MULTIVARIATE MULTIPLE COMPARISONS BY BOOTSTRAP AND PERMUTATION TESTS

Elisa Norberto Ferreira SANTOS¹
Daniel Furtado FERREIRA²

ABSTRACT: The present work proposes a multivariate multiple comparison procedure for multidimensional vectors using bootstrap and permutation. The procedure employs computationally intensive methods and is evaluated by Monte Carlo simulation. The idea consists of obtaining the distance $T^2$ of Hotelling for all pairs of means and comparing it to distributions of bootstrap or permutation distances obtained by resampling. The experimentwise type I error rates and the power of the proposed tests were computed in order to evaluate their performance in $N$ Monte Carlo simulations. The results of the simulations of proposed tests demonstrated good test size and power properties. The tests also showed robust power performance when the homoscedastic supposition was violated. The R code are available in the appendix.

KEYWORDS: Experimentwise type I error rates; multiple comparisons; bootstrap; permutation.

1 Introduction

Multivariate statistics describe a collection of procedures which involve analysis of several variables measured in the same experimental unity Mardia et al. (1979). When the data come from experiments with replications the primary purpose of the researchers is to test hypothesis of equality of the treatment effects considering all dimensions simultaneously.

In the univariate case, the analysis of variance (ANOVA) F-test is used to test the hypothesis of equality of treatment effects. If the null hypothesis is

¹Universidade Federal de Lavras – UFLA, Programa de Pós-Graduação em Estatística e Experimentação Agronômica, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil. E-mail: norelisa@hotmail.com

²Universidade Federal de Lavras – UFLA, Departamento de Ciências Exatas – DEX, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil. E-mail: danielff@dex.ufla.br. Bolsista CNPq.
rejected, multiple comparisons gives us a way of determining which pairs of means are significantly different. Multiple comparison tests are designed to allow the researchers to compare all possible pairs of group means using an appropriate and controlled significance level. These tests are usually performed after an ANOVA, where the null hypothesis that all population means are equal is rejected. Therefore, multiple comparison procedures are used to determine which means differ. The researchers have to choose the appropriate procedure for the comparisons of the treatment means for their particular circumstances (Hochberg, 1987; Hsu, 1996; Hinkelmann and Kempthorne, 2005).

On the other hand, in the multivariate case, researchers apply tests for the hypothesis of equality of treatment mean vectors, similar to the univariate F-test. These procedures give statistical judgment whether statistical significant difference exist among mean vectors. The first step is to apply a multivariate analysis of variance (MANOVA). There are several statistic tests for the null hypothesis (Rencher, 2002) of no differences in mean vectors for the different treatment groups analogous to the F-test in ANOVA. This test criterion are the Wilks likelihood ratio statistic, the Pillai trace statistic, the Lawley-Hotelling trace statistic and the Roy's greatest root statistics. If this overall null hypothesis is rejected by one of these tests, the next step is to conduct follow-up tests to identify the contrasts among treatment mean vectors that are responsible for the hypothesis rejection. There is little consensus to the proper methods for further identifying and interpreting group differences in the multivariate MANOVA problem. When is the interest to compare several mean vectors, but not the individuals means components within a vector, the linear contrast is commonly used. But the use of contrast is limited to cases where the treatment structure suggests a priori planned comparisons and the group levels are qualitative (Rencher, 2002).

The Bonferroni-type inequalities is the most commonly used procedure for multivariate multiple comparison, although this adjustment is very conservative in general. Kakizawa (2009) constructed asymptotically conservative 100(1 − α)% simultaneous confidence intervals among mean vectors for cases of comparisons with a control and pairwise comparisons, but the mathematical property of these procedures was not yet clarified theoretically, since the intervals were sometimes liberal. To overcome the complexity of the (exact or asymptotic) distribution theory of $T^{2}_{\text{max}}$-type statistics that appeared in his simultaneous confidence intervals of mean vectors improved Bonferroni-type inequalities were applied to construct asymptotically conservative simultaneous confidence intervals for pairwise comparisons as well as for comparisons with a control.

Imada and Yamamoto (2010) discussed the multiple comparison procedure with a control for a multivariate one-sided test in each pairwise comparison. This is a multivariate analog of the one-side multiple comparison procedure with a control and is applied to comparing several new treatments with a standard treatment simultaneously based on the multivariate normal response. They derived a formula to determine the critical value for pairwise comparisons satisfying a specified significance level. Then, they formulated the power of the test and gave some
numerical examples regarding the critical values and the power of the test and investigated their characteristics. Unfortunately, in this study, they assumed that the common covariance matrix was known, that is an unrealistic situation. All the above procedures may have restrictions and not fulfill the requirements of the researchers.

