Acta Scientiarum
http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v42i1.44456

Comparison of tests on covariance structures of normal populations

Isabella Marianne Costa Campos, Denismar Alves Nogueira*, Eric Batista Ferreira and Davi Butturi Gomes

Departamento de Estatística, Universidade Federal de Alfenas, Rua Gabriel Monteiro da Silva, 700, 37130-001, Alfenas, Minas Gerais, Brasil. *Author for correspondence. E-mail: denismar.nogueira@unifal-mg.edu.br

ABSTRACT. In some studies, there is interest in testing the variance structure, as in the context of multivariate or modelling techniques. Therefore, the importance of using hypothesis tests on covariance structures is emphasized. The purpose of this study was to perform a detailed performance study regarding the power and type I error rate of some existing identity and sphericity tests, considering the scenarios with different numbers of variables (2 to 64) and sample sizes (5 to 100). The proposal of Ledoit and Wolf (2002) is the most appropriate to test the identity structure. For the sphericity test, the version of John (1972), modified by Ledoit and Wolf (2002), followed by the proposal of Box (1949), were the ones with the best performance.

Keywords: power; type i error rate; likelihood ratio; monte carlo simulation.

Received on September 5, 2018.
Accepted on April 2, 2019.

Introduction

In data analysis, for the univariate or multivariate analysis to be applied more effectively, some assumptions must hold. Some of these assumptions refer to the covariance structure. In order to ensure the assumptions about covariance, it is necessary to perform some hypothesis test capable of evaluating the structure of the matrix under study, which can be presented in different formats. When you have some hypothesis testing options available, you need to choose the most appropriate test, which allows for greater confidence in the results. The selection of the test is made with the knowledge of the performance, that is, of the control of the type I error rate and power. The presence of such a detailed and descriptive study of the behaviour of the tests of interest, considering its importance, was not verified in the specific literature. This study aimed to carry out performance in relation to the power and type I error rate of the existing identity and sphericity tests, in addition to classifying the tests as conservative, exact or liberal.

The use of a test with inadequate power and type I error rate affects decision-making and, for example, dependence or independence between variables, the applicability of multivariate techniques, and adequate modelling of covariance structure, among others. The importance of knowledge on type I error and power of any hypothesis test are directly related to success in decision-making. According to Cantelmo and Ferreira (2007), a perfect hypothesis test is one that never rejects a true null hypothesis and always rejects a false null hypothesis, a situation considered unreal. In practice, an ideal test is one that has type I error rate close to the level of significance adopted and power approaching 100%. For this purpose, the particular cases of the likelihood ratio tests, the identity and sphericity tests for a normally distributed population were used. These tests were compared considering different simulated scenarios.

Methods such as regression require the assumption for the covariance matrix when independence is assumed. The F-test requires a structure in the form of composite symmetry or HF to be valid, (Huynh & Feldt, 1976; Littell, Henry, & Ammerman, 1998). The time series can be adjusted by regression models (Kedem & Fokianos, 2005) and the dependence controlled by the covariance matrix (uncorrelated waste), several structures can occur in these series. Generalized and mixed linear models expand this application allowing the use of other covariance structures. In the study of (Gouveia, Silva, Ferreira, Gadelha, & Lima Filho, 2015), it intended to estimate the volume of *Eucalyptus* clones using mixed models, the heterogeneous first order autoregressive structure was adopted. Xavier and Dias (2001) verified through the observation of cases in which the covariance matrix satisfies or not the sphericity condition. If the matrix does not meet the sphericity condition, corrections should be used for the degrees of freedom of intra-

individual factors, or else, opt for multivariate or mixed models. In the work of Alves, Tavares, Tavares, Lobato and Oliveira (2015), we can mention another application in agronomy of the use of covariance structures that, with the objective of plant genetic improvement, evaluated the vegetative development of 25 progenies of *cupuaçu* trees, through the analysis of repeated measures, verifying, through the sphericity test, which type of statistical analysis is most appropriate (time-subdivided plots or mixed models). Regarding the variable plant height, because they are repeated measures, it may not meet the independence and, therefore, it does not pass in the criterion of HF.

According to Gouvêa, Prearo and Romeiro (2012), there is a growing use of multivariate methods; as well as misuses of these methods, either in the inadequacy of the objectives of use of the tools or regarding the violation of their premises. Among the multivariate techniques that have some assumption about the covariance structure, for the evaluation of interdependence, we cite the technique of factorial analysis that according to Fávero, Belfiore, Silva, and Chan (2009), seeks to synthesize, reduce data and create indicators or factors. This method requires non-independent variables. Another technique used to evaluate interdependence is the Principal Component Analysis (PCA), which, according to Santo (2012), is used to reduce the data that allow identifying patterns. The hypothesis to be tested in the covariance matrix is independence.

The Multivariate Analysis of Variance (Manova), according to Reis (1997), presents difference for the univariate case (ANOVA) whereas it evaluates the differences of the means of groups only for the response variable. In the univariate case, it demands structures of the type Composite Symmetry, Spherical or HF. According to Fávero, Belfiore, Takamatsu, and Suzart (2017), in addition to the premise of multivariate normality of the dependent variables, Manova presupposes equality of its covariance matrices and requires structure of the type Composite Symmetry or Spherical.

