Testing the independence of variables for specific covariance structures: a simulation study

Filipe J. Marques[a,b,*], Joana Diogo[a], Mina Norouzirad[b], Regina Bispo[a,b]

[a] NOVA University of Lisbon, FCT NOVA, Portugal
[b] Center for Mathematics and Applications, NOVA Math, Portugal

Abstract
In this work we show how it is possible to test the nullity of covariances, in a set of variables, using a simple univariate procedure. The methodology proposed enable us to preform the multivariate test of independence of several variables, under specific conditions for the covariance structure. The methodology proposed may be used in the high dimensional setting and, given its simplicity, allows to overcome the difficulties in using the exact distribution of the statistic used in the likelihood ratio testing procedure. A simulation study is provided to assess the power and significance level, in different scenarios, of the testing procedure proposed when compared with different likelihood ratio tests and to the testing methodology in Schott (2005).

Keywords Chi-square tests, high dimensional setting, likelihood ratio tests.

1 Introduction

The independence of a set of variables is a key assumption in different statistical methods. Under the assumption of multivariate Normality, to test the independence of a set of variables is equivalent to test if the covariance matrix is diagonal. The distribution of the likelihood ratio statistic, Λ, used to test the diagonal structure of a covariance matrix may be represented as the distribution of the product of independent Beta random variables (Coelho and Marques, 2010; Marques et al., 2011). However, as it is well known (Marques et al., 2011; Marques and Coelho, 2020) this distribution is not easy to handle in practice. There are several approximations available for the distribution of this likelihood ratio statistic, including the χ^2 (Wilks, 1938), Box type (Box, 1949) and saddle point (Butler et al., 1993) approximations but these, in some scenarios, may not be precise enough, for example when the samples are small or when there is a large number of variables. More recently, in Coelho (2004), the author developed new approximations, the so-called near-exact distributions, which may be used to develop more precise approximations for the likelihood ratio statistic used to test the independence structure (Coelho, 2004; Coelho and Marques, 2010). This test was also addressed in Coelho and Marques (2010) as one of the two tests composed to obtain the statistic used to test the null hypothesis of a spherical structure. The same technique was also used in Anderson (2003).

*Corresponding author: Filipe J. Marques, Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal (*fjm@fct.unl.pt*)

As it is well-known, the likelihood ratio testing approach can only be considered when the sample size, n, is greater than the number of variables, p. However, in these cases, even if p and n are large, the traditional approximations may be inaccurate or may become excessively time consuming. These scenarios may arise in big data problems. Likelihood ratio tests can not be used when $p > n$ because the sample covariance matrix is not definite positive. Thus, our main objective with this work is to provide a simple testing procedure which may be used to test the independence of a set of variables under specific assumptions, when $n \geq p$ or $n < p$, .

This problem of testing the independence of a set of variables, also designated by complete independence, has also been address in Schott (2005). The author considers a statistic based on the sum of squares of sample correlations and shows that its distribution converges to a Normal distribution when both the sample size and the number of variables go to infinity. A similar approach was followed by Srivastava (2005). The results in Schott (2005) will be considered in the simulation section of this work. Ledoit and Wolf (2002) investigated the consistency property and limiting distribution of several test statistics as dimensionality and sample size go to infinity together, with their ratio converging to a finite nonzero limit. In Li and Liu (2016) a permutation test is considered, which is based on the maximum between the largest off-diagonal entry and the largest eigenvalue of the sample correlation matrix, which have as limiting distributions, respectively, a type I extreme value distribution and Tracy–Widom law of type I distribution. In this work we propose a different testing procedure that is based on a fundamental univariate test and therefore quite straightforward. The test statistic has an exact χ^2 distribution and can be easily used and implemented.

Thus, the main goals of this work are: (i) to develop a univariate procedure for testing the independence of a set of variables under the assumption that their covariances are all non-positive or non-negative, (ii) to apply this procedure in two scenarios, one non-realistic with known variances and one realistic with unknown variances, (iii) to show that this methodology may be applied when $n < p$, and (iv) finally, to show that the new test is more robust than the existing ones for departures from normality.

