTB as it controlled the TEE exactly in practically all configurations, unlike HTNTB, which presented more exceptions. In the cases where ATB was not exact, it was slightly liberal, while HTNTB was significantly liberal in most exceptions, which occurred for a heavier-tailed distribution, the multivariate t_3. It is worth recalling that the good performance of the proposed bootstrap tests, HTNTB and ATB, was maintained in cases with a larger number of treatments. Their respective asymptotic competitors, HTNT and AT, were significantly liberal in practically all configurations where $g = 60$ and 100.

Therefore, for homoscedastic cases, the test with the best performance was ATB, as it controlled the TEE exactly and its power was, in general, close to or superior to the others. Moreover, the ATB test maintained this excellent performance even in high-dimensionality contexts, thus remaining as the most robust test in this scenario as well.

In heteroscedastic cases, the performance of the tests varied considerably according to the adopted configurations. This fluctuation can be explained by the complexity of these situations, which are very comprehensive, as they involve not only the heterogeneity of covariance matrices but also the consideration of non-normal distributions, unbalanced data, and high dimensionality.

In low-dimensionality scenarios with a smaller number of treatments (3 and 6), the tests HTNT, AT, HTNTB, and ATB showed quite similar power performances, with HTNT being slightly less powerful than the others. Regarding TEE control, the TA test showed the best performance, as it achieved more consistent control across different configurations. However, the proposed tests HTNTB and ATB also presented satisfactory results, being surpassed by their respective asymptotic competitors (HTNT and TA) only in situations involving a larger number of variables.

For a higher number of treatments (60 and 100), the HTNTB test showed superior performance in TEE control compared to its asymptotic competitor, HTNT. The ATB test performed better than AT when the number of variables was small ($p = 5$), while for $p = 250$, the opposite result was observed. The percentages presented in Table 4 help visualize these conclusions. Considering the three multivariate distributions and the three nominal significance levels α used in the simulations, heteroscedasticity, low dimensionality, $g = 60$ and 100, this table presents the percentages of simulated configurations in which the tests HTNT, AT, HTNTB, and ATB were classified into two groups: (I) significantly liberal and (II) exact, conservative, or slightly liberal.

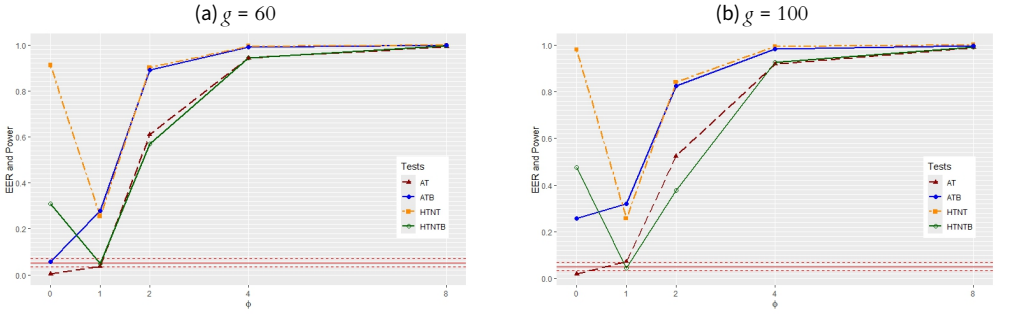

Figure 6. Power and FWE of HTNT, AT, HTNTB and ATB tests, considering heteroscedasticity, high dimensionality, multivariate t_3, number of treatments $g \in \{60, 100\}$, number of variables $p = 250$, sample sizes $r_i = 3$, $\forall\, i = 1, \ldots, g$, nominal significance level $\alpha = 0.05$ and number of standard errors $\phi \in \{0, 1, 2, 4, 8\}$.

Table 4. Percentages of simulated configurations in which the tests HTNT, AT, HTNTB, and ATB were classified into two groups: (I) significantly liberal and (II) exact, conservative, or slightly liberal, considering all distributions, all αs, under heteroscedasticity, low dimensionality, $g = 60$ and 100

p	g	HTNT		HTNTB		AT		ATB	
		I	II	I	II	I	II	I	II
5	60	88.9%	11.1%	22.2%	77.8%	88.9%	11.1%	–	100.0%
	100	88.9%	11.1%	22.2%	77.8%	88.9%	11.1%	11.1%	88.9%
250	60	100.0%	–	55.6%	44.4%	–	100.0%	88.9%	11.1%
	100	100.0%	–	100.0%	–	55.6%	44.4%	100.0%	–

Therefore, in general, the proposed bootstrap tests, HTNTB and ATB, performed relatively better than their respective asymptotic competitors, HTNT and AT, in these situations.

In heteroscedastic cases with high dimensionality, the tests HTNT, AT, HTNTB, and ATB showed quite similar behaviors for the multivariate normal and uniform (0, 1) distributions, with the AT test generally showing the best performance, both in TEE control and power. However, for the

multivariate t_3 distribution, which has heavier tails, the proposed tests, HTNTB and ATB, showed significant improvement in their performances, even surpassing their asymptotic competitors in most configurations, especially regarding power values. The AT test, in turn, was significantly conservative in practically all configurations for this distribution and, consequently, showed lower power than the other tests.