The choice of the appropriate structure, according to (Gouveia et al., 2015), directly affects the parameter estimates and the standard errors of fixed and random effects. The structures of interest of this study are: Structure of Identity: For this case, it is assumed that the variables are independent and that the variances are homogeneous equal to one. Structure of Sphericity: For this case, it is assumed that the variables are independent and that the variances are homogeneous. According to Cecon et al. (2008), the spherical structure imposes equal variances on all occasions of measurements and independent observations with a single parameter. Structure of Independence: For this case, it is assumed that the variables are independent and that the variances are homogeneous or heterogeneous.

According to Casella and Berger (2010), a hypothesis is a statement about a parameter of the population and has interest in carrying out statistical tests on this statement. The purpose of any hypothesis test is to decide, based on a population sample, which of two complementary hypotheses is true.

When deciding whether to accept or reject the null hypothesis H_0, an experimenter may be making a mistake. Generally, hypothesis tests are evaluated and compared through their probabilities of resulting in errors. A hypothesis test of H_0: $\theta \in \Theta_0$ *versus* H_1: $\theta \in \Theta_0^c$ can result in one of two types of errors. Traditionally, these two types of errors are named type I error and type II error. If $\theta \in \Theta_0$, but the hypothesis test incorrectly decides to reject H_0, then the test made a type I error. If, on the other hand, $\theta \in \Theta_0^c$, but the test decides not to reject H_0, a type II error has occurred. Following a hypothesis test and comparing the type I error rate with the pre-established level of significance, according to Biase and Ferreira (2011), one can name them as conservative, liberal or exact. A test is considered conservative when the type I error rate is lower than the level of significance adopted, considered liberal when it is higher and exact when the type I error rate coincides with the level of significance adopted. Doria Filho (1999) defines it as the probability of correctly rejecting H_0, given that it is really false, that is, it is related to the ability of the test to identify differences.

The tests of interest in this work are all based on the principle of likelihood ratio that compares the likelihood function under null hypothesis with the likelihood function along the whole parameter space (the alternative hypothesis), assuming multivariate normality. The likelihood quotient, in general, is represented by the Λ statistic. Distribution of Λ statistic is often unknown, but for large samples and under very general conditions, Wald (1945) showed that $-2\log(\Lambda)$ converges to a chi-square distribution under the true null hypothesis. The degrees of freedom, denoted by *f,* are obtained by subtracting the number of independent parameters for the whole parametric space minus the number of independent parameters under the null hypothesis. The increase or decrease in the number of parameters directly affects the degrees of freedom (Timm, 2002).

For the particular case of testing the hypothesis H_0: $\Sigma = I$ *versus* H_1: $\Sigma \neq I$, where I is an identity matrix, two possibilities have been found in the literature of Ferreira (2008), both follow an asymptotic chi-square

distribution. The test presented below is a particular case of the test proposed by Korin (1968), according Equation 1:

$$\chi_c^2 = \left[(n-1) - \frac{1}{6}\left(2p + 1 - \frac{2}{p+1}\right)\right] \times tr(S) - nlog|S| - p \dot{\sim} \chi_f^2, \;\; f = \frac{p(p+1)}{2} \qquad (1)$$

where:

tr( ) denotes the trace of a matrix, (×) the multiplication mathematical operator, S refers to the estimator of the biased sample covariance matrix, being this statistic asymptotically distributed ($\dot{\sim}$) with f degrees of freedom.

Ledoit and Wolf (2002) also presented an alternative test for this case, also suitable for the situation where the number of variables p exceeds the sample size n. This test is not limited by the uniqueness of the sample covariance matrix and the test statistic is given by Equation 2:

$$\chi_c^2 = \frac{n}{2} tr[(S-I)^2] - \frac{p^2}{2}\left[\frac{1}{p} tr(S)\right]^2 + \frac{p^2}{2} \dot{\sim} \chi_f^2, \; f = \frac{p(p+1)}{2} \qquad (2)$$

According to Ferreira (2008), for the independence test it is assumed that population covariance is null. Among its many uses, according to Li and Yao (2016), we can mention the particular interest of the biological area to test the genetic independence in genomic studies that inspired a range of discussions about the importance of tests of covariance matrix structures. In this case, there are still two situations to consider. Variances may be different, variances can be the same or homogeneous, H_0: $\Sigma = \Sigma_0 = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \vdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$.

