






## ARTICLE

# Proposition of Bootstrap Tests for Comparisons Between Two Independent Mean Vectors in High Dimensionality

 Miguel Carvalho Nascimento,<sup>1</sup>  Lúcia Helena Costa Braz,<sup>\*,1</sup> and  Daniel Furtado Ferreira<sup>2</sup>

<sup>1</sup>Postgraduate Program in Statistics and Agricultural Experimentation, Federal University of Lavras, Lavras, MG, Brazil.

<sup>2</sup>Department of Statistics, Federal University of Lavras, Lavras, MG, Brazil.

\*Corresponding author. Email: lucia.helena@ifmg.edu.br

(Received: September 26, 2024; Revised: January 29, 2025; Accepted: February 11, 2025; Published: May 21, 2025)

## Abstract

Inference regarding the comparison of mean vectors between two independent populations is of great interest in applied fields, especially in scenarios where high-dimensional data analyses are common. In low-dimensional cases with the multivariate Behrens-Fisher problem, there are numerous solutions, but most test statistics have asymptotic distributions. In multivariate procedures, a problem arises when the number of variables,  $p$ , is greater than or equal to the sample size,  $n$ . In this case, it is not possible to use the few existing methods, as they rely on the inverse of the sample covariance matrix, which cannot be obtained in this situation ( $p \geq n$ ) since the covariance matrix is singular. In most cases, asymptotic tests are very liberal, particularly in small samples and specifically in multivariate cases when the dimensionality is high. The bootstrap method is one of the main computationally intensive methods, where its key advantage is that it does not require knowledge of the population probability distribution. Additionally, when the conditions assumed for the application of a test are violated, the bootstrap makes the problem extremely simple to address. Based on this, the present study aimed to propose multivariate comparison tests between two independent mean vectors: the Ahmad Bootstrap Test (ABT) and the Hyodo, Takahashi, and Nishiyama Bootstrap Test (HTNBT), in high-dimensional settings, for balanced or unbalanced, non-normal and normal data, under the multivariate Behrens-Fisher problem. The performance of these tests was evaluated and compared with tests indicated by the literature, namely Hotelling's  $T^2$ , the modified Nel and Merwe (MNV) test proposed by Krishnamoorthy and Yu, the test proposed by Ahmad (AT), and the test proposed by Hyodo, Takahashi, and Nishiyama (HTNT), using Monte Carlo simulation. Power and Type I error rate were considered as evaluation measures. Comparisons were conducted in various scenarios, such as cases of homoscedasticity and heteroscedasticity of covariance matrices, in low and high dimensionality for multivariate normal,  $t$  with 3 degrees of freedom, and uniform (0, 1) distributions. In other words, scenarios in which the conditions assumed for the application of most tests are violated. The results showed that the ATB test was generally robust and consistent compared to its competitors in most

evaluated situations, while the HTNTB test was strongly conservative and had low power.

**Keywords:** Non-parametric Bootstrap; Test Evaluation; Monte Carlo Simulation.

## 1. Introduction

Inference concerning the comparisons of mean vectors between two independent populations is of significant interest in applied fields, especially in scenarios involving high-dimensional data analysis, where the number of variables is greater than or equal to the number of observations. In cases of low-dimensionality with the multivariate Behrens-Fisher problem, characterized by heterogeneity between the covariance matrices of two multivariate normal populations (Ferreira, 2018), the statistics of traditionally applied tests only have asymptotic distributions. Numerous solutions exist for this problem (Bennett, 1951; James, 1954; Yao, 1965; Johansen, 1980; Nel & Van der Merwe, 1986; Krishnamoorthy & Yu, 2004). Among them, Krishnamoorthy & Yu (2004) recommend using the test statistic proposed by Nel & Van der Merwe (1986) with a modification suggested by them.

In multivariate procedures, a problem arises when the number of variables  $p$  exceeds the sample size  $n$ . In such cases, few existing methods can be utilized because they rely on the inverse of the sample covariance matrix, which cannot be obtained when  $(p \geq n)$ , as the covariance matrix is singular. To address this issue, Hyodo *et al.* (2014) proposed using Dempster's trace criterion (Dempster, 1958; Dempster, 1960) for comparisons between mean vectors. Additionally, problems arise when multivariate data do not originate from multivariate normal distributions and when, analogous to the univariate case, the covariance matrices of treatments (or populations) are not homogeneous. Such cases, under generally unfavorable conditions, were addressed by Ahmad (2018), who presented asymptotic solutions based on chi-squared and standard normal distributions.

In most cases, asymptotic tests are overly liberal, meaning they exhibit Type I error rates that are considerably higher than the nominal significance levels adopted, especially in small samples and specifically in the multivariate case when dimensionality is high. In many instances, these tests are not efficient in controlling Type I error, as observed by Silva *et al.* (2008), who concluded that the bootstrap test studied was superior to the asymptotic tests, being considered robust with respect to the assumptions made for the test while controlling Type I error in a conservative manner.

The bootstrap technique is one of the main computationally intensive methods that, among its major advantages, does not require knowledge of the population probability distribution. Moreover, when the conditions assumed for the application of a test are violated, the bootstrap method makes addressing the problem extremely straightforward.

Therefore, the general objective of this work is to propose non-parametric bootstrap tests for comparisons between two independent mean vectors in high dimensionality, considering balanced or unbalanced, non-normal and normal data under the multivariate Behrens-Fisher problem. The specific objectives are to evaluate the performance of the proposed tests and compare it with the performance of tests present in the literature, including Hotelling's  $T^2$  test and the tests proposed by Krishnamoorthy & Yu (2004), Ahmad (2018), and Hyodo *et al.* (2014), considering various scenarios such as cases of homoscedasticity and heteroscedasticity between covariance matrices, in both low and high dimensionality for multivariate normal,  $t$ -distribution with 3 degrees of freedom, and uniform (0, 1) distributions.

## 2. Methods

Without loss of generality, consider the  $p$ -dimensional random vectors  $\mathbf{X}_{ik} = [X_{ik1}, \dots, X_{ikp}]^\top \sim \mathfrak{F}_i$ ,  $i = 1, 2$ ,  $k = 1, 2, \dots, n_i$  where  $\mathbf{X}_{ik}$  has a  $p$ -dimensional mean vector  $E(\mathbf{X}_{ik}) = \boldsymbol{\mu}_i$  and a positive definite  $p \times p$  symmetric covariance matrix  $\text{Cov}(\mathbf{X}_{ik}) = \boldsymbol{\Sigma}_i$ , with  $\mathfrak{F}_i$  representing the distribution

family for the  $i$ -th treatment or population. Cases where the distribution family  $\mathfrak{F}_i$  corresponds to the multivariate normal distribution were considered, as well as the more general case where  $\mathfrak{F}_i$  is a  $p$ -variate distribution that is not necessarily normal, with potentially unequal (heterogeneous) covariance matrices  $\Sigma_i$  and unbalanced data, where  $n_i$  are different ( $n_i$  is the sample size for the  $i$ -th treatment or population). The complete sample is the combination of the samples from the two populations into a single sample, with  $n$  being the total size of the combined sample, where  $n = \sum_{i=1}^2 n_i$ . Typical cases were considered with  $p < n-2$ , as well as cases where  $p \geq n-2$ , indicating high dimensionality. The inference of interest was on the  $p$ -dimensional parameter vectors  $\delta = \mu_2 - \mu_1$ .

Unbiased estimators for the mean vector  $\mu_i$  and the covariance matrix  $\Sigma_i$  of the  $i$ -th treatment (or population) are given by

$$\bar{X}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik} \quad \text{and} \quad S_i = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (X_{ik} - \bar{X}_i)(X_{ik} - \bar{X}_i)^\top, \quad i = 1, 2. \quad (1)$$

When the covariance matrices of the populations are equal, that is, in the homogeneous case where  $\Sigma_1 = \Sigma_2 = \Sigma$ , the estimator for the common covariance matrix  $\Sigma$  is given by

$$S_p = \frac{1}{n-2} \sum_{i=1}^2 (n_i - 1) S_i, \quad (2)$$

associated with  $\nu = n - 2$  degrees of freedom, where  $n = n_1 + n_2$ .

The general case of interest refers to linear combinations  $a^\top \delta$ , for a known non-zero vector  $a \in \mathbb{R}^p$ , where  $\delta = \mu_2 - \mu_1$  is the vector of differences between the two mean vectors. A case of particular interest is  $\delta$  *per se*, which includes all possible differences. The estimator of  $\delta$  is  $\hat{\delta} = \bar{X}_2 - \bar{X}_1$ .

The null hypothesis tested is  $H_0 : \delta = 0$ , with the alternative hypothesis given by  $H_1 : \delta \neq 0$ . Hotelling's  $T^2$  test was applied and served as a reference for comparisons with the other proposed tests. This test and the multivariate normality test (MNV) were applied only when  $p < n - 2$ . The asymptotic tests of Hyodo *et al.* (2014) and Ahmad (2018) were applied in all cases, as well as the two proposed tests. The following sections will describe the tests assuming the sample structures described above.

## 2.1 Tests

In this subsection, computations for the Hotelling's  $T^2$  test (HT), Krishnamoorthy & Yu (2004) (MNV), Ahmad's test (AT), Hyodo *et al.* (2014) (HTNT), and the proposed bootstrap tests are presented.

### 2.1.1 Hotelling's $T^2$ Test (HT)

The statistic  $T_c^2$ , for the HT test was computed as

$$T_c^2 = \frac{n_1 n_2}{n} (\bar{X}_1 - \bar{X}_2 - \delta_0)^\top S_p^{-1} (\bar{X}_1 - \bar{X}_2 - \delta_0). \quad (3)$$

Under  $H_0$ , the distribution of  $T_c^2$  is exact and proportional to a central  $F$  distribution, assuming multivariate normal samples and homoscedasticity, given by

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

The null hypothesis  $H_0 : \mu_1 = \mu_2 = \mu$  was rejected when the computed value of the statistic in (3) exceeded the critical value from the distribution (4) given by  $[(n_1+n_2-2)p]F_{\alpha,p,n_1+n_2-1-p/(n_1+n_2-1-p)}$ , where  $F_{\alpha,p,n_1+n_2-1-p}$  represents the upper quantile 100 $\alpha$ % of the  $F$  distribution with  $f_1 = p$  and  $f_2 = n_1 + n_2 - 1 - p$  degrees of freedom.