For these reasons, the present work proposes a multiple comparisons procedures for multidimensional vectors through bootstrap and permutation. The proposed procedures use intensive computational methods and their performance was evaluated by Monte Carlo simulations. The idea is based on obtaining Hotelling’s $T^2$ distances for all pairs of means and comparing them to the bootstrap and permutation distribution distances originated from a resampling procedure. The proposed procedure do not need that an initial omnibus test should be applied first. The experimentwise type I error rates and the power of the proposed tests were computed to evaluate their performance.

2 Proposed tests

Let $Y_{i1}, Y_{i2}, \ldots, Y_{ir}$ be a sample of size $r$ of random vectors of dimension $p \times 1$, from the $i$th non-specified multivariate distribution ($i = 1, 2, \ldots, k$) with mean vector $\mu_i$ and covariance matrix $\Sigma_i$. We propose a procedure of multivariate multiple comparisons as a follow-up test. Considering the family of $m = k(k-1)/2$ hypothesis

$$
\begin{cases}
H_0 : \quad \mu_i = \mu_{i'}, \quad i \neq i' = 1, 2, \ldots, k \\
H_1 : \quad \mu_i \neq \mu_{i'}, \quad \text{for some } i \neq i'
\end{cases}
$$

(1)

tests based on bootstrap and permutation were derived as follows. Initially, the estimates of the mean vector

$$\bar{Y}_i = \frac{1}{r} \sum_{j=1}^{r} Y_{ij}
$$

and the covariance matrix

$$S_i = \frac{1}{r-1} \left[ \sum_{j=1}^{r} Y_{ij}Y_{ij}^\top - \left( \sum_{j=1}^{r} Y_{ij} \right) \left( \sum_{j=1}^{r} Y_{ij} \right)^\top \right]$$

from the $i$th population, with $i = 1, 2, \ldots, k$, should be calculated.

As a sample measure of the difference of two vector means we considered by Mahalanobis distance between pairs of the estimated means of treatment. Thus the statistics of the new procedure were defined as follows

$$T^2_{ii'} = r(\bar{Y}_i - \bar{Y}_{i'})^\top (S_i + S_{i'})^{-1}(\bar{Y}_i - \bar{Y}_{i'}).$$

(2)
and should be calculated for all pair of sample mean vectors, \(i \neq i' = 1, 2, \ldots, k\).

A limitation of this procedure is that the inverse matrix \((S_i + S_{i'})^{-1}\) will only exist if \(p \leq 2(r - 1)\). A possibility to overcome the problem for those circumstances was the use of generalized inverses of \(S_i + S_{i'}\). Another possibility is the use of pooled covariance matrix estimator \(S_p\) from the \(k\) populations, given by

\[
S_p = \frac{\sum_{i=1}^{k} (r - 1)S_i}{k(r-1)} = \frac{\sum_{i=1}^{k} S_i}{k}.
\]

So, for the circumstances where \(p > 2(r - 1)\) we adopted the alternative expression to the test statistic given by

\[
T_{ii'}^2 = \frac{r}{2}(\overline{Y}_i - \overline{Y}_{i'})^\top S_p^{-1}(\overline{Y}_i - \overline{Y}_{i'}),
\]

for \(i \neq i' = 1, 2, \ldots, k\).

In the multivariate non-normality case or even under normality, the joint distribution of \(T_{ii'}^2\) for all of pairs of mean vectors is unknown. We consider the possibility to obtain the sampling null distribution \(T_{ii'}^2\) by means of bootstrap and permutation. The imposition of null hypothesis of equality of the \(k\) population mean vectors \((H_0 : \mu_1 = \mu_2 = \cdots = \mu_k)\) was made by gathering the \(rk\) multivariate observations in a single group. This combined sample was resampled with and without replacement, recreating the original sample structure of the \(k\) populations with \(r p\)-variate observations.

For the \(m\)th pooled sample the set \(\Omega\) was formed of all pairs of Hotelling’s \(T^2\) resampling distances and is given by

\[
\Omega = \{T_{12}^2, T_{13}^2, \ldots, T_{1km}^2; T_{23}^2, \ldots, T_{(k-1)km}^2\}
\]

and the maximum, defined as follows

\[
\Gamma_m = \max_{i \neq i'} \Omega = \max_{i \neq i'} \{T_{12}^2, T_{13}^2, \ldots, T_{1km}^2; T_{23}^2, \ldots, T_{(k-1)km}^2\},
\]

was retained.