This paper is organized as follows: in Section 2, we present the usual likelihood ratio procedure used to test the independence of a set of variables, and we also present the novel approach suggested in this study. In Section 3, we use simulations to compare the estimated power and significance level of the new testing procedure with those of the usual likelihood ratio test and Schott's testing methodology (Schott, 2005). We examine two likelihood ratio tests of complete independence for specific covariance structures in Section 4 and compare their estimated power to the new approach. Finally, Section 5, is dedicated to the conclusions.

2 Methodology

In this section we present the likelihood ratio test and the new testing procedure, proposed in this work, to test the independence of several variables.

2.1 The likelihood ratio test

Let us suppose we have a sample of size n from a multivariate Normal population $\underline{X} = (X_1, \ldots, X_p)' \sim N_p(\underline{\mu}, \Sigma)$ with mean vector $\underline{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \ldots & \sigma_{1p}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \ldots & \sigma_{2p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}^2 & \sigma_{p2}^2 & \ldots & \sigma_{pp}^2 \end{pmatrix}. \tag{1}$$

We are interested in testing the null hypothesis

$$H_0 : \sigma_{ij}^2 = 0 \quad \text{for all } i, j \text{ with } i \neq j \tag{2}$$

which, under Normality, states that the random variables are independent. It is well-known that, under the Normal assumption, the likelihood ratio test statistic (Anderson, 2003) is

$$\Lambda = \left(\frac{|A|}{\prod_{i=1}^{p} a_{ii}} \right)^{n/2} \tag{3}$$

where A is the maximum likelihood estimator of Σ and a_{ii} is i-th element of the diagonal of A. As already mentioned the exact distribution of this statistic is difficult to handle in practice, so the implementation of the test requires the use of approximations, such as the χ^2 approximations (Anderson, 2003; Wilks, 1938; Rencher and Christensen, 2012; Mudholkar et al., 1982). We suggest the use of near-exact distributions (Coelho, 2004; Coelho and Marques, 2010) to approximate the distribution of the likelihood ratio statistic. In Coelho and Marques (2010), in Section 2, the authors developed near-exact approximations for the test statistic in (3) and in Subsection 4.1, of the same reference, the precision of the approximations developed is assessed when compared to the approximation proposed in Mudholkar et al. (1982).

The likelihood ratio statistic in (3) is used to test the independence without assuming a specific covariance structure under the alternative hypothesis. In Section 4, we present two likelihood ratio tests that can also be used to test the independence of variables against specific covariance structures in the alternative hypothesis.

2.2 Univariate reduction

Let $W = \sum_{i=1}^{p} X_i$. If the variables X_i $(i = 1, \ldots, p)$ are independent, then $V(W) = \sum_{i=1}^{p} V(X_i)$, where $V(\cdot)$ denotes de variance of a random variable. This result, in general, is not an equivalence. However, if we consider the assumption that the covariances are all non-positive or non-negative then, we have an equivalence, that is, if $V(W) = \sum_{i=1}^{p} V(X_i)$ then we have the nullity of covariances which, under the normality assumption, ensures the independence between random variables. Thus, this relation can thus be used to test the independence of the random variables X_i $(i = 1, \ldots, p)$. Consider an observed sample of size

n, $(\underline{x}_1, \ldots, \underline{x}_n)$, from a multivariate Normal population $\underline{X} = (X_1, \ldots, X_p)' \sim N_p(\underline{\mu}, \Sigma)$ with $\underline{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix Σ in (1), we may represent observed sample matrix as

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ x_{31} & x_{32} & \ldots & x_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

then:

(i) if the variances are unknown (the most realistic scenario), we may use the maximum likelihood estimators, S_{ii}^2, to estimate the variances, σ_{ii}^2, of each of the variables X_i $(i = 1, \ldots, p)$;