### 2.1.2 MNV Test

The statistic for the MNV test,  $T_c^{*2}$ , was computed as

$$T_c^{*2} = (\bar{X}_1, -\bar{X}_2, -\delta_0)^\top \left( \frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (\bar{X}_1, -\bar{X}_2, -\delta_0),$$

and the null hypothesis was rejected when

$$T_c^{*2} > \frac{\nu p}{(\nu + 1 - p)} F_{\alpha,p,\nu+1-p},$$

where the degrees of freedom adjustment is given by  $\nu$

$$\nu = \frac{p + p^2}{\sum_{i=1}^2 \frac{1}{n_i - 1} \left\{ \text{tr} \left[ \left( \frac{S_i S_e^{-1}}{n_i} \right)^2 \right] + \left[ \text{tr} \left( \frac{S_i S_e^{-1}}{n_i} \right) \right]^2 \right\}} \quad \text{and} \quad S_e = \frac{S_1}{n_1} + \frac{S_2}{n_2}.$$

### 2.1.3 Ahmad Test (AT)

The test statistic  $T_2$  for the AT was computed as

$$T_2 = 1 + \frac{nQ_0}{nQ_1/p}, \quad (5)$$

where the quantity  $Q_1$  is given by

$$Q_1 = \sum_{i=1}^2 Q_{i1},$$

with

$$Q_{i1} = \frac{E_i - U_{n_i}}{n_i}, \quad E_i = \sum_{k=1}^{n_i} \frac{\mathbf{X}_{ik}^\top \mathbf{X}_{ik}}{n_i} \quad \text{and} \quad U_{n_i} = \frac{1}{n_i(n_i-1)} \sum_{k=1}^{n_i} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{n_i} h(\mathbf{X}_{ik}, \mathbf{X}_{i\ell}),$$

and the quantity  $Q_0 = U_0/p$ , where

$$U_0 = \sum_{i=1}^2 U_{n_i} - 2U_{n_1 n_2},$$

with

$$U_{n_1 n_2} = \frac{1}{n_1 n_2} \sum_{k=1}^{n_1} \sum_{\ell=1}^{n_2} h(\mathbf{X}_{1k}, \mathbf{X}_{2\ell}),$$

that is,  $Q_0 \in U_0$  with kernels of  $U_{n_i}$  and  $U_{n_1 n_2}$  scaled by  $p$ ,

$$h(\mathbf{X}_{ik}, \mathbf{X}_{i\ell}) = \frac{\mathbf{X}_{ik}^\top \mathbf{X}_{i\ell}}{p} \quad \text{and} \quad h(\mathbf{X}_{1k}, \mathbf{X}_{2\ell}) = \frac{\mathbf{X}_{1k}^\top \mathbf{X}_{2\ell}}{p}.$$

The null hypothesis was rejected if  $T_2 \geq T_\alpha$ , where  $T_\alpha$  is given by  $T_\alpha = \chi_{\alpha;f}^2/f$ , with  $\chi_{\alpha;f}^2$  being the upper 100 $\alpha$ % quantile of the chi-squared distribution with degrees of freedom  $f$  given by

$$f = \frac{[tr(\mathbf{\Omega}_0)]^2}{tr(\mathbf{\Omega}_0^2)},$$

where

$$\mathbf{\Omega}_0 = \frac{n}{p} \hat{\mathbf{\Sigma}}_0 \quad \text{and} \quad \hat{\mathbf{\Sigma}}_0 = \sum_{i=1}^2 \frac{\mathbf{S}_i}{n_i}.$$

### 2.1.4 Hyodo, Takahashi, and Nishiyama Test (HTNT)

The  $D_2$  statistic for the HTNT was computed by

$$D_2 = \frac{p}{\hat{\sigma}} \left\{ \frac{(\bar{\mathbf{X}}_2 - \bar{\mathbf{X}}_1 - \boldsymbol{\delta})^\top (\bar{\mathbf{X}}_2 - \bar{\mathbf{X}}_1 - \boldsymbol{\delta})}{(1/n_1 + 1/n_2)tr(\mathbf{S})} - 1 \right\}, \quad (6)$$

where  $\hat{\sigma} = \sqrt{2p\hat{a}_2/\hat{a}_1^2}$ ,  $tr(\mathbf{S})$  corresponds to the trace of the matrix  $\mathbf{S}$ , and the constants  $\hat{a}_1$  and  $\hat{a}_2$  are given by

$$\hat{a}_1 = \frac{tr(\mathbf{S})}{p} \quad \text{and} \quad \hat{a}_2 = \frac{\nu^2}{(\nu+2)(\nu-1)p} \left[ tr(\mathbf{S}^2) - \frac{tr^2(\mathbf{S})}{\nu} \right],$$

with  $\nu = n - 2$ .

The null hypothesis was rejected if  $D_2 \geq z$ , where  $z = z(\alpha)$  is the upper 100 $\alpha$ % quantile of the  $D$  statistic, as given by (Hyodo *et al.*, 2014; Nishiyama *et al.*, 2014)

$$z(\alpha; \hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4) = z_\alpha + \frac{\sqrt{2}\hat{a}_3(z_\alpha^2 - 1)}{3\sqrt{p\hat{a}_2^3}} + \frac{1}{p} \left[ \frac{\hat{a}_4}{2\hat{a}_2^2} z_\alpha(z_\alpha^2 - 3) - \frac{2\hat{a}_3^2}{9\hat{a}_2^3} z_\alpha(2z_\alpha^2 - 5) \right] + \frac{1}{2\nu} z_\alpha,$$

where  $z_\alpha$  is the upper 100 $\alpha$ % quantile of the standard normal distribution. The remaining required quantities are given by

$$\hat{a}_3 = \frac{\nu^4}{(\nu+4)(\nu+2)(\nu-1)(\nu-2)p} \left[ tr(\mathbf{S}^3) - \frac{3tr(\mathbf{S}^2)tr(\mathbf{S})}{\nu} + \frac{2tr^3(\mathbf{S})}{\nu^2} \right],$$

$$\hat{a}_4 = \frac{\nu^3[b_1tr(\mathbf{S}^4) + b_2tr(\mathbf{S}^3)tr(\mathbf{S}) + b_3tr^2(\mathbf{S}^2) + b_4tr(\mathbf{S}^2)tr^2(\mathbf{S}) + b_5tr^4(\mathbf{S})]}{(\nu+6)(\nu+4)(\nu+2)(\nu+1)(\nu-1)(\nu-2)(\nu-3)p},$$

where the  $b_i$ s are

$$\begin{aligned} b_1 &= \nu^2(\nu^2 + \nu + 2), & b_2 &= -4\nu(\nu^2 + \nu + 2), & b_3 &= -\nu(2\nu^2 + 3\nu - 6), \\ b_4 &= 2\nu(5\nu + 6) & \text{and} & & b_5 &= -(5\nu + 6). \end{aligned}$$

### 2.1.5 Ahmad Test via Bootstrap (ATB)

For the computation of the non-parametric bootstrap proposal of the test by Ahmad (2018), the statistic  $T_2$  was first calculated on the original sample using expression (5). Using the estimates  $\bar{\mathbf{X}}_i$  from (1) of the original sample, the modified sample was obtained, given by

$$\mathbf{Y}_{ik} = \mathbf{X}_{ik} - \bar{\mathbf{X}}_i,$$

for  $i = 1, 2$  e  $k = 1, 2, \dots, n_i$ .

This sample was combined by grouping the  $n = \sum_{i=1}^2 n_i$  observations  $Y_{ik}$  into a single  $p$ -dimensional sample of size  $n$ , thus imposing the null hypothesis of equality of the mean vectors. This combined sample was resampled with replacement, recreating the structure of the original sample from two populations with  $n_i$   $p$ -varied observations from the  $i$ -th treatment or population, with  $i = 1, 2$ . This process was repeated  $B = 2,000$  times.

In each set of data originating from the bootstrap resampling, the statistic  $T_2$  in (5) was calculated, generating the final statistic  $T_{2\ell}$ , in the  $\ell$ -th bootstrap resampling were stored together with the original value, forming a vector of dimension  $\ell = B + 1$ , given by  $\mathbf{T} = [T_{21}, T_{22}, \dots, T_{2B}, T_{2(B+1)}]^\top$ . Subsequently, the  $p$ -value was computed by

$$p\text{-value} = \frac{\sum_{\ell=1}^{B+1} I(T_{2\ell} \geq T_{2(B+1)})}{B+1},$$

where  $I(T_{2\ell} \geq T_{2(B+1)})$  is the indicator function that returns 1 if  $T_{2\ell} \geq T_{2(B+1)}$  and 0 otherwise. The null hypothesis was rejected when the obtained  $p$ -value was less than or equal to the nominal significance level adopted, i.e., when  $p\text{-value} \leq \alpha$ .

### 2.1.6 Hyodo, Takahashi, and Nishiyama Test via Bootstrap (HTNTB)

Similarly to the statistic  $T_2$ , the use of the statistic by Hyodo *et al.* (2014) was considered. Thus, in the original sample, the statistics  $D_2$  were computed as in (6) and the modified sample was obtained, which was grouped into a single  $p$ -dimensional sample of size  $n$ . This combined sample was resampled with replacement, recreating the structure of the original sample from two populations with  $n_i$   $p$ -varied observations from the  $i$ -th treatment or population, with  $i = 1, 2$ . This process was also repeated  $B = 2,000$  times.

In each set of resampled data, the statistic  $D_2$  in (6) was obtained, generating the final statistic  $D_{2\ell}$ , in the  $\ell$ -th bootstrap resampling or in the original sample when  $\ell = B + 1$ . The values of the  $\ell$ -th bootstrap resampling were stored together with the original value, forming a vector of dimension  $B + 1$ , given by  $\mathbf{D} = [D_{21}, D_{22}, \dots, D_{2B}, D_{2(B+1)}]^\top$ . Subsequently, considering  $D_{2\ell}$  and  $D_{2(B+1)}$ , the  $p$ -value was computed in the same manner as described for ATB. The null hypothesis was rejected when the obtained  $p$ -value was less than or equal to the nominal significance level adopted.

Next, the strategies considered in this work to evaluate the performance of the tests are presented.

## 2.2 Simulations and test performance evaluation

The performance evaluation of the proposed tests, ATB and HTNTB, along with the tests HT, MNV, AT, and HTNT, was conducted via Monte Carlo simulation in two stages. In the first stage, where the simulations were performed under the null hypothesis  $H_0 : \mu_1 = \mu_2 = \mu$ , the proportion of rejections of the null hypothesis is related to the Type I error rate, and in the second stage, performed under the alternative hypothesis  $H_1$ , the proportion of rejections is related to the power.