This procedure was repeated \(B\) times. A set \(\Gamma\) of the \(B\) maximum values retained at each resampling step was formed by

\[
\Gamma = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_m, \ldots, \Gamma_B\}.
\]

The null hypothesis test could be finally concluded estimating the \(p\)-value by

\[
p_{ii'} - \text{value} = \frac{\sum_{m=1}^{B} I(\Gamma_m \geq T_{ii'}^2)}{B},
\]

384

where $I(\Gamma_m \geq T^2_{ii'})$ is the indicator function that has the value 1 if the inequality is verified and 0, otherwise and $T^2_{ii'}$ is the original sample distance between the sample mean from (3). The observed $p$-value of each pair of populations mean vectors was compared to the nominal significance level $\alpha$ to complete the null hypothesis test. If the $p$-values are smaller than or equal to the nominal significance level, the null hypotheses should be rejected.

A real example considering $k = 6$ land use systems (LUS) from Amazônia and $r = 10$ replications, characterized by Nobrega (2006) was used to illustrate the test. The LUS used were the forest (F), old secondary forest (OSF), young secondary forest (YSF), agroforestry (AF), pasture (P), agriculture (A). Two variates related to soil texture were considered in this example: sand and clay.

In the next section we presented a procedure to evaluate the proposed tests based on Monte Carlo simulation methods. It should be clear that two similar tests were constructed. The difference between the two procedures is only the resampling scheme that can be done with or without replacement. The idea behind these newer tests uses the fact that distances close to zero indicate no differences and larger distances of the extreme 100% upper tail of the resampling distance distribution indicate significant differences between the two mean vectors. For accounting for multiplicity we consider the distribution of the maximum distance, inspired by the studentized range distribution largely used in univariate multiple comparison procedures.

### 3 Monte Carlo evaluation

Two cases were considered to evaluate the properties of the two tests. The first aimed to verify the experimentwise type I error rates under the null hypothesis. The second evaluated the power of the proposed tests under the alternative hypothesis. Under the null hypothesis of equality of mean vectors ($\mu_1 = \mu_2 = \cdots = \mu_k = \mu$) an amount of $N$ Monte Carlo simulations were accomplished to evaluate the experimentwise type I error rates of the proposed tests. We considered samples of size $r$ of each multivariate normal population, obtaining in any Monte Carlo simulation the pooled sample $Y_{11}, Y_{12}, \ldots, Y_{1r}, Y_{21}, \ldots, Y_{r1}, \ldots, Y_{kr}$.

Every vector was simulated without loss of generality from normal multivariate distributions with common mean vector $\mu$, i.e., $Y_{ij} \sim N_p(\mu, \Sigma)$ where $p$ is the dimension (number of variables) and $\Sigma$, the $ith$ population covariance matrix $(p \times p)$ that we assume to be definite positive.

To accomplish the simulations we considered the cases of homoscedastic and heteroscedastic population covariance matrix patterns. In the first circumstance, the population covariance matrices were all equal to $\Sigma$, considering a compound symmetry structure, where $\Sigma = \sigma^2[(1 - \rho)I + \rho J]$ where $I$ is a $p \times p$ identity matrix, $J$ is a $p \times p$ unit matrix, $\sigma^2$ is the residual variance and $\rho$ is the population correlation coefficient.

For the heteroscedastic cases, the first population covariance matrix $\Sigma_1$ was settled in the same way as done for the homoscedastic case considering the
compound symmetry structure. The other population covariance matrices were established by

$$\Sigma_i = \left[1 + \left(\frac{\sqrt{\delta} - 1}{k - 1}\right)(i - 1)^{k - 1}\right] \Sigma_1,$$  \hspace{1cm} (7)

where $\delta$ is the degree of heterogeneity between two covariance matrices, defined by the ratio of the largest and smallest generalized variances given by $\delta = |\Sigma_k|/|\Sigma_1|$, for $i = 2, 3, \ldots, k$. The covariance heterogeneity increases from 1 to $\delta$ as $i$ increases from 1 to $k$. When $i = k$, then $\Sigma_k = \sqrt{\delta} \Sigma_1$ and therefore, the ratio of their determinants will be equal to $\delta$.

In the simulations we considered all combinations of the number of replications $r$, the number of populations $k$ and the number of variates $p$, where the chosen values were $r = 3, 10, 20$, $k = 5, 10, 50, 100$, and $p = 2, 5, 10$. Different values of the population correlation coefficient for the compound symmetry structure were considered, but the results did not differ, only the case of $\rho = 0.5$ was showed, since the test statistic is invariant under linear transformation (Mardia et al., 1979). The value of the residual variance $\sigma^2$ was settled for 10 and the population mean vector $\mu$ was settled for the null vector $0$, without loss of generality. The $\delta$ values that specify the heterogeneity degrees of the covariance matrices were also settled for 1, 2, 8 and 16. When $\delta = 1$ we could simulate the homoscedastic cases. We considered at last $N = 1,000$ Monte Carlo simulations and $B = 1,000$ bootstrap or permutation resamples.