(ii) we may obtain an observed sample for $W = \sum_{i=1}^{p} X_i$, which we will denote by $(w_1, w_2, \ldots, w_n)$, where $w_j = \sum_{i=1}^{p} x_{ji}$ $(j = 1, \ldots, n)$ and $(x_{1i}, \ldots, x_{ni})$ is the observed sample of the random variable X_i $(i = 1, \ldots, p)$. We may construct a new sample matrix with a new column for the observed sample of W, more precisely

$$\mathbf{X}^* = \left(\begin{array}{cccc|c} x_{11} & x_{12} & \ldots & x_{1p} & w_1 \\ x_{21} & x_{22} & \ldots & x_{2p} & w_2 \\ x_{31} & x_{32} & \ldots & x_{3p} & w_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} & w_n \end{array}\right);$$

(iii) using the maximum likelihood estimator, S_W^2, and the observed sample $(w_1, w_2, \ldots, w_n)$ we can estimate variance of W;

(iv) finally, we can consider an univariate test for the variance define as

$$H_0 : \sigma_W^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma_W^2 \neq \sigma_0^2 \tag{4}$$

where, under H_0, when the variances are known, $\sigma_0^2 = \sum_{i=1}^{p} \sigma_{ii}^2$ and when the variances are unknown $\sigma_0^2 = \sum_{i=1}^{p} s_{ii}^2$ where s_{ii}^2 is the observed value of S_{ii}^2. When the variances are known, the test statistics is

$$\frac{(n-1)S_W^2}{\sum_{i=1}^{p} \sigma_{ii}^2} \sim \chi_{n-1}^2 \tag{5}$$

where S_W^2 is the estimator of the variance of W. When the variances are unknown we replace, in (5), $\sum_{i=1}^{p} \sigma_{ii}^2$ by $\sum_{i=1}^{p} s_{ii}^2$. The distribution of the preceding test statistic relies on the assumption of normality of W, which is satisfied if $\underline{X} = (X_1, \ldots, X_p)' \sim N_p(\underline{\mu}, \Sigma)$.

If the variables X_i $(i = 1, \ldots, p)$ are not normal distributed, we may consider asymptotic results, such as:

- in the absence of the assumption of Normality, when the variables are identically distributed with finite variances $\sigma_{ii}^2 = \sigma^2$; if the variables are independent then, the Central Limit Theorem, ensures that

$$W = \sum_{i=1}^{p} X_i \underset{p\to\infty}{\overset{a}{\sim}} N\left(\sum_{i=1}^{p} \mu_i, \sum_{i=1}^{p} \sigma_{ii}^2 = p\sigma^2\right) ; \tag{6}$$

- there are several generalizations of the previous result that may help us to deal with other scenarios. For example, without the assumption of Normality, when the variables are not identically distributed with variances equal to σ_{ii}^2, with $|X_i|$ having finite moments of some order $2 + \delta$ for some $\delta > 0$, and the rate of growth of these moments limited by the Lyapunov condition, if the variables are independent, then by the Liapounov's, theorem we have

$$W = \sum_{i=1}^{p} X_i \underset{p\to\infty}{\overset{a}{\sim}} N\left(\sum_{i=1}^{p} \mu_i, \sum_{i=1}^{p} \sigma_{ii}^2\right) . \tag{7}$$

When the normality assumption is violated, the independence of the variables is no longer guaranteed when the covariances are null. Nevertheless, we believe that the test is robust to the lack of normality and we will examine this point in the simulation section.

Several additional comments are in order:

- if the variances are unknown, large sample sizes are advisable in order to improve the precision of the estimate of $\sum_{i=1}^{p} \sigma_{ii}^2$;
- if there is a departure from the Normality assumption, a large number of variables improve the approximating results in (6) and (7);
- this method is applicable in high-dimensional settings, i.e., when $n < p$ with large p;
- this approach can also be used in the presence of missing values for some of the variables.

Lastly, this approach is much simpler and more straightforward than the likelihood ratio testing procedure, as it reduces a multivariate testing procedure to a well-known univariate test for the variance of W.

The properties of this testing procedure will be investigated in the following section.