Under the null hypothesis, the likelihood function can be written by

$$L_0(\mathrm{X};\mu,\Sigma_0) = (2\pi)^{-np/2}|\Sigma_0|^{\frac{-n}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(X_j-\mu)\Sigma^{-1}(X_j-\mu)\right\}$$
$$= (2\pi\sigma^2)^{-np/2}\left(\prod_{k=1}^{p}\sigma_{kk}^{\frac{-n}{2}}\right)\exp\left\{-\frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{n}\frac{(X_{jk}-\mu_k)^2}{2\sigma_{kk}}\right\}$$

which denotes the covariance matrix specified in the null hypothesis, so that the likelihood ratio test statistics is given by Equation 3:

$$\Lambda = \frac{|S_n|^{\frac{n}{2}}}{[\prod_{k=1}^{p}\sigma_{kk}]^{\frac{n}{2}}} = |\mathbf{R}|^{\frac{n}{2}}, \qquad (3)$$

where:

$|S_n|^{\frac{n}{2}} = |\mathbf{R}|^{\frac{n}{2}}\prod_{k=1}^{p}\hat{\sigma}_{kk}$, so that the statistic has chi-square asymptotic distribution with $f = \frac{p(p+1)}{2}$ degrees of freedom i.e., according Equation 4:

$$\chi_c^2 = -2\log(\Lambda) \dot{\sim} \chi_f^2 \qquad (4)$$

Bartlett (1954) proposed a correction to improve the chi-square asymptotic approximation. Under the null hypothesis, it can be written as Equation 5:

$$\chi_c^2 = [(n-1) - \tfrac{2p+5}{6})]\times \mathrm{tr(S)} - n\log|\mathbf{R}| - p \dot{\sim} \chi_f^2 \qquad (5)$$

It refers to the sphericity tests on the structure of a covariance matrix whose variances are homogeneous. According to Ferreira (2008), the objective is to know if the study variables correlate considerably with the same variability. Some multivariate methodologies may require that the variables under study be correlated.

According to Malhotra (2012), Bartlett's sphericity test is used when one wants to verify the hypothesis that the variables are not correlated in the population, that is, the correlation matrix is a spherical structure and each variable correlates perfectly with itself, but does not correlate with the other variables under study. It is considered by hypothesis: null covariance, equal or homogeneous variances. Further cases of sphericity can be tested using diagonal structure covariance matrices where variance values are constant and known.

$$H_0: \Sigma = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \vdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I \quad \text{vs} \quad H_1: \Sigma \neq \sigma^2 I.$$

Under the null hypothesis, the likelihood function can be written by:

L_0 (X; μ,Σ_0) = $(2\pi)^{-np/2}|\sigma^2 I|^{-\frac{n}{2}}\exp\{-\frac{1}{2}\sum_{j=1}^{n}(X_j-\mu)\Sigma^{-1}(X_j-\mu)\}$

$= (2\pi\sigma^2)^{-np/2}\exp\{-\frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{n}\frac{(X_{jk}-\mu_k)^2}{2\sigma_{kk}}\}$.

The likelihood ratio test for the sphericity hypothesis is therefore

$\Lambda = \frac{|S_n|^{\frac{n}{2}}}{\left[\frac{tr(S_n)}{p}\right]^{\frac{n}{2}}} = \frac{|S|^{\frac{n}{2}}}{\left[\frac{tr(S)}{p}\right]^{\frac{n}{2}}}$, S_n is the unbiased covariance matrix estimator, S is the estimator of the biased covariance matrix, with asymptotic chi-square distribution, that is, Equation 6:

$$\chi_c^2 = -2\log(\Lambda) = n\left\{p\,log[\frac{tr(S)}{p}] - \log|S|\right\} \sim \chi_f^2, \quad f=\frac{p(p+1)}{2}-1 \qquad (6)$$

Box (1949) presented a proposal for better performance, in which the corrected statistic is given by Equation 7:

$$\chi_c^2 = -2\log(\Lambda) = -[(n-1)-\frac{2p^2+2p+2}{6p}]\left\{\log|S| - p\log\left[\frac{tr(S)}{p}\right]\right\} \sim \chi_f^2, f=\frac{p(p+1)}{2}. \quad (7)$$

where:

$p > n$, John (1972) and Ledoit and Wolf (2002) presented an alternative test developed with the aim of making a new correction for the existing one. The test statistic that makes it robust against high dimensionality is given by Equation 8:

$$\chi_c^2 = -2\log(\Lambda) = -\frac{n}{2}tr[\left(\frac{S}{\frac{tr(S)}{p}} - I\right)^2] \sim \chi_f^2, f=\frac{p(p+1)}{2}-1. \qquad (8)$$

The same test was performed in the work of Timm (2002), however a change was made in the degrees of freedom, because instead of n, n − 1 was used. This modification was proposed by Sugiura (1972). Sphericity tests are a particular case of composite symmetry tests.

Material and methods

The data used in this work are of a fictitious nature, obtained through Monte Carlo simulations. Observations of normal multivariate distributions were made $N_p(\mu, \Sigma)$ with the mean vector $\mu \equiv 0$, without loss of generality. Different sample sizes (n) and number of variables (p) were considered. In each scenario, X_{kl} independent observations were generated on index k (sample unit) k = 1, ..., n and with different correlation in each scenario (index l = 1, ..., p variables). The scenarios under study were created through several combinations of n and p, satisfying the restriction $p < n$.