The experimentwise type I error rates were computed obtaining the proportions of the $N$ simulated experiments under the null hypothesis that presented at least one difference between mean vectors significantly ($p$-value < $\alpha$) different from zero. The $p$-values were obtained through the equation (6) and the nominal significance levels $\alpha$ were 1% and 5%. The observed levels of significance where tested against the nominal level, i.e., $H_0 : \alpha = \alpha_0$, where $\alpha_0 = 5\%$ or 1%, using an exact binomial test.

Under the alternative hypothesis $H_1$, the simulations were accomplished to evaluate the power of the tests. The simulations were done following the procedures described previously for the type I error, except by the fact of $Y_{ij}$ be generated from normal multivariate distributions with different mean vectors, i.e., $Y_{ij} \sim N_p(\mu_i, \Sigma_i)$. The population mean vector $\mu_1$ was settled as the null vector and the other mean vectors $\mu_i$, $i = 2, 3, \ldots, k$, was $\mu_i = \mu_{i-1} + \psi \sqrt{\text{diag} (\Sigma_{i-1} + \Sigma_1)}/r$, where $\psi = 1, 2, 4, 8$ and 16. The $\text{diag}$ operator returns the vector corresponding to the main diagonal of the matrix used as its argument. The square root operator should be applied to all components of the vector used as its arguments.

There are other cases worth examining as, for instance, only one variate show mean differences, only one treatment differs from the rest, but on all variates, only one treatment mean in one variate differs, an so on. It is not possible to examine every potential situations, but our method could deal with them easily.
The power values were computed by the proportions of the $N \sim$ simulated experiments under the alternative hypothesis that presented at least one significantly ($p$-value $< \alpha$) difference between mean vectors different from zero, but considering only all pairs whose parametric differences between two means vectors were $\psi$. Therefore, we considered only $k-1$ pairs of vector means in each simulation. The R code are available in the appendix A.

4 Type I error rates

The observed experimentwise type I error rates were compared to the nominal significance levels using the binomial test with confidence coefficient of 99% to accomplish the effect of Monte Carlo errors. In the $k=5$ case, we used 2,000 Monte Carlo simulations and in the remaining ones, 1,000 simulations, as a consequence of the high number of days spent at simulations. Considering the different values of $k$, the null hypotheses $H_0: \alpha = 1\%$ and $H_0: \alpha = 5\%$ were tested in each simulation performed considering a 1\% significance level.

For $k=5$ (Table 1) the proposed bootstrap test for multivariate multiple comparisons controlled the experimentwise type I error rates in all cases in which $p < 2(r-1)$ with test size equal or significantly ($p$-value $< 0.01$) less than the nominal significance level of 1\% ($\alpha = 1\%$). For $\alpha = 5\%$, we observed the same performance, except for the cases $r = 10$, $p = 2$ with $\delta = 8, 16$, with real test size significantly ($p$-value $< 0.01$) higher than the nominal significance level, even though these rates were not expressively different from the nominal level. The tests were liberal when the assumption that $p \leq 2(r-1)$ did not hold, for both $\alpha = 1\%$ or $\alpha = 5\%$. In this case, the use of a generalized inverse of $S_i + S_j$ must be made to circumvent the problem of estimating the distances between two mean vectors by (2). Type I error rates larger than the nominal levels were observed as consequence. The strategy of using the expression (3) for the distances was worse (results not shown), since the experimentwise type I error rates were higher when $p > 2(r-1)$ and especially when $p > k(r-1)$.

If $p > k(r-1)$ the bootstrap test is not recommended for multivariate multiple comparisons, since the results were liberal in almost all circumstances. The type I error rates increase when the number of variates $p$ departs from the maximum allowed $k(r-1)$.

To solve the problem in the cases in which $p \leq k(r-1)$, but it remains close to the second term of the inequality, the multivariate multiple comparison permutation test was proposed, where the resamples are made without replacement. The expression (3) was used for all cases. The results were slightly better, for the case where $p < 2(r-1)$ and expression (2) was used to compute the distances between mean vectors. In Table 2 the results for experimentwise type I error rates for the permutation test using expression (3) were considered excellent, because almost all observed rates were not significantly different from the nominal significance levels. There were only two exceptions, but they did not differ expressively from the nominal significance levels and the reason might be due to the lack of multiplicity adjustments. This indicates that the test has an exact size.
Table 1 - Experimentwise type I error rates in percentage as a function of number of replications $r$, number of variables $p$ and heterogeneity of covariance matrices $\delta$ considering the multivariate normal distribution, estimated from $B = 1,000$ resamples with replacement and from $N = 2,000$ Monte Carlo simulations, for $k = 5$ populations