Two random samples of independent and identically distributed vectors of sizes  $n_1$  and  $n_2$ , were considered, where  $n_1 \in \{10, 20, 50\}$ ,  $n_2 = 2n_1$ , with each dimension  $p \in \{2, 10, 50, 100, 300\}$ , generated from the multivariate normal,  $t$ -distribution with 3 degrees of freedom, and uniform  $(0, 1)$ , distributions. The covariance matrix structures  $\Sigma_i$ ,  $i = 1, 2$ , considered were compound symmetry (CS) and first-order autoregressive (AR(1)), defined, respectively, by  $\Sigma_i = \sigma^2[(1-\rho)\mathbf{I} + \rho\mathbf{J}]$  and  $\text{Cov}(X_k, X_l) = \kappa\rho^{|k-l|}$ ,  $\forall k, l$ , where  $\mathbf{I}$  is an identity matrix and  $\mathbf{J}$  is a matrix of ones, both  $p \times p$ . Specifically, for homoscedastic cases, only  $p \in \{10, 300\}$  with  $(n_1, n_2) = (20, 40), (40, 40)$  were considered.

In heteroscedastic cases, two configurations were considered for  $\Sigma_i$ , the first with both matrix structures being equicorrelated, i.e., CS, with  $\rho = 0.5$  e  $\rho = 0.8$  for  $\Sigma_1$  and  $\Sigma_2$ , respectively; and the second, CS and AR(1), respectively, both with  $\rho = 0.5$ . In the homoscedastic cases, CS was considered, with  $\rho = 0.5$ .

To evaluate the Type I error rate, the simulations were generated under the complete null hypothesis, i.e., with both populations having the same parametric mean vectors. Thus, the rejection of the null hypothesis was considered a Type I error. The probability of committing a Type I error was estimated by the proportion of experiments in which a significant difference between means was incorrectly detected relative to the total of  $N = 2,000$  simulated experiments.

The obtained Type I error rates were compared among themselves and with those obtained by Ahmad (2018). Since they were estimated via Monte Carlo simulation, they were not free from error. Therefore, the exact binomial test, with 99% confidence (Oliveira & Ferreira, 2010), was used to verify whether the tests are liberal, conservative, or exact. The hypotheses of the test are

$$\begin{aligned} H_0 : \alpha &= 0.05 \text{ (or } 0.01 \text{ or } 0.10) \\ &\text{versus} \\ H_1 : \alpha &\neq 0.05 \text{ (or } 0.01 \text{ or } 0.10). \end{aligned}$$

If the null hypothesis was rejected and the observed Type I error rates were significantly ( $p\text{-value} < 0.01$ ) lower than the nominal significance level, the test in question was considered conservative; if the observed Type I error rates were significantly ( $p\text{-value} < 0.01$ ) higher than the nominal level, the test was considered liberal; and if the observed Type I error rates were not significantly ( $p\text{-value} < 0.01$ ) different from the nominal level, the test was considered exact (Oliveira & Ferreira, 2010). Considering  $\gamma$  as the number of rejections of  $H_0$  for  $N = 2,000$  Monte Carlo simulations at the nominal significance level  $\alpha$ , the test statistic can be obtained using the relationship between the  $F$  and binomial distributions, with a success probability of  $\alpha$ .

The test statistic was computed as

$$F_b = \left( \frac{\gamma - 1}{N - \gamma} \right) \left( \frac{1 - \alpha}{\alpha} \right),$$

which, under the null hypothesis  $H_0$ , follows an  $F$  distribution with  $\nu_1 = 2(N - \gamma)$  and  $\nu_2 = 2(\gamma + 1)$  degrees of freedom. If  $F_b \leq F_{\nu_1, \nu_2}(\alpha/2)$  or  $F_b > F_{\nu_1, \nu_2}(1 - \alpha/2)$ , then the null hypothesis is rejected at the 1% significance level, 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  $\nu_1$  and  $\nu_2$  degrees of freedom (Oliveira & Ferreira, 2010).

In the second stage, for power evaluation, the simulations were conducted following the procedures described for Type I error, except that  $X_{ik}$  was generated from multivariate distributions with different mean vectors, i.e.,  $X_{ik} \sim \mathfrak{F}_i(\mu_i, \Sigma_i)$ . The mean vector of population 1,  $\mu_1$ , was defined according to the simulation distribution in each case, and the other mean vector was set as

$$\mu_2 = \mu_1 + \delta,$$

where  $\delta$  is defined as  $\delta = \Delta p_1$ , with  $\Delta = 0.2(0.2)1$  and the vector  $p_1 \in p_1 = [1/p, 2/p, \dots, p/p]$ , as presented in Ahmad (2018).

### 3. Results and Discussion

An evaluation of the results obtained from the simulations was conducted for the three significance levels adopted,  $\alpha \in \{0.10, 0.05, 0.01\}$ , which revealed similar behaviors when considering the same configurations for each  $\alpha$ . Therefore, the simulation results for  $\alpha = 0.05$ , will be presented and



discussed, as in Ahmad (2018) and Krishnamoorthy & Yu (2004), to enable a direct comparison between the tests studied. Occasionally, when the pattern does not hold, results for other significance levels will be discussed.

The first part of the test evaluation was based on the results of type I error rates for samples simulated under  $H_0$ , that is, when  $\mu_1 = \mu_2$ . These results were marked with symbols to classify the tests as liberal (+), conservative (−) or exact (without a symbol), according to the exact binomial test (Oliveira & Ferreira, 2010). In the second part, the tests were compared by examining the power, where the samples were simulated under  $H_1$ , that is,  $\mu_1 \neq \mu_2$ , as described in 2.

The tests were evaluated in various configurations to verify their robustness. Scenarios of ideal cases were considered, in which the basic assumptions of classical tests were met, i.e., homoscedastic balanced and unbalanced cases with multivariate normal distribution. However, more general cases were also evaluated, including heteroscedastic and unbalanced situations, where the distributions were not necessarily multivariate normal. In both cases, low and high-dimensional scenarios were considered.

### 3.1 Homoscedasticity of Covariance Matrices

To evaluate performance, the type I error rates of the proposed tests were compared with those of the HT, MNV, HTNT, and AT tests. Initially, the tests were subjected to cases where the basic assumptions of classical tests were met.

#### 3.1.1 Type I error rate

In the case of homoscedasticity of covariance matrices, with  $\alpha = 0.05$ , as presented in Table 1, when analyzing the ideal scenario for the tests,  $p = 10$ ,  $(n_1, n_2) = (40, 40)$ , where the populations (treatments) follow a multivariate normal distribution with  $\mu = \mathbf{0}$  and  $\sigma^2$  as described in 2, it can be observed that all tests controlled the Type I error rate, with most doing so exactly according to the exact binomial test. This corroborates the results found in the configurations studied by Gebert (2014) for the MNV test. It is worth noting that HTNTB was conservative in controlling the Type I error rate, with a rate of 0.026. When the nominal level of significance,  $\alpha$ , is changed to  $\alpha = 0.10$ , in addition to HTNTB, the AT test was also conservative, though not as strongly conservative as HTNTB, which had a Type I error rate of 0.041, while AT presented a rate of 0.075, being considered slightly conservative, as it is close to the confidence interval for being considered exact (0.0834, 0.1185) for  $\alpha = 0.10$ . Another important point to highlight is that for  $\alpha = 0.10$ , HTNT showed the opposite result to its bootstrap version, HTNTB, being liberal, i.e., it did not control the Type I error rate. For  $\alpha = 0.01$ , all tests controlled the Type I error rate exactly. In the unbalanced case,  $(n_1, n_2) = (20, 40)$ , the behavior of the tests was similar to that observed in the balanced case, with the main difference being that AT for  $\alpha = 0.10$ , became exact, but for  $\alpha = 0.01$  it was the only test that did not control the Type I error rate, while its bootstrap version, ATB, was exact in all previously mentioned scenarios, proving to be a consistent test, similar to more traditional tests like HT and MNV.

For the high-dimensionality case,  $p = 300$ ,  $(n_1, n_2) \in \{(40, 40), (20, 40)\}$ , still in Table 1, it is possible to observe that this consistency of ATB remains, being an exact test in controlling the Type I error rate for all  $\alpha$  levels, which is not the case for AT. In this scenario, AT did not control the Type I error rate with  $\alpha = 0.01$  in the unbalanced case. HTNT controlled the Type I error rate only at  $\alpha = 0.01$ , while its bootstrap version controlled it in all scenarios, although it was conservative at  $\alpha = 0.05$ , in the balanced case, and at  $\alpha = 0.10$ , in both scenarios.

When testing populations following a multivariate  $t_3$  distribution with three degrees of freedom, thereby violating some assumptions of the tests, the overall performance was similar to the multivariate normal case. As expected, the HT and MNV tests showed changes in their Type I error control, becoming conservative in some cases. Notably, the AT test began to control the Type I



**Table 1.** Type I Error Rates for the Tests HT, MNV, HTNT, AT, HTNTB and ATB, considering the number of variables ( $p$ ), sample sizes ( $n_1, n_2$ ), covariance matrix structures (CS), multivariate distributions, and a nominal significance level ( $\alpha$ ) of 0.05 under  $H_0$ , for homoscedastic cases.