The behaviour of the effective type-I error rate, for a nominal value α = 5% in scenarios with increasing n (5, 10, 20, 30, 50 and 100) and increasing p (2, 4, 8, 16, 32 and 64) was evaluated. After obtaining the n observations of the p variables, the statistics of the identity and sphericity tests presented were obtained using the software R (R Core Team, 2018). For the simulation of the data, it was used the *mvtnorm* function of the computational language R, proposed by Genz and Bretz (2009), which generates random data through a number of data (n) and pre-established variables (p). One thousand simulated samples were generated, depending on the matrix Σ that originated the data under the null hypothesis and under the alternative hypothesis. After the simulation, the analysis and interpretation of the test was performed in relation to the type-I error rate (with nominal probability α) and power (complementary to type-II error) in the proposed scenarios.

Study 1: power of the tests for identity

H_0: Σ = I *versus* H_1: Σ ≠ I. To obtain the type I error rate, we first started the scenario in H_0 which consists of an identity matrix (p×p). For the construction of the power curve, an arbitrary measure (δ) was used that characterizes the ratio of determinants between the observed matrix and H_0. Where $\delta = |\Sigma_i|/|\Sigma_1|$ with i = 2, 3, ..., k-1 refers to the difference between the last matrix generated on the matrix specified in H_0, which characterizes a δ (delta) variability ratio. The maximum value set for the δ was 64, so that all tests reach the maximum power. The tests involved in this study were: Korin (1968) - [Equation 2 - Korin I code] and Ledoit and Wolf (2002) - [Equation 3 - Ledoit I code].

Study 2

Power of the sphericity tests: H_0: $\Sigma = 64I$ *versus* H_1: $\Sigma \neq \sigma^2 I$. Starting from H_0 (which consists of the spherical covariance structure obtained with the highest value of δ in Study 1), the deviation of H_0 was caused by the gradual increase of the value of ρ from 0 to 0.9 in steps of 0.1. Negative ρ values were not considered. In this study, the following sphericity tests were evaluated: Original likelihood ratio test [Equation (4) - code: LRT II]; Bartlett (1954) [Equation 5 - code: bartlett]; Test of original likelihood ratio sphericity [Equation 6 - code: LRT I]; Box (1949) [Equation 7 - code: Box]; Ledoit and Wolf (2002) and John (1972) [Equation 8 - jlw] and Sugiura (1972) [Equation 8 - code: sug].

In the process of classification of the test between conservative, exact and liberal, the exact confidence interval was used for the proportion based on the binomial distribution obtained in the Sisvar software (Ferreira, 2014). The range used for the level of significance was 99% CI (α): [0.0339; 0.0705] because it has the true accuracy ratio of a test at 1% significance, values below the lower limit denote a conservative test and values above the upper limit, a liberal test, in a simulation process with 1,000 repetitions. The criterion used to indicate the most powerful test was: given that a test was considered exact, the test of greater power was recommended.

Results and discussion

The order of the studies was followed for the presentation of the results. In Study 1, two specific tests for identity, Ledoit I and Korin I, were evaluated. For the $p = 2$ scenario, both tests preserved the type I error rate equal to the level of significance adopted. Thus, they were considered exact for every n evaluated. The scenarios in which the sample size is small $n = 5$ and $n = 10$, the Ledoit I test was more powerful for every δ value, and should be recommended in these situations. According to Nogueira and Pereira (2013), it is important to have tests with levels of significance close to the α adopted a priori and that the power is high, even in situations of small samples. It can be summarized, in this scenario, that the power of the tests increases with the decrease of δ as the sample of data increases. In the case $p = 2$, regardless of sample size, the Ledoit I test is recommended.

For the $p = 4$ case, all sample sizes indicated that the Ledoit I and Korin I tests showed exact behaviour. In these situations, the Ledoit I test was more powerful for all δ values in relation to Korin I. The power curves of the tests only approach for $n = 100$. After analysis, regardless of all possible combinations, the use of the Ledoit I test is indicated because it presents the type I error rate close to the level of significance adopted and higher levels of power.

For $p = 8$, Ledoit I is liberal and Korin I exact in the scenarios $n = 20$ and $n = 30$. Increasing the sample size to $n \geq 50$ caused the decrease in the type I error rate of both tests, Ledoit I that was liberal became exact and considered more powerful. For $n = 100$, the values $\delta = 2$ to 16, the Ledoit I test was more powerful, for other variations of δ, both tests reach the maximum probability of power. The results suggest the Ledoit I test as the most appropriate.

In the $p = 16$ case considering $n = 30$, a liberal behaviour of both tests was obtained. The Ledoit I test was more powerful for every δ, but did not approximate the maximum probability of rejection of H_0 in the scenario evaluated. Remembering that, in this case, it is a false power. For $n = 50$, Ledoit I is liberal and Korin I exact, but both do not reach the maximum probability of power. In $n = 100$ for all points of the power curve, the Ledoit I test was considered the most powerful. The power curve showed a similar behaviour for all n used. The increase in n caused the decrease in the value of the type I error rate and improvement of power in both tests.

When considering $p = 32$ and $n = 50$, the Ledoit I and Korin I tests were evaluated as liberal. The Korin I test was more powerful than Ledoit I. The higher the value of the type I error rate, the more power it has, but this is not a reliable result. For the $n = 100$ scenario, the Ledoit I test was liberal and Korin I was rated as exact and most powerful. Setting $p = 32$, increasing the sample size from 50 to 100 provided a reduction in the type I error rate of both tests where they presented similar behaviour.