<table>
<thead>
<tr>
<th>$r$</th>
<th>$p$</th>
<th>$\delta$</th>
<th>1%</th>
<th>5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.00**</td>
<td>0.70**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.05**</td>
<td>1.40**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.00**</td>
<td>0.70**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.00**</td>
<td>0.85**</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2.95**</td>
<td>15.10**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.35**</td>
<td>14.90**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>3.60**</td>
<td>14.50**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>2.80**</td>
<td>14.30**</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2.45**</td>
<td>10.90**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.80**</td>
<td>11.45**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2.05**</td>
<td>11.00**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>2.20**</td>
<td>11.25**</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.85</td>
<td>4.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.80</td>
<td>4.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1.45</td>
<td>6.40**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>1.50</td>
<td>6.85**</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1.15</td>
<td>5.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.15</td>
<td>5.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>1.25</td>
<td>5.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>1.20</td>
<td>6.00</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.50**</td>
<td>4.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.45**</td>
<td>4.55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.30**</td>
<td>4.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.45**</td>
<td>4.50</td>
<td></td>
</tr>
</tbody>
</table>

** and ** significantly ($p$-value < 0.01) lower and higher, respectively, to the nominal significance levels of $\alpha = 5\%$ and $\alpha = 1\%$.

In Tables 3 and 4 we show the experimentwise type I error rates for $k = 50$ and $k = 100$ populations, respectively, for the bootstrap test. Similar response patterns showed for the $k = 5$ populations were observed. In the homoscedastic circumstances ($\delta = 1$), the bootstrap test was exact or conservative ($k = 50$, $r = 10$, $p = 2$ and $\alpha = 1\%$). This case ($\delta = 1$) is considered ideal for the test application in multiple comparisons. With the increasing of the covariance heteroscedastic, there was an increasing of liberal results. In the circumstances of $\delta \geq 8$, the bootstrap test was liberal, except for $k = 50$, $r = 10$, $p = 5$ and $\delta = 8$, for all $\alpha$ considered.
Table 2 - Experimentwise type I error rates in percentage as a function of number of replications \( r \), number of variables \( p \) and heterogeneity of covariance matrices \( \delta \) considering the multivariate normal distribution, estimated from \( B = 1,000 \) resamples without replacement and from \( N = 1,000 \) Monte Carlo simulations, for \( k = 5 \) populations

<table>
<thead>
<tr>
<th>( r )</th>
<th>( p )</th>
<th>( \delta )</th>
<th>( \alpha ) 1%</th>
<th>( \alpha ) 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>2</td>
<td>1.50</td>
<td>5.80</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>16</td>
<td>1.20</td>
<td>6.10</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>2</td>
<td>2.20**</td>
<td>6.50</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0.80</td>
<td>4.90</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>16</td>
<td>1.50</td>
<td>5.90</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>8</td>
<td>1.00</td>
<td>5.50</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>16</td>
<td>0.60</td>
<td>6.10</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>1</td>
<td>1.10</td>
<td>4.80</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>2</td>
<td>0.80</td>
<td>5.60</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>16</td>
<td>1.00</td>
<td>5.30</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>16</td>
<td>1.20</td>
<td>6.20</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>2</td>
<td>1.20</td>
<td>4.90</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>16</td>
<td>1.60</td>
<td>6.00</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>16</td>
<td>2.10</td>
<td>7.00</td>
</tr>
</tbody>
</table>

** significantly (\( p \)-value < 0.01) higher to the nominal significance levels of \( \alpha = 5\% \) and \( \alpha = 1\% \).

In the other simulated cases, \( k = 50 \) and \( k = 100 \), (Tables 3 and 4), the response pattern was very similar to those presented for \( k = 5 \). Its important to emphasize that the experimentwise type I error rates of the bootstrap test, when it was considered liberal, tend to increase as the number of populations \( k \) increases. In this liberal circumstance and also in the heteroscedastic covariance structure, another aspect to be considered was the tendency of decreasing the type I error rates as the number of variates increases. Though these results have been observed, it is interesting to emphasize this performance behavior pattern is not very usual, but sounds as good property of the proposed tests.
Table 3 - Experimentwise type I error rates in percentage as a function of number of replications \( r \); number of variables \( p \) and heterogeneity of covariance matrices \( \delta \) considering the multivariate normal distribution, estimated from \( B = 1,000 \) resamples with replacement with the pooled covariance matrices estimates in the test statistic and from \( N = 1,000 \) Monte Carlo simulations, for \( k = 50 \) populations.