3 Simulations

To understand the properties of this testing procedure, we consider samples from a multivariate Normal population with a null mean vector $\underline{\mu}$ and a covariance matrix Σ assuming the following structures:

- Equivariance-equicorrelation or compound symmetry (CS)

$$\Sigma = \Sigma_{CS} = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}. \tag{8}$$

- Autoregressive (AR)

$$\Sigma = \Sigma_{AR} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{|p-1|} \\ \rho & 1 & \rho & \dots & \rho^{|p-2|} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{|p-1|} & \rho^{|p-2|} & \rho^{|p-3|} & \dots & 1 \end{bmatrix}. \tag{9}$$

- Circular (C)

$$\Sigma = \Sigma_C = \begin{bmatrix} \sigma^2 & b_1 & \dots & b_{p-1} \\ b_{p-1} & \sigma^2 & \dots & b_{p-2} \\ \vdots & \vdots & & \vdots \\ b_1 & b_2 & \dots & \sigma^2 \end{bmatrix} \tag{10}$$

  where $b_j = b_{p-j}$ for $j = 1, \dots, \lfloor p/2 \rfloor$, with $\lfloor \cdot \rfloor$ represents the largest integer that is not greater than the argument.

Note that in structure (CS), the correlations (or covariances) always have the same sign, whereas in structures (AR) and (C), we must consider values that ensure this assumption.

The (CS) and the (AR) covariance structures are two of the most frequently assumed covariance structures (see, for example, Chan and Choy, 2008; Littell et al., 2000). The circular structure (C) is more generic and includes the former two. The choice of these structures allowed us to easily define values for their parameters ensuring that the assumption required for the proposed methodology was verified.

Additionally, we must ensure that the covariance matrices are positive definite, so the choice of covariance matrix values is not arbitrary. On the $x-$axis of all figures presented in the next subsections, we have the values of ρ as identified in structures (CS) and (AR) in (8) and (9), respectively. Regarding the structure in (10), the value of ρ, in the figures, can not be directly identified in the matrix. The value of ρ shown in the $x-$axis of figures is the one used to generate the values in a matrix with the structure (C) in (10). More precisely, we use the function circulant from the package scipy.linalg of Python to generate different circular matrices. The values of ρ are chosen to ensure that in the simulations: (i) the covariance matrices are positive definite, and (ii) that we range from matrices whose covariances are all negative and then move on to matrices with all positive covariances. Although the covariances in this process are guaranteed to have values very close to zero, the value $\rho = 0$ does not ensure that they are all zero. All simulations and calculations were performed using Python 3.7 through Google Colab platform.

3.1 Plots for the estimated powers

In this subsection, we examine various choices for n and p. We compute the estimated power for (i) the likelihood ratio test described in Subsection 2.1, (ii) the new univariate methodology presented in Subsection 2.2 and (iii) the testing methodology developed by Schott (2005), which uses an approximation to the Normal distribution and may also be applied in the high dimensional setting (see Schott, 2005). The power is estimated by simulating 2000 samples for each combination of p and n.

3.1.1 Case I: $n > p$

In Figures 1, 2, and 3, the estimated test powers considering the covariance structures (8), (9), and (10) are depicted, respectively. In both Figures 1 and 3, it can be seen that the estimated power is higher in all situations for the methodology proposed in this paper. The autoregressive covariance structure (Figure 2) may allow us to see that the trends have changed and that the differences are not as noticeable. Additionally, it is important to note that, for the univariate method suggested in this study, there are typically no differences between the solid and dotted lines representing the estimated powers when the variances are known or unknown. In many plots, we can only see the solid line. Thus, we can conclude that, for the univariate procedure introduced in Subsection 2.2, there are no differences in the estimated power, regardless of whether the variances are known or unknown, particularly when the sample size and number of variables are large.