CS with $\rho = 0.5$								
$p$	$\alpha$	$n_1, n_2$	HT	MNV	HTNT	AT	HTNTB	ATB
Multivariate Normal - Low Dimensionality								
10	0.10	40, 40	0.1020	0.1005	0.1215 <sup>+</sup>	0.0750 <sup>-</sup>	0.0410 <sup>-</sup>	0.1000
		20, 40	0.1010	0.1055	0.1195 <sup>+</sup>	0.0895	0.0565 <sup>-</sup>	0.0990
	0.05	40, 40	0.0525	0.0520	0.0540	0.0410	0.0260 <sup>-</sup>	0.0435
		20, 40	0.0520	0.0520	0.0545	0.0595	0.0315 <sup>-</sup>	0.0525
	0.01	40, 40	0.0090	0.0080	0.0090	0.0150	0.0090	0.0095
		20, 40	0.0090	0.0115	0.0100	0.0180 <sup>+</sup>	0.0075	0.0140
Multivariate Normal - High Dimensionality								
300	0.10	40, 40	–	–	0.1265 <sup>+</sup>	0.0880	0.0545 <sup>-</sup>	0.0955
		20, 40	–	–	0.1490 <sup>+</sup>	0.1090	0.0700 <sup>-</sup>	0.1175
	0.05	40, 40	–	–	0.0710 <sup>+</sup>	0.0495	0.0345 <sup>-</sup>	0.0445
		20, 40	–	–	0.0845 <sup>+</sup>	0.0630	0.0460	0.0575
	0.01	40, 40	–	–	0.0075	0.0135	0.0080	0.0075
		20, 40	–	–	0.0150	0.0230 <sup>+</sup>	0.0185 <sup>+</sup>	0.0125
Multivariate $t_3$ - Low Dimensionality								
10	0.10	40, 40	0.0775 <sup>-</sup>	0.0750 <sup>-</sup>	0.1245 <sup>+</sup>	0.0815 <sup>-</sup>	0.0460 <sup>-</sup>	0.0930
		20, 40	0.0865	0.0840	0.1330 <sup>+</sup>	0.0865	0.0460 <sup>-</sup>	0.0940
	0.05	40, 40	0.0380 <sup>-</sup>	0.0340 <sup>-</sup>	0.0555	0.0450	0.0310 <sup>-</sup>	0.0475
		20, 40	0.0410	0.0345 <sup>-</sup>	0.0630	0.0495	0.0305 <sup>-</sup>	0.0465
	0.01	40, 40	0.0065	0.0065	0.0105	0.0135	0.0095	0.0125
		20, 40	0.0065	0.0055	0.0125	0.0135	0.0105	0.0090
Multivariate $t_3$ - High Dimensionality								
300	0.10	40, 40	–	–	0.1115	0.0790 <sup>-</sup>	0.0490 <sup>-</sup>	0.0885
		20, 40	–	–	0.1095	0.0745 <sup>-</sup>	0.0485 <sup>-</sup>	0.0840
	0.05	40, 40	–	–	0.0605	0.0435	0.0325 <sup>-</sup>	0.0435
		20, 40	–	–	0.0490	0.0435	0.0340 <sup>-</sup>	0.0440
	0.01	40, 40	–	–	0.0090	0.0140	0.0105	0.0105
		20, 40	–	–	0.0115	0.0155	0.0120	0.0105
Multivariate Uniform (0, 1) - Low Dimensionality								
10	0.10	40, 40	0.1055	0.1035	0.1290 <sup>+</sup>	0.0965	0.0590 <sup>-</sup>	0.0865
		20, 40	0.0935	0.1005	0.1310 <sup>+</sup>	0.0920	0.0585 <sup>-</sup>	0.1015
	0.05	40, 40	0.0540	0.0540	0.0645 <sup>+</sup>	0.0555	0.0325 <sup>-</sup>	0.0500
		20, 40	0.0445	0.0505	0.0580	0.0565	0.0370 <sup>-</sup>	0.0475
	0.01	40, 40	0.0115	0.0110	0.0140	0.0180 <sup>+</sup>	0.0135	0.0075
		20, 40	0.0105	0.0100	0.0120	0.0215 <sup>+</sup>	0.0095	0.0145
Multivariate Uniform (0, 1) - High Dimensionality								
300	0.10	40, 40	–	–	0.1295 <sup>+</sup>	0.0910	0.0600 <sup>-</sup>	0.0985
		20, 40	–	–	0.1305 <sup>+</sup>	0.0995	0.0700 <sup>-</sup>	0.1075
	0.05	40, 40	–	–	0.0710 <sup>+</sup>	0.0530	0.0360 <sup>-</sup>	0.0475
		20, 40	–	–	0.0725 <sup>+</sup>	0.0655 <sup>+</sup>	0.0470	0.0600
	0.01	40, 40	–	–	0.0095	0.0180 <sup>+</sup>	0.0115	0.0070
		20, 40	–	–	0.0165	0.0250 <sup>+</sup>	0.0185 <sup>+</sup>	0.0115

<sup>-</sup>: Significantly (p-value < 1%) lower than  $\alpha$ .  
<sup>+</sup>: Significantly (p-value < 1%) higher than  $\alpha$ .

Source: Author (2024).

error rate at  $\alpha = 0.01$ . In the high-dimensional case, a significant difference was observed in HTNT, which controlled the Type I error rate in all scenarios, while AT became conservative at  $\alpha = 0.10$ . Finally, ATB was the only test that controlled the Type I error rate exactly across all scenarios under the multivariate  $t_3$  distribution.

An additional scenario, which further deviates from the ideal conditions for performing the tests, occurs when the populations follow a multivariate uniform distribution  $(0, 1)$ . In the low-dimensional case, the tests showed similar performance to the multivariate normal distribution, with statistically significant differences. The HTNT test failed to adequately control the Type I error rate in the balanced case for  $\alpha = 0.05$ , while the AT test precisely controlled it at  $\alpha = 0.10$ , but failed to maintain control at  $\alpha = 0.01$ . In the high-dimensional case, the similarity persisted in most scenarios; however, AT failed to control the Type I error rate in the unbalanced case at  $\alpha = 0.05$  and in both cases for  $\alpha = 0.01$ . In general, for homoscedasticity of covariance matrices, the ATB test stood out as the only one to be exact in all simulated scenarios, consistently controlling the Type I error rate.

### 3.1.2 Power

When evaluating the power of the tests for the scenarios presented in the Type I error rate analysis above, similar behavior was observed across all  $\alpha$  levels. Therefore, only  $\alpha = 0.05$  will be presented and discussed. In Figure 1, from left to right, the scenarios of low and high dimensionality are displayed, respectively, and from top to bottom, the multivariate normal,  $t_3$  and uniform  $(0, 1)$  distributions are shown. The last case is unbalanced because, despite showing very similar power to the balanced case (as will be discussed below), AT did not control the Type I error rate according to the results presented in 3.1.1. In the ideal case, with normality and low dimensionality, as expected, the classical tests HT and MNV had higher power than the other tests, both in balanced and unbalanced cases. However, this difference in power was not as pronounced when the mean differences were small ( $\Delta \in \{0.2, 0.4\}$ ) or for larger differences ( $\Delta = 1.0$ ). Another anticipated result was that tests that showed conservative control of the Type I error rate exhibited lower power. For the multivariate  $t_3$  the tests performed very similarly to the multivariate normal case, though generally, the tests were slightly less powerful. Finally, for the multivariate uniform  $(0, 1)$  distribution, the power of the tests was significantly higher than in the normal case, reaching 100% power for  $\Delta = 0.4$  in the balanced case and  $\Delta = 0.6$  in the unbalanced case. The performance for high dimensionality was similar to that for low dimensionality.

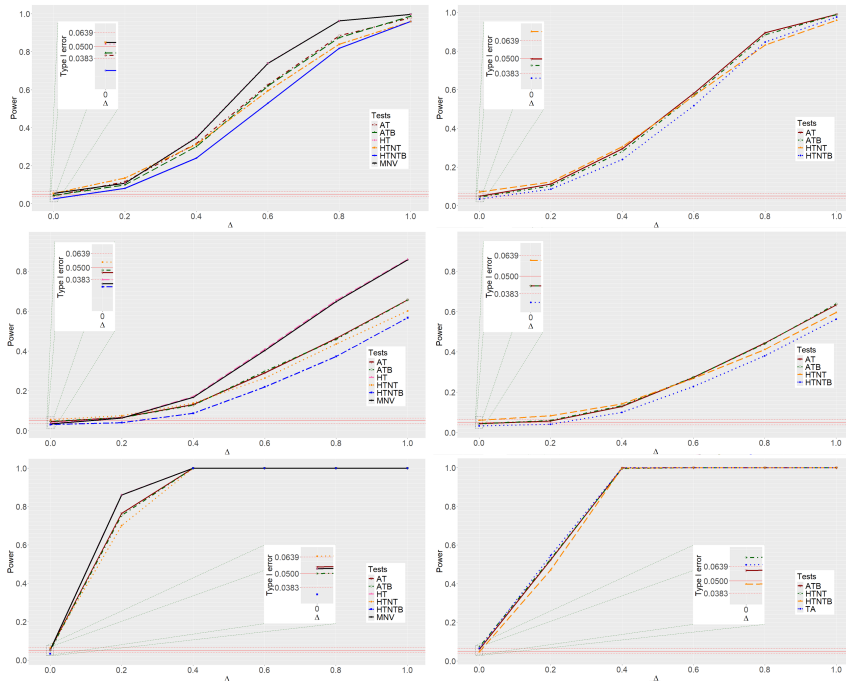
Overall, as expected, in the ideal scenarios, HT and MNV showed the best performance, i.e., exact control of the Type I error rate and higher power compared to their competitors. However, these tests were not robust, as they were not consistent across all scenarios. Additionally, these tests have limitations for high-dimensional cases, as their test statistics depend on the inverse of the covariance matrix, which is not possible when  $p \geq n - 2$ . Thus, the test with the best overall performance in the homoscedastic case was ATB, as it was robust and consistent in all the simulated cases in this study.

## 3.2 Heteroscedasticity of Covariance Matrices

To better assess the robustness of the tests, it is necessary to move further away from the ideal scenario. Thus, the tests were evaluated in two heteroscedastic scenarios, as described in 2.2. The first scenario involves CS and CS with different  $\rho$  values, introducing a slight variation between the covariance matrix structures of the two populations. The second scenario involves CS and AR(1) creating a more drastic difference between the structures.

### 3.2.1 Type I Error Rates (CS, CS)

In the scenario where the two populations/treatments have different compound symmetry structures, it was generally observed that the behavior of the tests, whether they controlled the Type I error rate or not, was very similar across the nominal significance levels ( $\alpha$ ) studied. There were only a few differences. Generally, at  $\alpha = 0.10$  tests that did not control the Type I error rate exactly at  $\alpha = 0.05$ , tended to show more cases of liberal or conservative behavior. Therefore, only  $\alpha = 0.05$  will be discussed here, as the behavior at other  $\alpha$  levels was similar to what was observed in 3.1.2.



Source: Author (2024).

**Figure 1.** Power and Type I Error Rates for the Hotelling's  $T^2$  Test (HT), MNV, HTNT, Ahmad's Test (AT), HTNTB, and ATB, considering  $p \in \{10, 300\}$ ,  $(n_1, n_2) = (40, 40), (20, 40)$ ,  $\alpha = 0.05$ , homoscedasticity, multivariate normal,  $t_3$ , and uniform (0, 1) distributions in low and high dimensionality cases.