<table>
<thead>
<tr>
<th>( r )</th>
<th>( p )</th>
<th>( \delta )</th>
<th>( \alpha = 1% )</th>
<th>( \alpha = 5% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.90</td>
<td>5.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.20</td>
<td>5.80</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>2.00**</td>
<td>8.40**</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>3.60**</td>
<td>11.30**</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.50††</td>
<td>4.20</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.00</td>
<td>4.20</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>3.00**</td>
<td>11.30**</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>3.30**</td>
<td>10.60**</td>
<td></td>
</tr>
</tbody>
</table>

++ and ** significantly (\( p \)-value < 0.01) lower and higher, respectively, to the nominal significance levels of \( \alpha = 5\% \) and \( \alpha = 1\% \).

Table 4 - Experimentwise type I error rates in percentage as a function of number of replications \( r \); number of variables \( p \) and heterogeneity of covariance matrices \( \delta \) considering the multivariate normal distribution, estimated from \( B = 1,000 \) resamples with replacement with the pooled covariance matrices estimates in the test statistic and from \( N = 1,000 \) Monte Carlo simulations, for \( k = 100 \) populations.

<table>
<thead>
<tr>
<th>( r )</th>
<th>( p )</th>
<th>( \delta )</th>
<th>( \alpha = 1% )</th>
<th>( \alpha = 5% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.10</td>
<td>5.50</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.50</td>
<td>7.20</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>2.80**</td>
<td>10.30**</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>2.80**</td>
<td>12.20**</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.00</td>
<td>5.70</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.20</td>
<td>6.10</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>4.90**</td>
<td>13.10**</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>4.20**</td>
<td>14.60**</td>
<td></td>
</tr>
</tbody>
</table>

** significantly (\( p \)-value < 0.01) higher to the nominal significance levels of \( \alpha = 5\% \) and \( \alpha = 1\% \).

In Table 5 the experimentwise type I error rates were shown to the permutation test for the same case showed in Table 4 for the bootstrap test. The results were very similar. Therefore, the two test are very close when there are no problems with respect to the inverse of the pooled covariance matrix or the number of variables \( p \) is smaller when compared with the population sample sizes. These cases occur when \( p \leq k(r - 1) \). For the cases where this condition is achieved, but \( p \) is in the neighborhood of this inequality, the permutation test has better control of the type
I error. It is worth to emphasize that the two test have very similar performance in large sample sizes and should be considered equivalent.

Table 5 - Experimentwise type I error rates in percentage as a function of number of replications \( r \); number of variables \( p \) and heterogeneity of covariance matrices \( \delta \) considering the multivariate normal distribution, estimated from \( B = 1,000 \) resamples without replacement (permutation test) with the pooled covariance matrices estimates in the test statistic and from \( N = 1,000 \) Monte Carlo simulations, for \( k = 100 \) populations

<table>
<thead>
<tr>
<th>( r )</th>
<th>( p )</th>
<th>( \delta )</th>
<th>( 1% )</th>
<th>( 5% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1.09</td>
<td>5.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1.49</td>
<td>7.22**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>2.81**</td>
<td>10.32**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>2.80**</td>
<td>11.51**</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1.02</td>
<td>5.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1.18</td>
<td>6.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>4.92**</td>
<td>12.99**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>4.19**</td>
<td>13.65**</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1.08</td>
<td>5.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1.11</td>
<td>5.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>2.39**</td>
<td>7.29**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>2.38**</td>
<td>8.51**</td>
</tr>
</tbody>
</table>

** significantly \((p\text{-value} < 0.01)\) higher to the nominal significance levels of \( \alpha = 5\% \) and \( \alpha = 1\% \).

5 Power

Considering the nominal significance level of 5% (dotted line) and 1% (full line) (Figure 1) and \( k = 5 \) we verify that there was a greater power superiority for \( \alpha = 5\% \) than for \( \alpha = 1\% \), as expected, while the differences among the mean vectors increased. For both levels, under \( H_0 \) (Table 1), the test was extremely conservative. This partially explains that lower power found.

In Figure 2 the values of power in function of \( \psi \) and \( \alpha \) for \( k = 100 \) are presented, with \( r = 10 \) and \( \delta = 1 \) and 16. The number of populations had large effect in the power of the multivariate multiple comparison bootstrap test. The values of power found in the present work in the multivariate context are comparable to the ones found by Silva (1998) for univariate Scott and Knott test (Scott and Knott, 1974) and higher than those of Ramos (Ramos, 2007) for the univariate original and bootstrap version of Calinski and Corsten test.