When assessing $p = 64$ and $n = 100$, it was observed that both are liberal. The Korin I test was the most powerful for all δ values assessed, but also the most liberal. Both presented low power, which is contradictory with the expected, moreover, liberal tests are naturally more powerful. In this scenario, the discussion arises that committing type I error may be more serious than type II error. It was observed in Study 1 (Figure 1) that as the value of n and p increases, both tests become less powerful. In the classification of the type I error rate ($\delta = 1$), the Ledoit I test is the most indicated for Study 1 as presented in Table 1. A fact noted is that the Ledoit I test created for situations where $p > n$, in this study where $n > p$,

situation that avoids non-zero eigenvalues in the numerator of the statistic Λ according to Wang and Yao (2013), obtained a satisfactory result.

In Study 2, the sphericity tests were evaluated. In the $p = 2$ case, considering $n = 5$ the LRT II, jlw and sug tests are conservative, Box and Bartlett exact and LRT I is considered liberal. In this scenario, given the results obtained, the most indicated test is Bartlett for better control of the type I error rate and greater power between the exact ones, followed by Box.

Wang and Yao (2013) proposed corrections for the likelihood ratio test and John's (1972) test for large sphericity.

The performance of these tests was evaluated in situations of normality and non-normality. It was concluded that when the sample size n is fixed, the power of the test decreases when the value of p is close to n; this could be explained by the fact that some of the eigenvalues of the maximum likelihood estimator S_n approach zero, making the test almost degenerate and lose power. In the $n = 5$ case, where the sample size is small and approaches the number of p variables, the jlw test also did not perform satisfactorily. In the $n = 10$ scenario, none of the assessed tests were liberal, the jlw, sug, Bartlett and LRT II tests were conservative, Box and LRT I exact. From the results, it was concluded that the most indicated test is the test proposed by Box (1949), to lead to a better approximation for the chi-square distribution in small samples.

In the $n = 20$ and 30 scenarios, the tests were classified in relation to the type I error rate: LRT I, Bartlett and LRT II as conservatives, and Box, jlw and sug as exact tests. For $\rho = 0.9$, all tests reached the maximum probability of power. As in the previous scenario, the most effective test was the Box test. For the $n = 50$ case, LRT I, Bartlett and LRT II were considered conservative; Box, jlw and sug were considered exact. None of the tests was considered liberal, for $\rho = 0.7$ only Bartlett and LRT II do not reach the maximum probability of power, only occurring for all in $\rho = 0.9$. Keeping the same number of variables $p = 2$ and increasing the sample number to $n = 100$, the classification in relation to the type I error rate was that the tests LRT I, Bartlett and LRT II are conservative and Box, jlw and sug are exact. For $\rho = (0.6, ..., 0.9)$, all tests reached the maximum probability of power. For $p = 2$, it was observed that the increase in the number of n provided the tests with a power towards the maximum probability with this simulation proposal.

According to Hair, Black, Babin, and Anderson (2010), when the sample size is small the hypothesis test is little sensitive, it also states that increasing the sample size will imply increasing the power of the test, which was noted in this scenario. The most suitable test for $n > 5$ is, unanimously, the Box test.

Increasing the number of variables for $p = 4$ and in the $n = 10$ scenario, sug and LRT II tests were conservative; Box, jlw and Bartlett exact and LRT I liberal. From the sphericity tests evaluated, considering the cases for $n \geq 20$, LRT I, Box, jlw and sug were classified as exact, while Bartlett and LRT II as conservative. For $n = 20$, in $\rho = 0.8$, all tests reach the maximum probability, with the exception of LRT II. Increasing the value of n to 30 and keeping the number of variables $p = 4$ for $\rho \geq 0.7$, all reached maximum power. In the scenarios $n = 50$ of $\rho = 0.1$ to 0.4 the sequence of greatest power is jlw, sug, Box, LRT I, Bartlett and LRT II. It was found that, by increasing the sample size, the jlw test achieved higher levels of power when compared to the others in most scenarios.

In the $p = 4$ scenario, the increase in the number of variables provided a decrease in the type I error rate and an improvement in the power of the tests. The LRT I test which was a liberal test became exact, LRT II remained as conservative, sug changed from conservative to exact, Bartlett changed the classification from exact to conservative, Box and jlw remained exact. In these scenarios the jlw test, a comprehensive test created even for non-singular matrices and $p > n$ (Ferreira, 2008), is a satisfactory and recommended test.

Table 1. Test recommendations in Study 1.