In Table 2, it can be observed that in the multivariate normal case, the performance of the tests changed when the structure of one of the covariance matrices was modified. Comparing with  $p = 10$  as shown in Table 1 it can be noted that the tests most negatively affected by this change were HT, MNV and HTNT, meaning they did not control the Type I error rate, yielding rates significantly higher than expected. For instance, HT and MNV were considered strongly liberal, while HTNT was slightly liberal, meaning it was close to the confidence interval of the exact binomial test for these cases. Slightly liberal results may still be acceptable for use in research if they fall within an acceptable risk in an experiment, as a higher power for the tests might be expected as a consequence. The goal is to find exact or conservative tests, provided that conservative tests have power as high or very close to those that are exact in controlling the Type I error rate. Other tests affected were HTNTB and ATB, which became, respectively, strongly and slightly conservative in some cases, with HTNTB being strongly conservative in all scenarios. This trend observed in the multivariate normal distribution extends to the other multivariate distributions studied here.

Also in Table 2, it is clear that traditional tests like HT and MNV were more affected. Moreover, another factor that exacerbated poor performance in controlling the Type I error rate for these tests

was the increase in  $p$  (the number of variables observed in the populations). When  $p$  increased, the tests tended to produce increasingly unacceptable Type I error rates, with probabilities as high as 47.00% or 78.65%, making it more advisable to flip a coin for decision-making. For MNV, as presented by Krishnamoorthy & Yu (2004), with  $p = 2$ , the test controlled the Type I error rate exactly, but as  $p$  increased, the performance of the test declined, particularly when  $p$  approached one of the sample sizes  $n_i$ . For example, with  $p = 10$ ,  $(n_1, n_2) = (10, 20)$ , MNV exhibited slightly liberal behavior, but it began controlling the Type I error rate exactly when the sample sizes were increased, thereby distancing  $p$  from  $n_i$ , which corroborates the results presented by Gebert (2014), where the test showed liberal behavior for  $p = 7$ ,  $(n_1, n_2) = (8, 30)$  for all levels of heterogeneity between the covariance matrices studied.

The other tests remained more stable as  $p$  increased. Sample size was not as relevant a factor as  $p$  in this Type I error rate analysis, not contributing significantly for most tests based on the simulations studied in this work. However, it is worth mentioning that Takahashi *et al.* (2013) recommend the use of HTNT for small sample sizes, likely because the largest sample size the authors used was 40. In Table 2, it can be observed that in low-dimensionality settings, HTNT performed worse with  $(n_1, n_2) = (50, 100)$ , which may reinforce the authors' recommendation to use the test only with small sample sizes. However, this poor performance with  $(n_1, n_2) = (50, 100)$  was not observed in high-dimensional settings.

**Table 2.** Type I Error Rates for the Tests HT, MNV, HTNT, AT, HTNTB, and ATB, considering the number of variables ( $p$ ), sample sizes ( $n_1, n_2$ ), covariance matrix structures (SC, SC), multivariate normal and uniform distributions, low and high dimensionality, and a nominal significance level of 0.05 ( $\alpha$ ), under  $H_0$ .

$p$		$\Sigma_1$ : CS, $\rho = 0.5$ and $\Sigma_2$ : CS, $\rho = 0.8$					
		HT	MNV	HTNT	AT	HTNTB	ATB
Multivariate Normal - Low Dimensionality							
2	10, 20	0.0690 <sup>+</sup>	0.0470	0.0540	0.0530	0.0155 <sup>-</sup>	0.0535
	20, 40	0.0675 <sup>+</sup>	0.0430	0.0595	0.0440	0.0130 <sup>-</sup>	0.0470
	50, 100	0.0690 <sup>+</sup>	0.0450	0.0660 <sup>+</sup>	0.0485	0.0110 <sup>-</sup>	0.0440
10	10, 20	0.1550 <sup>+</sup>	0.0900 <sup>+</sup>	0.0545	0.0505	0.0180 <sup>-</sup>	0.0345 <sup>-</sup>
	20, 40	0.1535 <sup>+</sup>	0.0540	0.0585	0.0550	0.0155 <sup>-</sup>	0.0350 <sup>-</sup>
	50, 100	0.1475 <sup>+</sup>	0.0410	0.0650 <sup>+</sup>	0.0565	0.0135 <sup>-</sup>	0.0380 <sup>-</sup>
50	20, 40	0.1830 <sup>+</sup>	0.2925 <sup>+</sup>	0.0615	0.0505	0.0175 <sup>-</sup>	0.0395
	50, 100	0.4195 <sup>+</sup>	0.1625 <sup>+</sup>	0.0600 <sup>+</sup>	0.0455	0.0105 <sup>-</sup>	0.0420
100	50, 100	0.4700 <sup>+</sup>	0.7865 <sup>+</sup>	0.0695 <sup>+</sup>	0.0580	0.0160 <sup>-</sup>	0.0425
Multivariate Normal - High Dimensionality							
50	10, 20	–	–	0.0505	0.0535	0.0180 <sup>-</sup>	0.0290 <sup>-</sup>
100	10, 20	–	–	0.0505	0.0515	0.0170 <sup>-</sup>	0.0345 <sup>-</sup>
	20, 40	–	–	0.0585	0.0575	0.0195 <sup>-</sup>	0.0395
300	10, 20	–	–	0.0665 <sup>+</sup>	0.0610	0.0235 <sup>-</sup>	0.0360 <sup>-</sup>
	20, 40	–	–	0.0585	0.0580	0.0210 <sup>-</sup>	0.0410
	50, 100	–	–	0.0610	0.0490	0.0140 <sup>-</sup>	0.0370 <sup>-</sup>
Multivariate Uniform (0, 1) - Low Dimensionality							
2	10, 20	0.0795 <sup>+</sup>	0.0535	0.0560	0.0640 <sup>+</sup>	0.0165 <sup>-</sup>	0.0475
	20, 40	0.0785 <sup>+</sup>	0.0545	0.0685 <sup>+</sup>	0.0610	0.0190 <sup>-</sup>	0.0460
	50, 100	0.0775 <sup>+</sup>	0.0505	0.0775 <sup>+</sup>	0.0495	0.0130 <sup>-</sup>	0.0390
10	10, 20	0.1505 <sup>+</sup>	0.0780 <sup>+</sup>	0.0530	0.0535	0.0195 <sup>-</sup>	0.0345 <sup>-</sup>
	20, 40	0.1585 <sup>+</sup>	0.0515	0.0565	0.0590	0.0155 <sup>-</sup>	0.0415
	50, 100	0.1515 <sup>+</sup>	0.0435	0.0700 <sup>+</sup>	0.0505	0.0155 <sup>-</sup>	0.0385
50	20, 40	0.2370 <sup>+</sup>	0.3705 <sup>+</sup>	0.0570	0.0475	0.0130 <sup>-</sup>	0.0375 <sup>-</sup>
	50, 100	0.4540 <sup>+</sup>	0.1525 <sup>+</sup>	0.0610	0.0480	0.0120 <sup>-</sup>	0.0405
100	50, 100	0.5890 <sup>+</sup>	0.8095 <sup>+</sup>	0.0730 <sup>+</sup>	0.0520	0.0195 <sup>-</sup>	0.0380 <sup>-</sup>
Multivariate Uniform (0, 1) - High Dimensionality							
50	10, 20	–	–	0.0570	0.0595	0.0220 <sup>-</sup>	0.0415
100	10, 20	–	–	0.0580	0.0595	0.0220 <sup>-</sup>	0.0395
	20, 40	–	–	0.0450	0.0430	0.0130 <sup>-</sup>	0.0325 <sup>-</sup>
300	10, 20	–	–	0.0580	0.0560	0.0140 <sup>-</sup>	0.0305 <sup>-</sup>
	20, 40	–	–	0.0665 <sup>+</sup>	0.0575	0.0175 <sup>-</sup>	0.0385
	50, 100	–	–	0.0615	0.0540	0.0200 <sup>-</sup>	0.0435

–: Significantly (p-value < 1%) lower than  $\alpha$ .  
+: Significantly (p-value < 1%) higher than  $\alpha$ .

Source: Author (2024).

Overall, for low-dimensionality in the multivariate normal case, the HT, MNV and HTNT tests were liberal. HT changed from slightly liberal at  $p = 2$  to strongly liberal at  $p \geq 10$ , and MNV showed similar behavior for  $p = 10$  and lower sample sizes, which was relevant for this test in this structure, and for  $p \in \{50, 100\}$ . HTNT, although it did not control the error rate in some cases at  $\alpha = 0.05$  and increased the number of liberal cases at  $\alpha = 0.10$ , controlled the Type I error rate in all scenarios for  $\alpha = 0.01$ , being slightly conservative in only one case:  $p = 50$  with  $(n_1, n_2) = (50, 100)$ . Its bootstrap version, HTNTB, was strongly conservative across all  $\alpha$  levels and, like for  $\alpha = 0.05$ , was not affected by increases in  $p$  or  $n$ , indicating that using the bootstrap methodology in this test was important for stability in controlling the Type I error rate, though this test is expected to have low power. The AT showed exact control for both  $\alpha = 0.05$  and  $\alpha = 0.10$ , but had a few cases of being slightly liberal for  $\alpha = 0.01$  at  $p = 10$ ,  $(n_1, n_2) = (10, 20)$  and  $p = 100$ ,  $(n_1, n_2) = (50, 100)$ . Its bootstrap version, ATB, controlled the Type I error rate in all scenarios, being slightly conservative in some cases. It is important to note that, as in 3.1.1, the Type I error rate of the tests were similar between the multivariate normal and  $t_3$  distributions.

When studying the tests in the multivariate uniform (0, 1) distribution, an increase in scenarios where the tests were liberal was observed, along with an overall increase in error rates. Most of the tests exhibited similar behavior in terms of Type I error rate control as observed in the case of the

multivariate normal distribution, with the exception of the AT test, which failed to control the rate for  $p = 2$  and  $(n_1, n_2) = (10, 20)$ . Additionally, there was a higher number of failures in controlling the rate for  $\alpha = 0.01$  compared to the normal distribution for the same significance level.