3.2.4 Application to an actual dataset

The multivariate multiple comparison tests analyzed in this work were applied to an actual dataset consisting of $g = 6$ land use systems (LUSs) in the Amazon: primary forest (FP), early-stage regenerating secondary forest (FI), advanced-stage regenerating secondary forest (FA), pasture (P), agroforestry (A), and agriculture, annual and semi-perennial crops (AG), with sample sizes $(r_1, \ldots, r_6) = (r_A, r_{AG}, r_{FA}, r_{FI}, r_P, r_{FP}) = (10, 18, 10, 30, 13, 17)$ and $p = 2$ variables related to soil texture (sand and clay), thus a low-dimensionality context. To identify which of the scenarios analyzed in this study the dataset belongs to, Royston's multivariate Shapiro-Wilk test (Royston, 1983) and Bartlett's test using F approximation (Box, 1949) were applied, and based on the results, we have non-normal multivariate distribution and heteroscedasticity.

The objective, when applying the tests studied and evaluated in this work, is to verify whether the mean vector of the control treatment FP is statistically equal to the other mean vectors. Before applying the tests, the nominal significance level was set at 5%. The p-values obtained for each pair of comparisons with the control FP, for each test, are presented in Table 5.

Table 5. p-values for each pair of comparisons between mean vectors of LUSs with the control treatment FP, for each test

Comparison pair	T2FO	HTNT	AT	HTNTB	ATB
A - FP	0.4063	0.5744	0.4346	0.9820	0.9005
AG - FP	0.0380	0.0550	0.0048*	0.1524	0.0375**
FA - FP	0.4613	0.6299	0.5088	0.9930	0.9475
FI - FP	0.0249	0.0016*	0.0034*	0.0680	0.0345**
P - FP	0.0008*	0.0000*	0.0000*	0.0075**	0.0025**

*: significantly different (p-value < α/m) from the mean vector of control FP.

**: significantly different (p-value < α) from the mean vector of control FP.

It is observed that the T2FO and HTNTB tests identified the same set of statistically different pairs, indicating exclusively that the mean vector corresponding to the pasture LUS (P) differs significantly from that of the primary forest LUS (FP). On the other hand, the HTNT test, in addition to pointing out this distinction, also showed statistically significant differences between the mean vectors of the pair FI - FP. Finally, the AT and ATB tests presented equal results, identifying three pairs with statistically relevant differences: AG-FP, FI-FP, and P-FP. Note that the mean vector of the pasture LUS (P) was identified as statistically different from the mean vector of the primary forest LUS (FP) by all tests, a result that corroborates those found by Nóbrega (2006) and (Santos & Ferreira (2012).

4. Conclusions

The proposed bootstrap tests, HTNTB and ATB, showed the best performances in homoscedastic scenarios, with ATB surpassing HTNTB. The ATB test showed more effective TEE control, in addition to having higher power, regardless of whether the distribution was normal or not, and in both low and high-dimensionality contexts. Thus, the ATB test stands out as the most recommended option in these situations.

In contrast, it was not possible to identify a test that stood out as the most effective for heteroscedastic cases, which can be explained by the complexity and comprehensiveness of these situations. However, it is relevant to emphasize that, in several circumstances, the proposed bootstrap tests performed better than their respective existing asymptotic versions. In general, there are no tests that are uniformly more powerful and that simultaneously guarantee type I error rate control in general multiple comparison scenarios.

Acknowledgments

The authors acknowledge the financial support of the CAPES and CNPq agencies and the support of IFMG Formiga *campus*. Also, authors thank reviewers and editors for their valuable comments.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

Not Applicable.

Author Contributions

Conceptualization: BRAZ, L. H. C.; NASCIMENTO, M. C.; FERREIRA, D. F. **Data curation:** BRAZ, L. H. C.; NASCIMENTO, M. C. **Formal analysis:** BRAZ, L. H. C.; FERREIRA, D. F. **Investigation:** BRAZ, L. H. C.; NASCIMENTO, M. C.; FERREIRA, D. F. **Methodology:** BRAZ, L. H. C.; NASCIMENTO, M. C.; FERREIRA, D. F. **Software:** BRAZ, L. H. C.; FERREIRA, D. F. **Resources:** FERREIRA, D. F. **Supervision:** FERREIRA, D. F. **Validation:** FERREIRA, D. F. **Visualization:** BRAZ, L. H. C.; NASCIMENTO, M. C.; FERREIRA, D. F. **Writing original draft:** BRAZ, L. H. C.; NASCIMENTO, M. C. **Writing-review and editing:** BRAZ, L. H. C.; NASCIMENTO, M. C.; FERREIRA, D. F.