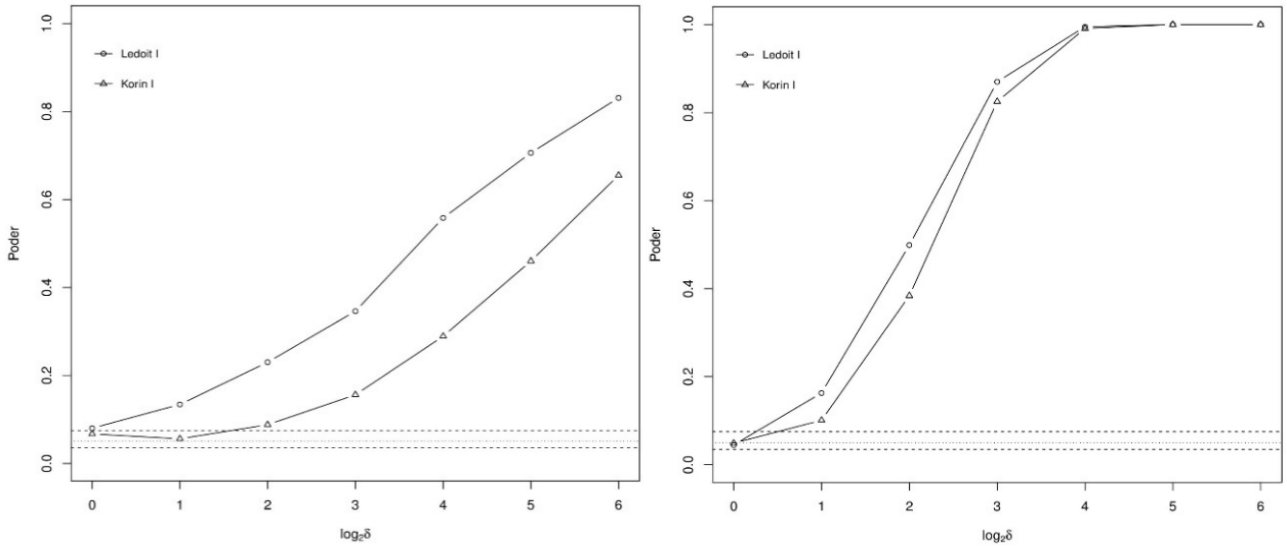
n	p					
	2	4	8	16	32	64
5	Ledoit I					
10	Ledoit I	Ledoit I				
20	Ledoit I	Ledoit I	Korin I			
30	Ledoit I	Ledoit I	Korin I	Ledoit I		
50	Ledoit I	Ledoit I	Ledoit I	Korin I	Ledoit I	
100	Ledoit I	Ledoit I	Ledoit I	Ledoit I	Korin I	Ledoit I

Considering the sample size n = 20 and the increase in the number of variables for p = 8, LRT II showed conservative behaviour, Box and sug exact behaviour, LRT I, jlw and Bartlett liberal behaviour. The LRT I test showed liberal behaviour and of course, higher level of power is expected. In these situations, jlw was considered more appropriate for maintaining the control of the type I error rate. With the increase of the sample number and variables, the tests were more powerful and some of them more conservative, and better controlling the type I error rate.

In the p = 16 case and n = 30, LRT II is the only conservative test, sug is the only exact, LRT I, Box, jlw and Bartlett liberal. At n = 50, Box and jlw become exact and sug liberal, the decreasing sequence of power is LRT I, jlw, sug, Bartlett, Box, and LRT II. For the p = 32 case, jlw remained the most appropriate test for controlling the type I error rate and having the best power curve.

In the work of Wang and Yao (2013), when studying the behaviour of sphericity tests in large dimensions for the normality situation when n is equal to 64 or 128, with variable values of p and below the two alternatives mentioned above, the power of the tests increases with the sample size. For the large sample size n = 256 and p ranging from 16 to 240, all the evaluated tests of interest showed power around 1. Similar situation occurs in this study, when n = 100, the maximum sample size evaluated, the tests present high power close to 1 as the correlation increases. Assuming p = 64 and n = 100 sug is an exact test, all the other tests are liberal, in heterogeneity ρ = 0.1, all tests reached the maximum power level except the LRT II test. In Study 2, regardless of the proposed combinations of n and p given in Table 2, the test that was most indicated disregarding the situations of false power is the jlw. [See Figure 2].

In this study, as in the work of Lim, Li, and Lee (2010), it was possible to observe that, in general, the power performance of the likelihood ratio tests were low. According to the authors, it is well known that the majority of the likelihood ratio tests based on the chi-square approximation limit have a high likelihood of rejection, the modifications applied to the test statistics improved performance based on chi-square.



(a) $n = 20$ and $p = 8$ (b) $n = 100$ and $p = 8$

Figure 1. Type I error rate and power for Study 1, p = 8: n = 20 (a) and n = 100 (b), dashed lines denote: thin line: level of significance of α = 5%, thick lines: confidence interval.