In high-dimensionality, as shown in Table 2, the performance of the tests was very similar to the low-dimensionality case, both in terms of  $\alpha$  levels and the multivariate distributions studied in this work. The main differences were that HTNT achieved better control of the Type I error rate, i.e., it had fewer liberal cases. Its bootstrap version performed similarly in high-dimensionality as observed in low-dimensionality, being strongly conservative in most scenarios. The AT also exhibited behavior similar to that seen in low-dimensionality, still showing liberal cases at  $\alpha = 0.01$ , while its bootstrap version controlled the error rate in all cases, though there were more slightly conservative cases.

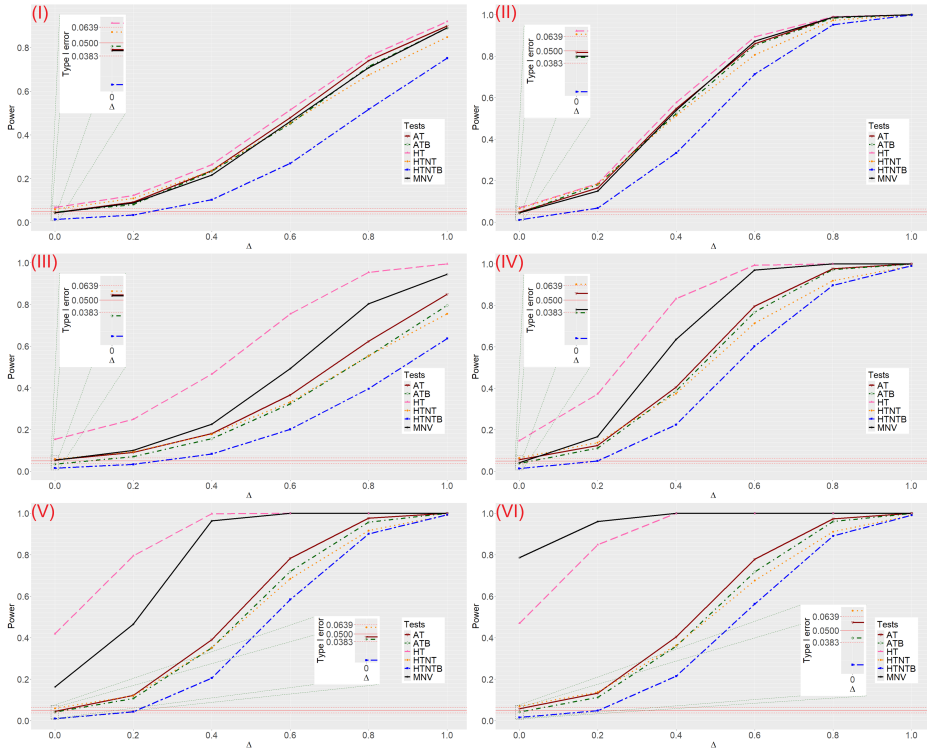
### 3.2.2 Power (CS, CS)

The power of the tests changed significantly in some cases compared to the previously studied scenarios in 3.1.2. Overall, the power of the tests was very similar across the different  $\alpha$  levels, as was observed in 3.1.2, so only the results for  $\alpha = 0.05$  will be presented. Additionally, the phenomena of decreased test power in the multivariate  $t_3$  distribution compared to the multivariate normal and the significant increase in power in the multivariate uniform  $(0, 1)$  were similar to those discussed in 3.1.2. For this reason, most of the results and discussions for the multivariate  $t_3$  and uniform  $(0, 1)$  distributions will be omitted. Ahmad (2018) considered the multivariate  $t$  distribution with 7 degrees of freedom and obtained power values similar to those of the multivariate normal distribution.

In Figure 2 the following scenarios for the multivariate distribution are presented from top to bottom and right to left:  $p = 2$  with  $(n_1, n_2) = (20, 40)$  (I),  $p = 2$  with  $(n_1, n_2) = (50, 100)$  (II),  $p = 10$  with  $(n_1, n_2) = (20, 40)$  (III),  $p = 10$  with  $(n_1, n_2) = (50, 100)$  (IV),  $p = 50$  with  $(n_1, n_2) = (50, 100)$  (V) and  $p = 100$  with  $(n_1, n_2) = (50, 100)$  (VI). In scenario (I), for the multivariate normal distribution, among the tests that controlled the Type I error rate, most showed similar power, except for HTNTB. For mean differences  $\Delta \in \{0.8, 1.0\}$ , the power of the tests diverged, with a drop in HTNT's rate and an increase in AT's rate, making AT the most powerful among the tests that controlled the Type I error rate. This was followed by ATB and MNV, both almost as powerful as AT. Lastly, HTNTB, being strongly conservative as expected, was the least powerful test. In scenario (II), where only the sample sizes  $n$  were increased, the tests showed higher power, with performance similar to that of scenario (I) but with slightly higher power rates. The only exception was HTNT, which, like HT, did not control the Type I error rate and still showed lower power for  $\Delta \geq 0.4$ .

When  $p$  increased to 10, a more noticeable change in test power was observed. Among those that controlled the Type I error rate, MNV was the most powerful, with a significant difference from the other tests. However, it is important to note that, according to Table 2, MNV did not control the Type I error rate for  $p = 10$  and  $(n_1, n_2) = (10, 20)$ . The other tests, except for HT, which did not control the Type I error rate, and HTNTB, which was strongly conservative, had similar power levels, with AT being the most powerful and ATB and HTNT being very similar. When the sample sizes  $n$  were increased while keeping  $p$  constant, in scenario (IV), the tests showed the same effect as described in Figure 2, in scenario (II).

Finally, the last two scenarios highlight the effect of  $p$  on traditional tests like HT and MNV, which were extremely affected, becoming increasingly liberal and powerful and, therefore, not recommended. Another test affected by the increase in  $p$  was HTNT. Takahashi *et al.* (2013) had already observed that HTNT was sensitive to large  $p$  values in situations of homoscedasticity and heavy-tailed distributions. Thus, in these last two scenarios, the tests AT, ATB, and HTNTB were consistent in controlling the Type I error rate. Among them, AT was the most powerful, but ATB had power levels very close to AT, while HTNTB remained the least powerful.



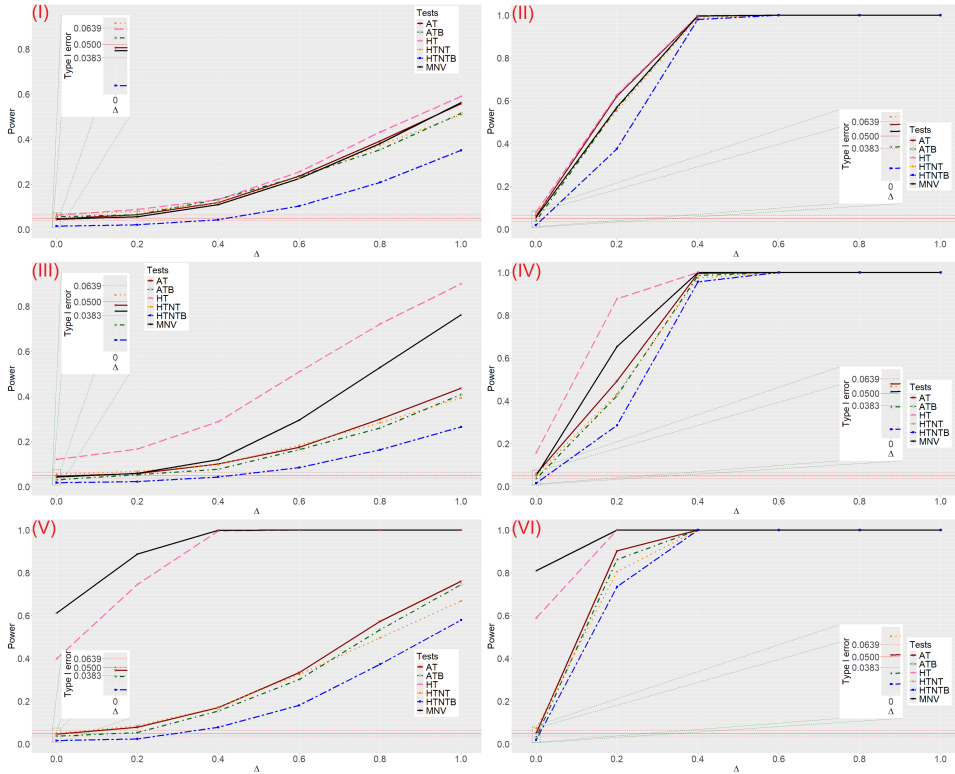
Source: Author (2024).

**Figure 2.** Power and Type I Error Rates for Hotelling's  $T^2$  Test (HT), MNV, HTNT, Ahmad's Test (AT), HTNTB, and ATB, considering  $\alpha = 0.05$ , heteroscedasticity, multivariate normal distribution, and low dimensionality.

In Figure 3, the following scenarios for the multivariate  $t_3$  and uniform  $(0, 1)$  distributions are presented:  $t_3$  with  $p = 2$  and  $(n_1, n_2) = (20, 40)$  (I), uniform  $(0, 1)$  with  $p = 2$  and  $(n_1, n_2) = (20, 40)$  (II),  $t_3$  with  $p = 10$  and  $(n_1, n_2) = (20, 40)$  (III), uniform  $(0, 1)$  with  $p = 10$  and  $(n_1, n_2) = (20, 40)$  (IV),  $t_3$  with  $p = 100$  and  $(n_1, n_2) = (50, 100)$  (V) and uniform  $(0, 1)$  with  $p = 100$  and  $(n_1, n_2) = (50, 100)$  (VI). Firstly, in the multivariate  $t_3$  distribution, a very similar behavior to that observed in the multivariate normal case was seen for the tests. However, it is worth noting that, in general, the tests were slightly less powerful than in the normal distribution. In the multivariate uniform  $(0, 1)$  distribution, similar to what was observed in 3.1.2, the tests exhibited high levels of power. Among those that controlled the Type I error rate in all scenarios, the power values were quite similar for each  $\Delta$  value, except for HTNTB, which showed lower power for  $\Delta = 0.2$ . The MNV test demonstrated high power in scenario (IV), but only for  $\Delta = 0.2$ . As expected, the HT test did not control the Type I error rate in any of the scenarios.