References

1. Ahmad, M. R. A unified approach to testing mean vectors with large dimensions. *AStA Advances in Statistical Analysis* **103,** 593–618 (2018).
2. Ahmad, M. R. Multiple comparisons of mean vectors with large dimension under general conditions. *Journal of Statistical Computation and Simulation* **89,** 1044–1059 (2019).
3. Box, G. P. A general distribution theory for a class of likelihood criteria. *Biometrika* **36,** 317–346 (1949).
4. Bretz, F, Hothorn, T & Westfall, P. *Multiple comparisons using R* 182 (Chapman and Hall, USA, 2011).
5. Dean, A & Voss, D. *Design and analysis of experiments* 740 (Springer, New Jersey, 1999).
6. Dean, A, Voss, D & Draguljić, D. *Design and analysis of experiments* 840 (Springer, New York, 2016).
7. Dempster, A. P. A high dimensional two sample significance test. *The Annals of Mathematical Statistics* **29,** 995–1010 (1958).
8. Dempster, A. P. A significance test for the separation of two highly multivariate small samples. *Biometrics* **16,** 41–50 (1960).

9. Gentle, J. E. *Random number generation and Monte Carlo methods* 2nd ed., 381 (Springer, New York, 2003).
10. Hinkelman, K & Kempthorne, O. *Design and analysis of experiments* 2nd ed., 632 (John Wiley and Sons, New Jersey, 2008).
11. Hochberg, Y & Tamhane, A. C. *Multiple comparisons procedures* 450 (John Wiley and Sons, Canadian, 1987).
12. Hsu, J. C. *Multiple comparisons - theory and methods* 277 (Chapman and Hall, USA, 1999).
13. Hyodo, M, Takahashi, S & Nishiyama, T. Multiple comparisons among mean vectors when the dimension is larger than the total sample size. *Communications in Statistics - Simulation and Computation* **43,** 2283–2306 (2014).
14. Kakizawa, Y. Multiple comparisons of several heteroscedastic multivariate populations. *Statistics and Probability Letters* **78,** 1328–1338 (2008).
15. Kakizawa, Y. Multiple comparisons of several homoscedastic multivariate populations. *Annals of the Institute of Statistical Mathematics* **61,** 1–26 (2009).
16. Machado, A. A., Silva, J. G. C., Demétrio, C. G. & Ferreira, D. F. in *Reunião anual da região brasileira da sociedade internacional de biometria* 50, 290 (Simpósio de estatística aplicada a experimentação agronômica, Londrina, 2005).
17. Manly, B. F. J. *Randomization, bootstrap and Monte Carlo methods in biology* 2nd ed., 356 (University of Otag, New Zealand, 1997).
18. Nascimento, M. C., Braz, L. H. C. & Ferreira, D. F. Proposition of Bootstrap Tests for Comparisons Between Two Independent Mean Vectors in High Dimensionality. *Brazilian Journal of Biometrics* **43,** 1–21. https://doi.org/10.28951/bjb.v43i3.772 (2025).
19. Nishiyama, T, Hyodo, M & Seo, T. Recent developments of multivariate multiple comparisons among mean vectors. *SUT Journal of Mathematics* **50,** 247–270 (2014).
20. Nóbrega, R. S. A. *Efeito de sistemas de uso da terra na Amazônia sobre atributos do solo, ocorrência, eficiência e diversidade de bactérias que nodulam caupi [Vigna unguiculata (L.) Walp]* Doutorado (Universidade Federal de Lavras, Lavras, MG, 2006).
21. Oliveira, I. R. C. & Ferreira, D. F. Multivariate extension of chi-squared univariate normality test. *Journal of Statistical Computation and Simulation* **80,** 513–526 (2010).
22. Royston, J. P. Some techniques for assessing multivariate normality based on the Shapiro-Wilk *W*. *Applied Statistics - Journal of the Royal Statistical Society - Series C* **32,** 121–133 (1983).
23. Santos, E. N. F. & Ferreira, D. F. Multivariate multiple comparisons by bootstrap and permutation tests. *Revista Brasileira de Biometria* **30,** 381–400 (2012).
24. Seo, T, Mano, S & Fujikoshi, Y. A generalized Tukey conjecture for multiple comparisons among mean vectors. *Journal of the American Statistical Association* **89,** 676–679 (1994).
25. Seo, T & Nishiyama, T. On the conservative simultaneous confidence procedures for multiple comparisons among mean vectors. *Journal of Statistical Planning and Inference* **138,** 3448–3456 (2008).
26. Staffa, S. J. & Zurakowski, D. Strategies in adjusting for multiple comparisons. A primer for pediatric surgeons. *Journal of Pediatric Surgery* **55,** 1699–1705 (2020).
27. Takahashi, S, Masashi, H., Takahiro, N. & Pavlenko, T. Multiple comparisons procedures for high-dimensional data and their robustness under non-normality. *Journal of the Japanese Society of Computational Statistics* **26,** 71–82 (2013).

28. Westfall, P. H. On using the bootstrap for multiple comparisons. *Journal of Biopharmaceutical Statistics* **21,** 1187–1205 (2011).

29. Westfall, P. H. & Young, S. S. *p* Value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association* **84,** 780–786 (1989).