Table 2. Test recommendations in Study 2.

n	p					
	2	4	8	16	32	64
5	bartlett					
10	box	jlw				
20	box	jlw	box			
30	box	jlw	jlw	sug		
50	box	jlw	jlw	jlw	jlw	
100	box	jlw	jlw	jlw	jlw	sug

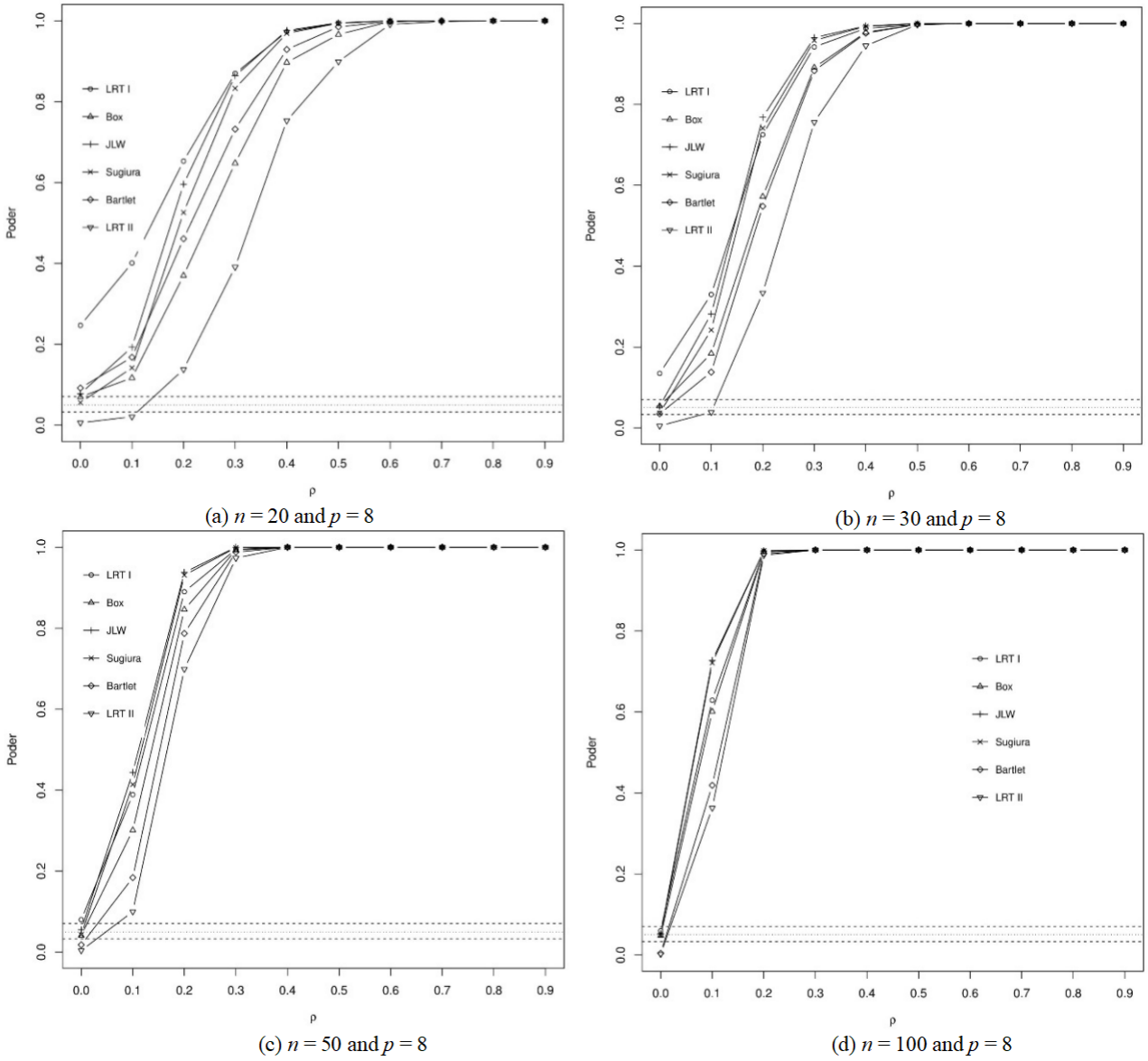


Figure 2. Type I error rate and power for Study 1, $p = 8$: $n = 20$ (a), $n = 30$ (b), $n = 50$ (c) and n = 100 (d), dashed lines denote: thin line: level of significance of $\alpha = 5\%$, thick lines: confidence interval.

Conclusion

It was found that the modifications collaborated to increase the power of the tests in Studies 1 and 2. In order to evaluate identity, Ledoit and Wolf's (2002) proposal was the most appropriate one; for sphericity, the version of John (1972) modified by Ledoit and Wolf (2002) followed by Box´s (1949) proposal were the ones with the best performance. Bartlett's proposal should be used only for small samples and small number of variables.

Acknowledgement

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brasil (Capes) - Finance Code 001.

References

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society - Series B (Methodological)*, *16*(2), 296–298.

Alves, R. M., Tavares, M. R. M., Tavares, H. R., Lobato, T. C., & Oliveira, T. F. D. (2015). Modelo de efeitos fixos com medida repetida aplicado em experimentos de melhoramento genético do cupuaçuzeiro. *Revista Brasileira de Fruticultura,* 37(4), 993-1000. doi: 10.1590/0100-2945-234/14

Biase, N. G., & Ferreira, D. F. (2011). Testes de igualdade e de comparações múltiplas para várias proporções binomiais independentes. *Revista Brasileira de Biometria, 29*(4), 549-570.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 36*(3-4), 317-346. doi: 10.2307/2332671

Cantelmo, N. F., & Ferreira, D. F. (2007). Desempenho de testes de normalidade multivariados avaliado por simulação Monte Carlo. *Ciência e Agrotecnologia, 31*(6), 1630-1636. doi: 10.1590/S1413-70542007000600005

Casella, G., & Berger, R. L. (2010). *Inferência estatística* (2a ed.). São Paulo, SP: Cengage Learning.