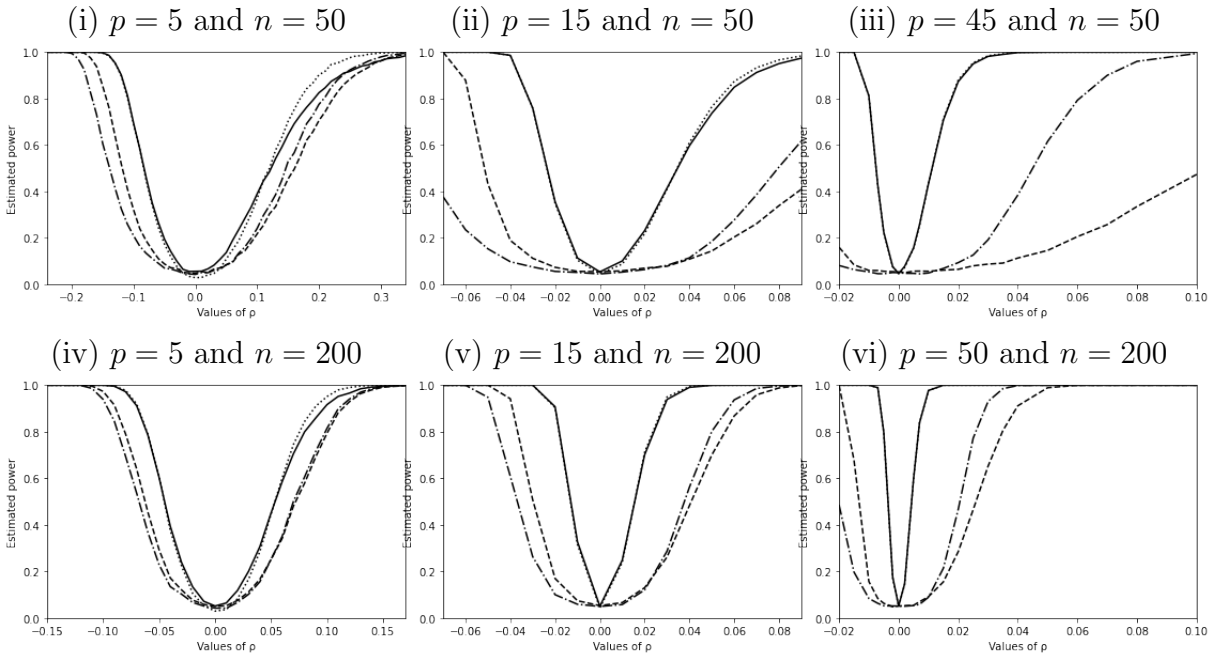

Figure 1: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively), (ii) likelihood ratio test (dashed line), and (iii) testing procedure presented in Schott (2005) (dotted-dashed line) as a function of ρ for random samples from a multivariate Normal population with $\Sigma = \Sigma_{CS}$ in (8).