In high dimensionality, Figure 4 presents the following scenarios for multivariate normal and uniform  $(0, 1)$  distributions: normal with  $p = 50$  and  $(n_1, n_2) = (10, 20)$  (I), uniform  $(0, 1)$  with  $p = 50$  and  $(n_1, n_2) = (10, 20)$  (II), normal with  $p = 100$  and  $(n_1, n_2) = (10, 20)$  (III), uniform  $(0, 1)$  with  $p = 100$  and  $(n_1, n_2) = (10, 20)$  (IV), normal with  $p = 300$  and  $(n_1, n_2) = (10, 20)$  (V) and uniform  $(0, 1)$  with  $p = 300$  and  $(n_1, n_2) = (10, 20)$  (VI). There is a very strong similarity with the low dimensionality case, both in the multivariate normal scenarios (I), (III), and (V), and in the multivariate uniform  $(0, 1)$  scenarios (II), (IV), and (VI). Overall, the most powerful test was AT, followed by HTNT and ATB, and finally HTNTB, which remained highly conservative in all cases. Another important point to note, which becomes more evident in high dimensionality for



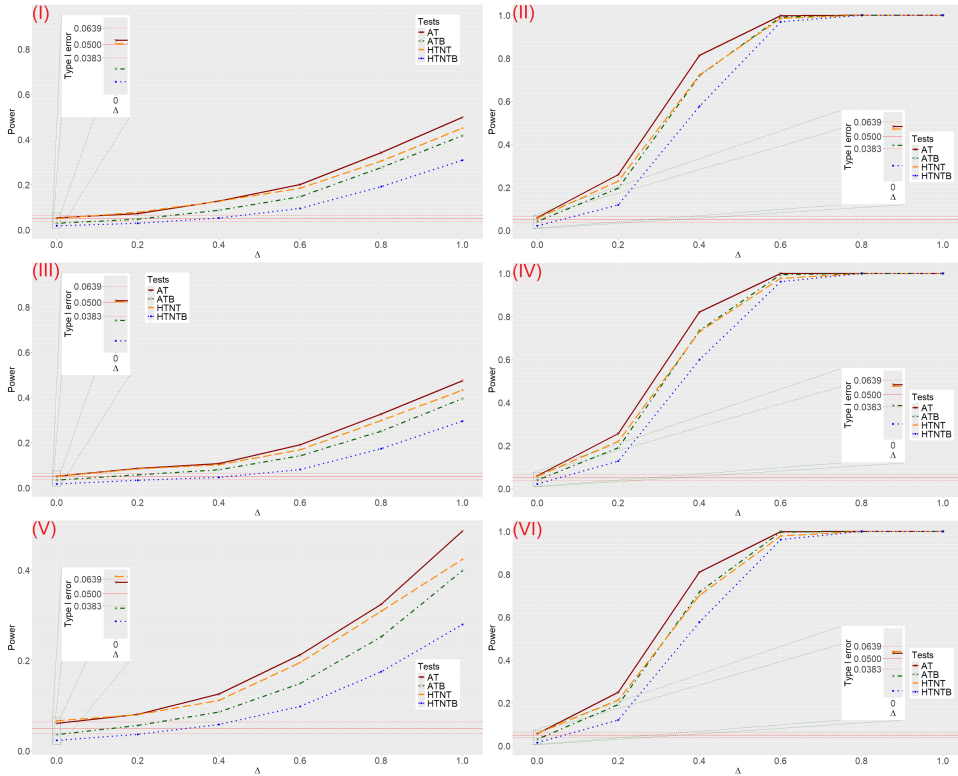


Source: Author (2024).

**Figure 3.** Power and Type I Error Rates for Hotelling's  $T^2$  Test (HT), MNV, HTNT, Ahmad's Test (AT), HTNTB, and ATB, considering  $\alpha = 0.05$ , heteroscedasticity, multivariate  $t_3$  and uniform  $(0, 1)$  distributions, and low dimensionality.

this heteroscedasticity structure, is the consistency in the tests' powers, highlighting their robustness to changes in  $p$ .

In general, for this scenario of heteroscedasticity in the covariance matrices, it can be concluded that the tests that controlled the Type I error rate in all scenarios were ATB and HTNTB, with ATB having the highest power among them, slightly lower than that presented by AT but more robust in this scenario, as was the case with homoscedasticity. HTNTB, on the other hand, was an extremely conservative test with very low power values.



Source: Author (2024).

**Figure 4.** Power and Type I Error Rates for Hotelling's  $T^2$  Test (TH), MNV, HTNT, Ahmad's Test (TA), HTNTB, and ATB, considering  $\alpha = 0.05$ , heteroscedasticity, multivariate normal and uniform (0, 1) distributions, and high dimensionality.

### 3.2.3 Type I Error Rates (CS, AR(1))

Moving further away from the ideal scenario, we now consider the case where the structures of the covariance matrices are significantly different, making this the most extreme scenario in this study. It is worth highlighting here that even in this structure, the multivariate normal and  $t_3$  distributions produced very similar results, although in  $t_3$  the tests had slightly fewer liberal cases, displaying the same pattern as in the normal distribution. Therefore, only the multivariate normal and uniform (0, 1) distributions are presented in Table 3.

As can be seen in Table 3, the performance of most tests was quite different from the previously studied cases. In this scenario, it is generally noticeable that the HT, HTNTB, and ATB tests were the most affected by the change in the covariance matrix structure. It is noteworthy that HT was the only test that, in low dimensionality, controlled the Type I error rate in all cases. Other tests that performed well in controlling the Type I error rate in low dimensionality were the MNV and AT tests, which had few liberal cases. Among these, MNV stood out at  $\alpha = 0.01$  and AT at  $\alpha = 0.10$ , with one outperforming the other in controlling the Type I error rate at each of these  $\alpha$  levels and being similar at  $\alpha = 0.05$ . Still in low dimensionality, HTNT had a performance similar to that observed in 3.2.1, just like AT, although HTNT had slightly more liberal cases. Another highlight was HTNTB, which shifted from being highly conservative to highly liberal, and, finally, ATB, which had controlled the Type I error rate in all cases in 3.2.1, became liberal in many cases in this scenario.

**Table 3.** Type I Error Rates for the Tests TH, MNV, HTNT, AT, HTNTB, and ATB, considering the number of variables ( $p$ ), sample sizes ( $n_1, n_2$ ), covariance matrix structures (CS, AR(1)), multivariate normal and uniform (0, 1) distributions, low and high dimensionality, and a nominal significance level ( $\alpha$ ) of 0.05, under  $H_0$ .

		$\Sigma_1$ : CS, $\rho = 0.5$ and $\Sigma_2$ : AR(1), $\rho = 0.5$					
$p$	$n_1, n_2$	HT	MNV	HTNT	AT	HTNTB	ATB
Multivariate Normal - Low Dimensionality							
2	10, 20	0.0530	0.0535	0.0520	0.0555	0.0190 <sup>-</sup>	0.0500
	20, 40	0.0470	0.0505	0.0615	0.0500	0.0200 <sup>-</sup>	0.0560
	50, 100	0.0490	0.0480	0.0770 <sup>+</sup>	0.0530	0.0170 <sup>-</sup>	0.0545
10	10, 20	0.0410	0.0535	0.0715 <sup>+</sup>	0.0645 <sup>+</sup>	0.0590	0.0700 <sup>+</sup>
	20, 40	0.0390	0.0480	0.0475	0.0515	0.0520	0.0665 <sup>+</sup>
	50, 100	0.0405	0.0460	0.0535	0.0580	0.0600	0.0660 <sup>+</sup>
50	20, 40	0.0365 <sup>-</sup>	0.0430	0.0825 <sup>+</sup>	0.0615	0.1100 <sup>+</sup>	0.1065 <sup>+</sup>
	50, 100	0.0315 <sup>-</sup>	0.0740 <sup>+</sup>	0.0695 <sup>+</sup>	0.0535	0.1095 <sup>+</sup>	0.0900 <sup>+</sup>
100	50, 100	0.0305 <sup>-</sup>	0.1125 <sup>+</sup>	0.0770 <sup>+</sup>	0.0490	0.1195 <sup>+</sup>	0.0980 <sup>+</sup>
Multivariate Normal - High Dimensionality							
50	10, 20	–	–	0.1055 <sup>+</sup>	0.0715 <sup>+</sup>	0.1210 <sup>+</sup>	0.0900 <sup>+</sup>
100	10, 20	–	–	0.1055 <sup>+</sup>	0.0595	0.1325 <sup>+</sup>	0.0835 <sup>+</sup>
	20, 40	–	–	0.0960 <sup>+</sup>	0.0670 <sup>+</sup>	0.1390 <sup>+</sup>	0.1065 <sup>+</sup>
300	10, 20	–	–	0.1315 <sup>+</sup>	0.0660 <sup>+</sup>	0.1600 <sup>+</sup>	0.1015 <sup>+</sup>
	20, 40	–	–	0.1130 <sup>+</sup>	0.0590	0.1480 <sup>+</sup>	0.1130 <sup>+</sup>
	50, 100	–	–	0.1065 <sup>+</sup>	0.0500	0.1485 <sup>+</sup>	0.1170 <sup>+</sup>
Multivariate Uniform (0, 1) - Low Dimensionality							
2	10, 20	0.0495	0.0540	0.0580	0.0710 <sup>+</sup>	0.0210 <sup>-</sup>	0.0480
	20, 40	0.0600	0.0595	0.0655 <sup>+</sup>	0.0595	0.0215 <sup>-</sup>	0.0565
	50, 100	0.0470	0.0475	0.0625	0.0485	0.0165 <sup>-</sup>	0.0595
10	10, 20	0.0515	0.0660 <sup>+</sup>	0.0615	0.0625	0.0545	0.0675 <sup>+</sup>
	20, 40	0.0440	0.0540	0.0540	0.0565	0.0520	0.0710 <sup>+</sup>
	50, 100	0.0445	0.0510	0.0465	0.0550	0.0575	0.0635
50	20, 40	0.0400	0.0585	0.0875 <sup>+</sup>	0.0585	0.1205 <sup>+</sup>	0.0955 <sup>+</sup>
	50, 100	0.0250 <sup>-</sup>	0.0725 <sup>+</sup>	0.0705 <sup>+</sup>	0.0550	0.1065 <sup>+</sup>	0.0895 <sup>+</sup>
100	50, 100	0.0265 <sup>-</sup>	0.0960 <sup>+</sup>	0.0905 <sup>+</sup>	0.0650 <sup>+</sup>	0.1435 <sup>+</sup>	0.1135 <sup>+</sup>
Multivariate Uniform (0, 1) - High Dimensionality							
50	10, 20	–	–	0.1005 <sup>+</sup>	0.0515	0.1155 <sup>+</sup>	0.0770 <sup>+</sup>
100	10, 20	–	–	0.1195 <sup>+</sup>	0.0675 <sup>+</sup>	0.1430 <sup>+</sup>	0.0935 <sup>+</sup>
	20, 40	–	–	0.0990 <sup>+</sup>	0.0660 <sup>+</sup>	0.1315 <sup>+</sup>	0.1045 <sup>+</sup>
300	10, 20	–	–	0.1300 <sup>+</sup>	0.0735 <sup>+</sup>	0.1665 <sup>+</sup>	0.1090 <sup>+</sup>
	20, 40	–	–	0.1115 <sup>+</sup>	0.0695 <sup>+</sup>	0.1590 <sup>+</sup>	0.1160 <sup>+</sup>
	50, 100	–	–	0.1085 <sup>+</sup>	0.0610	0.1450 <sup>+</sup>	0.1125 <sup>+</sup>

–: Significantly (p-value < 1%) lower than  $\alpha$ .