In Figure 3 the values of power of the permutation test as a function of \( \psi \) and \( \alpha \) for \( k = 100 \) are presented, with \( r = 10 \) and \( \delta = 1 \) and 16. This case use
Figure 1 - Power of the multivariate multiple comparison bootstrap test in function of the differences $\psi$ between two population means and of the nominal significance level $\alpha$, for $k = 5$, $r = 3$, $\rho = 0.5$, $p = 2$ and (a) $\delta = 1$ and (b) $\delta = 16$. 
Figure 2 - Power of the multivariate multiple comparison bootstrap test in function of the differences $\psi$ between the two population means and of the nominal significance level $\alpha$, for $k = 100$, $r = 10$, $\rho = 0.5$, $p = 5$ and (a) $\delta = 1$ and (b) $\delta = 16$. 

Figure 3 - Power of the multivariate multiple comparison permutation test in function of the differences $\psi$ between the two population means and of the nominal significance level $\alpha$, for $k = 100$, $r = 10$, $\rho = 0.5$, $p = 5$ and (a) $\delta = 1$ and (b) $\delta = 16$. 
the same configuration of that of Figura 2. When the results of both tests were compared, we can conclude that the observed differences are very small and they were attributed to the Monte Carlo errors. The two figures cannot be differentiated visually. Therefore, for large samples where the number of variables $p$ is smaller than the number of populations $k$ or the sample sizes $r$, the two test can be considered equivalent, since they showed almost the same type I error rates and power. Many other simulations of parameters configurations showed response patterns similar to those presented and they were omitted. Unfortunately there are no results of multivariate multiple comparison procedures to compare with our results.

6 Real Example

A real example considering $k = 6$ systems of land use systems (LUS) from Amazônia, $r = 10$ replications, and $p = 2$ variates to soil texture, sand and clay, were considered to illustrate the multivariate multiple comparison bootstrap test proposed in this article. In Table 6 the results of the bootstrap test statistics between the LUS (above diagonal) and the respectively $p$-values (below diagonal) were shown.

Table 6 - Values of the multivariate multiple comparison bootstrap test statistic (above diagonal) and $p$-values (below diagonal) between LUS mean vectors

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>AF</th>
<th>YSF</th>
<th>OSF</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00</td>
<td>7.35</td>
<td>6.22</td>
<td>0.02</td>
<td>0.02</td>
<td>12.23</td>
</tr>
<tr>
<td>AF</td>
<td>0.27</td>
<td>0.00</td>
<td>1.66</td>
<td>7.15</td>
<td>6.55</td>
<td>7.08</td>
</tr>
<tr>
<td>YSF</td>
<td>0.38</td>
<td>0.97</td>
<td>0.00</td>
<td>5.74</td>
<td>5.65</td>
<td>14.54</td>
</tr>
<tr>
<td>OSF</td>
<td>1.00</td>
<td>0.28</td>
<td>0.44</td>
<td>0.00</td>
<td>0.04</td>
<td>12.82</td>
</tr>
<tr>
<td>F</td>
<td>1.00</td>
<td>0.34</td>
<td>0.45</td>
<td>1.00</td>
<td>0.00</td>
<td>11.39</td>
</tr>
<tr>
<td>P</td>
<td>0.04</td>
<td>0.29</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Considering the nominal significance level $\alpha$ of 5%, we verified that the A and P, YSF and P, and OSF and P land use systems showed mean vectors significantly ($p$-value $< 0.05$) different among themselves considering the bootstrap test. Similar results was observed in the permutation test (results not shown). The LUS pasture was present in all three pairs whose mean vectors differed significantly ($p$-value $< 0.05$). This result corroborates the results found by Nobrega (2006) using other multivariate techniques, such cluster analysis.

Conclusions

In the homoscedastic circumstances the tests had optimal performance with respect to the experimental type I error rates that were lower or identical to the nominal significance levels when $p \leq k(r - 1)$. The tests were liberal in
heteroscedastic cases. Multivariate methods for variates selection like principal components are recommended in the case in which $p > k(r-1)$, but close to $k(r-1)$ and both tests showed similar power and type I error rates for larger values of $k$ and $r$. The permutation test has better performance on the control of experimentwise type I error when $p \leq k(r-1)$. The multivariate multiple comparison bootstrap and permutation tests have power comparable to or higher than the univariate multiple comparison tests. The tests were robust concerning the power when the assumption of homoscedastic covariances is violated. The bootstrap and the permutation tests can be considered equivalents for larger samples ($k$ and $r$).

Acknowledgements

The work reported here has been developed under the financial support of the CNPq. We would like to thank for technical support to Vladimir Vietoris from Slovak Agricultural University of Nitra, Slovakia.