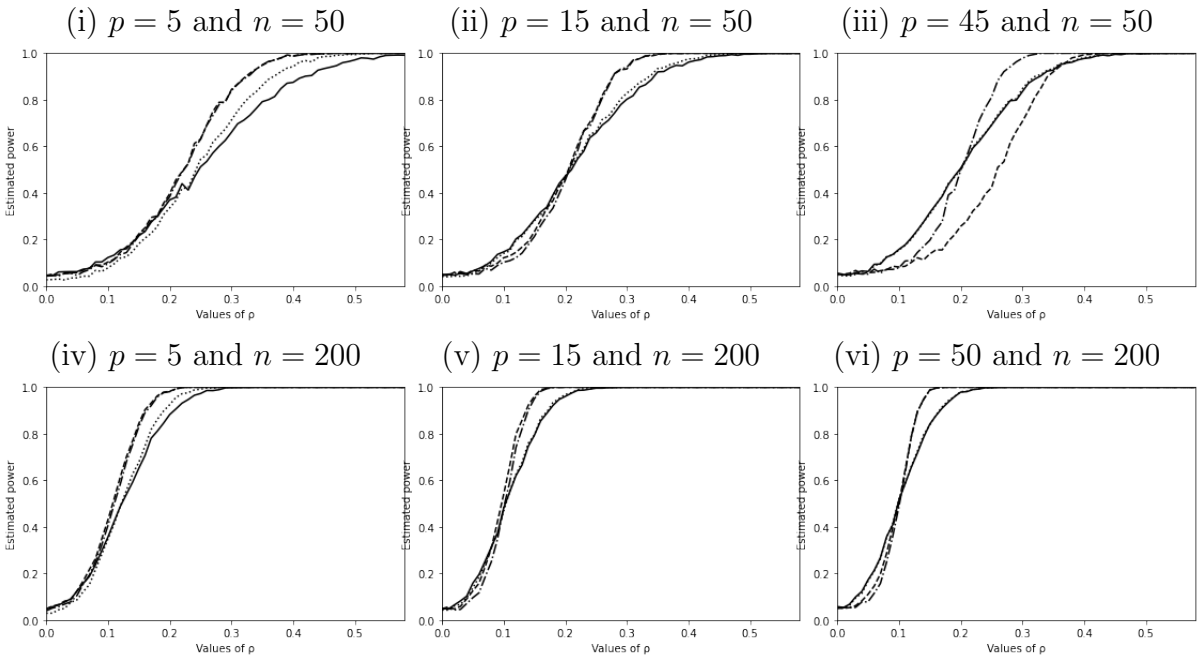

Figure 2: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively), (ii) likelihood ratio test (dashed line), and (iii) testing procedure presented in Schott (2005) (dotted-dashed line) as a function of ρ for random samples from a multivariate Normal population with $\Sigma = \Sigma_{AR}$ in (9).

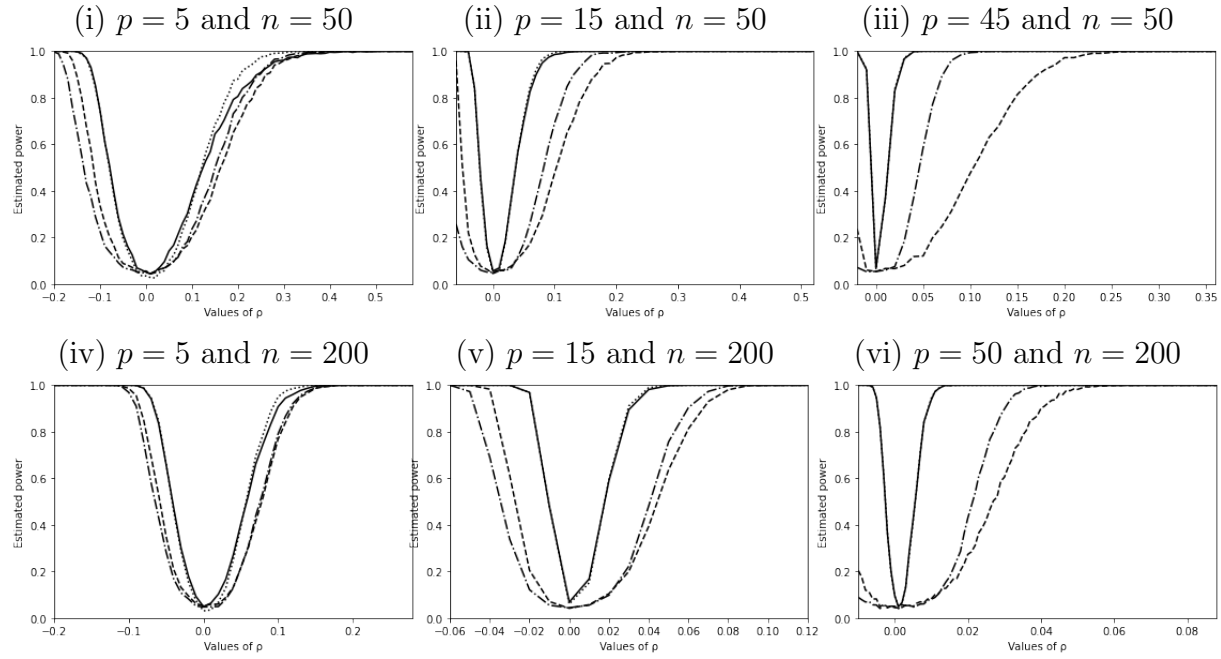

Figure 3: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively), (ii) likelihood ratio test (dashed line), and (iii) testing procedure presented in Schott (2005) (dotted-dashed line) as a function of ρ for random samples from a multivariate Normal population with $\Sigma = \Sigma_C$ in (10).