Cecon, P. R., Silva, F. F., Ferreira, A., Ferrão, R. G., Carneiro, A. P. S., Detmann, E., ... Morais, T. S. S. (2008). Análise de medidas repetidas na avaliação de clones de café 'Conilon'. *Pesquisa Agropecuária Brasileira, 43*(9) 1171-1176. doi: 10.1590/S0100-204X2008000900011

Doria Filho, U. (1999). *Introdução à bioestatística para simples mortais* (3a ed.). São Paulo, SP: Negócio Editora.

Fávero, L. P. L., Belfiore, P. P., Silva, F. P., & Chan, B. L. (2009). *Análise de dados: modelagem multivariada para tomada de decisões*. Rio de Janeiro, RJ: Elsevier.

Fávero, L. P. L., Belfiore, P. P., Takamatsu, R. T., & Suzart, J. (2014). *Métodos quantitativos com stata: procedimentos, rotinas e análise de resultados*. Rio de Janeiro, RJ: Elsevier.

Ferreira, D. F. (2008). *Estatística multivariada*. Lavras, MG: UFLA.

Ferreira, D. F. (2014). Sisvar: a Guide for its Bootstrap procedures in multiple comparisons. *Ciência e Agrotecnologia, 38*(2), 109-112. doi: 10.1590/S1413-70542014000200001

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities* (Vol. 195, Lecture Note in Statistics). Heidelberg, DE: Springer Science & Business Media. doi: 10.1007/978-3-642-01689-9

Gouvêa, M. A., Prearo, L. C., & Romeiro, M. C. (2012). Avaliação da aplicação de técnicas multivariadas de interdependência em teses e dissertações de algumas instituições de ensino superior. *FACEF Pesquisa-Desenvolvimento e Gestão, 15*(1), 107-124.

Gouveia, J. F., Silva, J. A. A., Ferreira, R. L. C., Gadelha, F. H. L., & Lima Filho, L. M. A. (2015). Modelos volumétricos mistos em clones de Eucalyptus no polo gesseiro do Araripe, Pernambuco. *Floresta, 45*(3), 587-598. doi: 10.5380/rf.v45i3.36844

Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Harlow, UK: Pearson Prentice Hall.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*(1), 69-82. doi: 10.2307/1164736

John, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika, 59*(1), 169-173. doi: 10.1093/biomet/59.1.169

Kedem, B., & Fokianos, K. (2005). *Regression models for time series* analysis. Wiley, NY. Wiley Series in Probability and Statistics.

Korin, B. P. (1968). On the distribution of a statistic used for testing a covariance matrix. *Biometrika, 55*(1), 171-178. doi: 10.1093/biomet/59.1.169

Ledoit, O., & Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics, 30*(4), 1081-1102. doi: 10.1214/aos/1031689018

Li, Z., & Yao, J. (2016). Testing the sphericity of a covariance matrix when the dimension is much larger than the sample size. *Electronic Journal of Statistics, 10*(2), 2973-3010. doi: 10.1214/16-EJS1199

Lim, J., Li, E., & Lee, S.-J. (2010). Likelihood ratio tests of correlated multivariate samples. *Journal of Multivariate Analysis, 101*(3), 541-554. doi: 10.1016/j.jmva.2009.10.011

Littell, R. C., Henry, P. R., & Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science, 76*(4), 1216-1231. doi: 10.2527/1998.7641216x

Malhotra, N. K. (2012). *Pesquisa de marketing: uma orientação aplicada*. Porto Alegre, RS: Bookman Editora.

Nogueira, D. A., & Pereira, G. M. (2013). Desempenho de testes para homogeneidade de variâncias em delineamentos inteiramente casualizados. *Sigmae, 2*(1), 7-22.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, AT: R Foundation for Statistical Computing.

Reis, E. (1997). *Estatística Multivariada Aplicada*, Lisboa, PT. Ed. Silabo.

Santo, R. E. (2012). Utilização da análise de componentes principais na compressão de imagens digitais. *Einstein, 10*(2), 135-139. doi: 10.1590/S1679-45082012000200004

Sugiura, N. (1972). Locally best invariant test for sphericity and the limiting distributions. *The Annals of Mathematical Statistics, 43*(4), 1312-1316. doi: 10.1214/aoms/1177692481

Timm, N. H. (2002). *Applied multivariate analysis*. New York, NY: Springer-Verlag.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16*(2), 117-186. doi: 10.1214/aoms/1177731118

Wang, Q., & Yao, J. (2013). On the sphericity test with large-dimensional observations. *Electronic Journal of Statistics, 7*, 2164-2192. doi: 10.1214/13-EJS842

Xavier, L. H., & Dias, C. T. S. (2001). Acurácia do modelo univariado para análise de medidas repetidas por simulação multidimensional. *Scientia Agricola, 58*(2), 241-250. doi: 10.1590/S0103-90162001000200005