RESUMO: O presente trabalho foi realizado com o objetivo de propor um procedimento de comparação múltipla para vetores multidimensionais via bootstrap e permutação. O procedimento foi implementado por meio de métodos de computação intensiva e avaliado por meio de simulação Monte Carlo. A idéia foi obter para todos os pares de médias a distância $T^2$ de Hotelling e compará-las com as distâncias obtidas na distribuição bootstrap e permutação, originada da reamostragem. As taxas de erro tipo I por experimento e o poder do teste proposto foram computados para avaliar o seu desempenho em $N = 1000$ simulações Monte Carlo. Pode-se concluir que os testes foram propostos com sucesso e os passos necessários para suas execuções foram implementados em R (funções no apêndice); os testes mostraram robustez no desempenho do poder quando a suposição de homocedasticidade é violada.

PALAVRAS-CHAVE: Erro tipo I por experimento; comparações múltiplas; bootstrap; permutação.

References


Received in 13.08.2012.

Approved after revised in 12.02.2013.
# auxiliary function to summarize the data
dados.sumario <- function(dados,r,k)
{
  ybar <- apply(dados[1:r,2:(p+1)],2,mean)
covi <- var(dados[1:r,2:(p+1)])
sumar <- cbind(ybar,covi)
  for (i in 2:k)
  {
    ybar <- apply(dados[((i-1)*r+1):(i*r),2:(p+1)],2,mean)
covi <- var(dados[((i-1)*r+1):(i*r),2:(p+1)])
sumari <- cbind(ybar,covi)
  sumar <- rbind(sumar,sumari)
  }
return(sumar)
}

# Computes T2 k x k matrix with T^2_{i,l} statistics.
T2il <- function(sumar,r,k,p)
{
  T2 <- matrix(0,k,k)
  Yb <- matrix(sumar[,1],k,p,byrow=TRUE)
  Aux <- matrix(t(sumar[,2:(p+1)]),k,p*p,byrow=TRUE)
  Sp <- matrix(apply(Aux,2,sum),p,p)/k
  G<-r/2*(Yb%*%ginv(Sp)%*%t(Yb))
  for (i in 1:(k-1))
  {for (ell in (i+1):k)
    {T2[i,ell] <- G[i,i]-2*G[i,ell]+G[ell,ell]
     T2[ell,i] <- T2[i,ell]
    }
  return(list(T2=T2,Sp=Sp))
  }

# Resample with and without replacement
reamostragem <- function(dados,r,k,boo)
{
  ii <- sample(1:(r*k), r*k, replace = boo)
  return(dados[ii])
}

# Computes the p-values and put in the
# lower triangular matrix $T^2$. The upper cases still contain the $T^2$ values.
valorp <- function(T2,Gamab,k,B)
{
  for (ii in 1:(k-1))
  {
    for (kk in (ii+1):k)
    {
      T2[kk,ii] <- length(Gamab[Gamab >= T2[ii,kk]])/B
    }
  }
  return(T2)
}

# Computes the bootstrap or permutation samples. If boo=TRUE then permutation distribution is returned, otherwise the bootstrap samples are obtained.
boot <- function(dados,r,k,p,B,boo)
{
  dadosb <- reamostragem(dados,r,k,boo)
  sumarb <- dados.sumario(dadosb,r,k)
  T2b <- T2il(sumarb,r,k,p)
  Gama <- max(T2b$T2)
  for (b in 2:B)
  {
    dadosb <- reamostragem(dados,r,k,boo)
    sumarb <- dados.sumario(dadosb,r,k)
    T2b <- T2il(sumarb,r,k,p)
    Gammab <- max(T2b$T2)
    Gama <- rbind(Gama,Gammab)
  }
  return(Gama)
}

r<-10;k<-6;p<-2 # Parameters of the data set
boo <- FALSE # FALSE for permutation and TRUE for bootstrap test
B<-10000 # number of resamples with or without replacement
# working diretory for the dataset -where the data set are setwd("C:/user")
# Data set. First column with the population identification: # 1, 2, ...k. Other columns with the variables. The first row # can represent the variable identifications or not.
dados <- read.table("DadosRafa.txt",header=T)
sumar <- dados.sumario(dados,r,k) # summary for other functions
T2 <- T2il(sumar,r,k,p) # T^2 statistics
Gamab <- boot(dados,r,k,p,B,boo) # bootstrap or permutation distributions
T2 <- valorp(T2$T2,Gamab,k,B) # T2 above diagonal and p-values below
T2 # Above diagonal-T2 Statistics; below - p-values
hist(Gamab) # histogram of the distribution
# density values
plot(density(Gamab),main="",xlab=expression(Gamma[m]^"*"),ylab="Density")