+: Significantly (p-value < 1%) higher than  $\alpha$ .

Source: Author (2024).

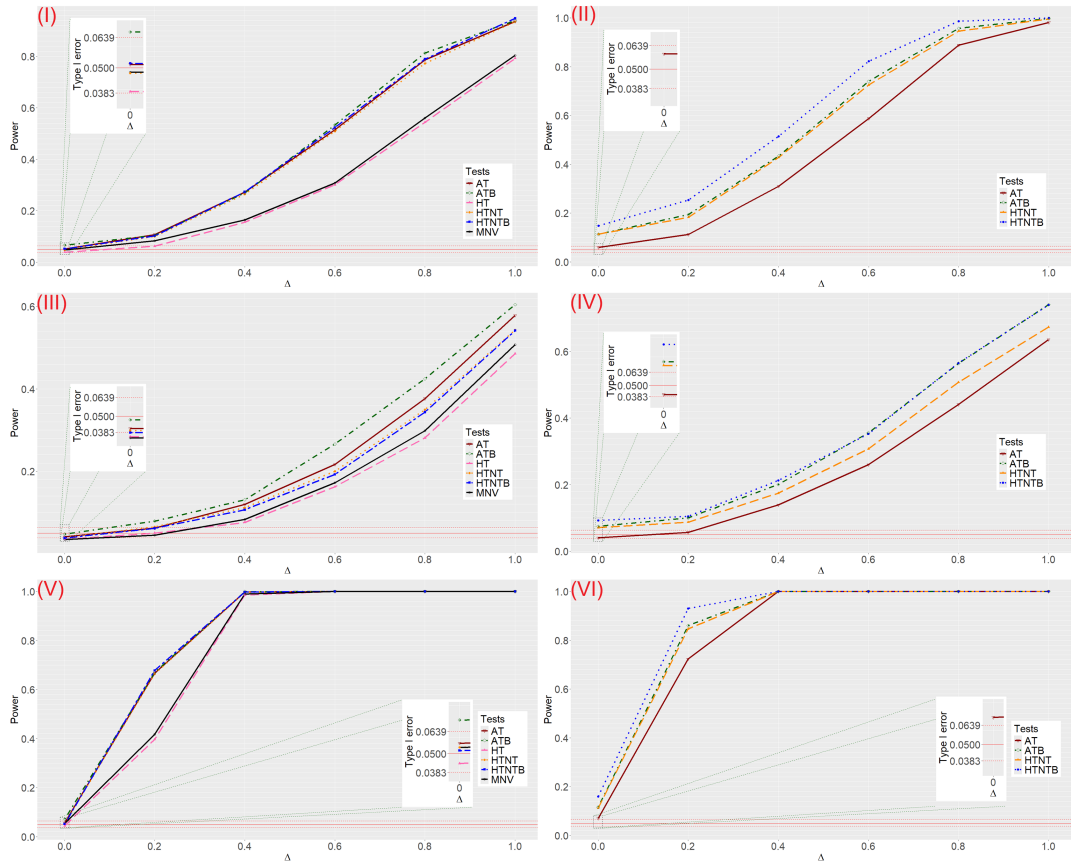
In high dimensionality, corroborating the positive results obtained by Ahmad (2018) for AT in controlling the Type I error rate, the test with the best performance was AT, which stood out by controlling the Type I error rate in all cases at  $\alpha = 0.10$ . However, even though it did not control the Type I error rate as desired for the other  $\alpha$  levels, it was the test with the fewest liberal cases. The other tests did not achieve an acceptable performance, as can be seen in Table 3.

### 3.2.4 Power (CS, AR(1))

In Figure 5, the following scenarios are presented for the multivariate normal,  $t_3$ , and uniform (0, 1) distributions: normal with  $p = 10$  and  $(n_1, n_2) = (20, 40)$  (I), normal with  $p = 300$  and  $(n_1, n_2) = (20, 40)$  (II),  $t_3$  with  $p = 10$  and  $(n_1, n_2) = (20, 40)$  (III),  $t_3$  with  $p = 300$  and  $(n_1, n_2) = (20, 40)$  (IV), uniform (0, 1) with  $p = 10$  and  $(n_1, n_2) = (20, 40)$  (V) and uniform (0, 1) with  $p = 300$  and  $(n_1, n_2) = (20, 40)$  (VI). In the normality scenarios, the tests in low dimensionality (I) can be divided into two groups regarding power values: the most powerful group (HTNT, AT, HTNTB, and ATB) and the less powerful group (HT and MNV). In the first group, as seen in 3.2.3, the best-performing test was AT, and in the second group, HT stood out. When moving to high dimensionality in the same distribution, it is noticeable that despite AT showing lower power, as observed in 3.2.3, it remained

the test with the best control over the Type I error rate.

Additionally, in Figure 5, we have the multivariate  $t_3$  distribution in (III) and (IV). In this case, the results differ slightly from the previous distribution, as the division into two groups of tests is not as clear. In low dimensionality (III), the test with the highest power was ATB, though this test showed more liberal cases than AT in other scenarios for this distribution. AT was the second most powerful test, but in general, the tests were less powerful in this distribution compared to the normal one, as observed in other cases. In high dimensionality, the pattern was more similar to the multivariate normal distribution, with conclusions similar to those drawn in (II).



Source: Author (2024).

**Figure 5.** Power and Type I Error Rates for Hotelling's  $T^2$  Test (TH), MNV, HTNT, Ahmad's Test (TA), HTNTB, and ATB, considering  $\alpha = 0.05$ , heteroscedasticity, multivariate normal,  $t_3$  and uniform  $(0, 1)$  distributions in low and high dimensionality cases.

Finally, the multivariate uniform  $(0, 1)$  distribution exhibited the same pattern as the multivariate normal distribution in both low and high dimensionality, but with two significant differences. The first is that power values were much higher, and the second is that there were very few scenarios in which the tests managed to control the Type I error rate.

## 4. Conclusions

The evaluation of the tests revealed clear differences in terms of Type I error rate control and power, particularly when varying the significance levels, dimensions, and assumptions of the tests.

The bootstrap tests, ATB and HTNTB, demonstrated superior performance in terms of robustness and consistency. ATB, in particular, stood out by controlling the Type I error rate in all scenarios, except in more extreme cases, such as when different covariance matrix structures were present. This makes it ideal for situations involving violations of classical assumptions and high dimensionality. In contrast, HTNTB, although conservative, exhibited lower power. Traditional tests, such as HT and MNV, showed greater power in ideal scenarios but revealed limitations in robustness, being more affected by heteroscedastic scenarios and high dimensionality, as they cannot be applied in such cases. The AT test presented a reasonable balance between Type I error rate control and power, but with some inconsistencies in non-ideal scenarios. Therefore, the choice of the test should take into account the specific application scenario, with ATB being the most versatile and robust option, while classical tests may be preferred in ideal, low-dimensionality situations where assumptions are met, and AT being suitable for more extreme cases in high dimensionality.

## Acknowledgments

The authors acknowledge the financial support of the CAPES and CNPq agencies and the support of IFMG Formiga *campus*.

## Conflicts of Interest

The authors declare no conflict of interest.

## Author Contributions

**Conceptualization:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Data curation:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Formal analysis:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Funding acquisition:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Investigation:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Methodology:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Project administration:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Software:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Resources:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Supervision:** FERREIRA, D. F. **Validation:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Visualization:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Writing – original draft:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.C.; FERREIRA, D. F. **Writing – review and editing:** CARVALHO NASCIMENTO, M.; BRAZ, L.H.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. Bennett, B. M. Note on a solution of generalized Behrens-Fisher problem. *Annals of the Institute of Statistical Mathematics* **2**, 87–90 (1951).
3. Dempster, A. P. A high dimensional two sample significance test. *The Annals of Mathematical Statistics* **29**, 995–1010 (1958).
4. Dempster, A. P. A significance test for the separation of two highly multivariate small samples. *Biometrics* **16**, 41–50 (1960).
5. Ferreira, D. F. *Estatística multivariada* 3rd ed., 622 (Editora UFPA, Lavras, 2018).

6. Gebert, D. M. P. Uma solução via bootstrap paramétrico para o problema de Behrens-Fisher multivariado. *Universidade Federal de Lavras*, 120 (2014).
7. 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).
8. James, G. S. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the populations variances are unknown. *Biometrika* **41**, 19–43 (1954).
9. Johansen, S. The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika* **67**, 85–92 (1980).
10. Krishnamoorthy, K. & Yu, J. Modified Nel and Van der Merwe test for the multivariate Behrens-Fisher problem. *Statistics and Probability Letters* **15**, 161–169 (2004).
11. Nel, D. G. & Van der Merwe, C. A. A solution to the multivariate Behrens-Fisher problem. *Communication in Statistics - Theory and Methods* **15**, 3719–3735 (1986).
12. Nishiyama, T, Hyodo, M & Seo, T. Recent developments of multivariate multiple comparisons among mean vectors. *SUT Journal of Mathematics* **50**, 247–270 (2014).
13. 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).
14. Silva, R. B. V., Ferreira, D. F. & Nogueira, D. A. Robustness of asymptotic and bootstrap tests for multivariate homogeneity of covariance matrices. *Ciência e Agrotecnologia* **32**, 157–166 (2008).
15. 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).
16. Yao, Y. An approximate degrees of freedom solution to the Behrens-Fisher problem. *Biometrika* **52**, 139–147 (1965).