3.1.2 Case II: $n < p$

In this subsection, we address high dimensional scenarios, so we will examine a variety of n and p such that $n < p$. In Figures 4, 5, and 6, we only present the estimated powers for the tests in Subsection 2.2 and Schott (2005) since the likelihood ratio test approach can not be used when $n < p$, as stated earlier. These figures may reveal patterns similar to those previously identified for $n > p$.

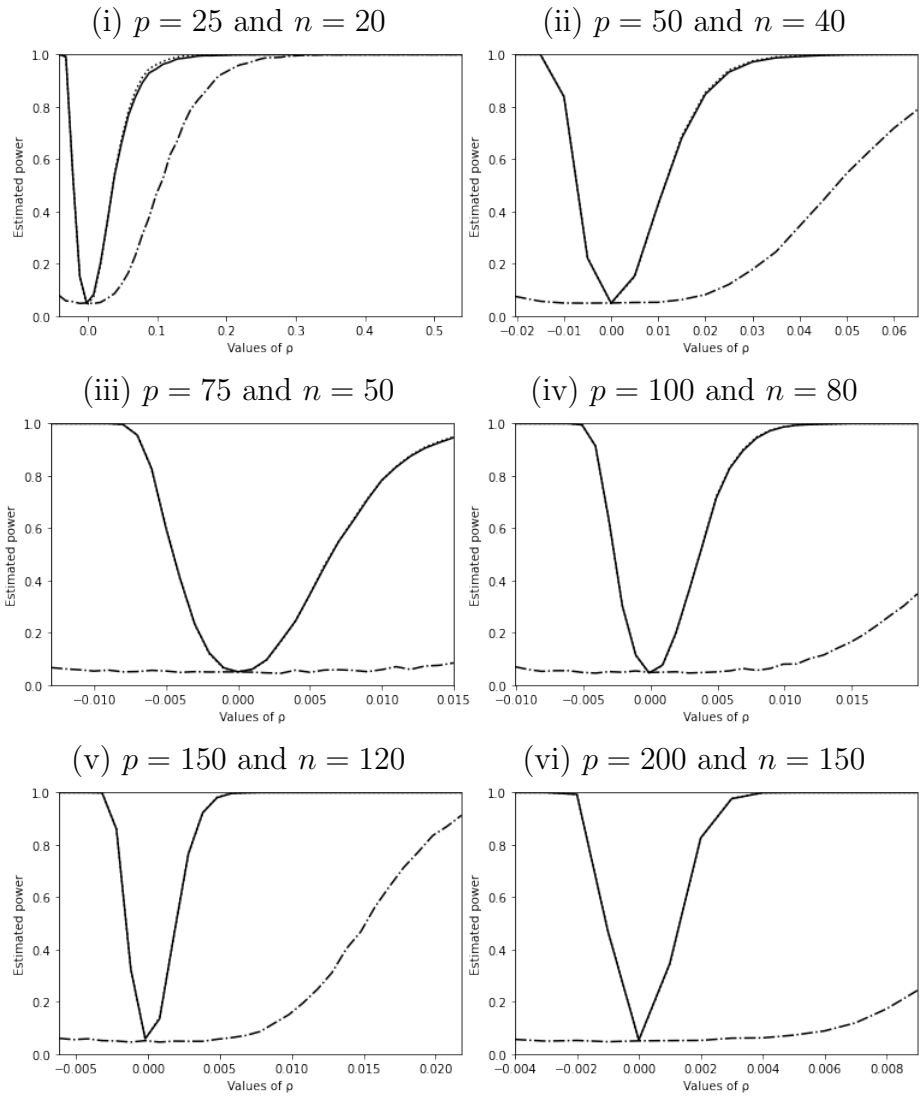

Figure 4: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively) and (ii) for the testing procedure presented in Schott (2005) (dotted-dashed line) as a function of ρ, considering random samples from a multivariate Normal population with $\Sigma = \Sigma_{CS}$ in (8).

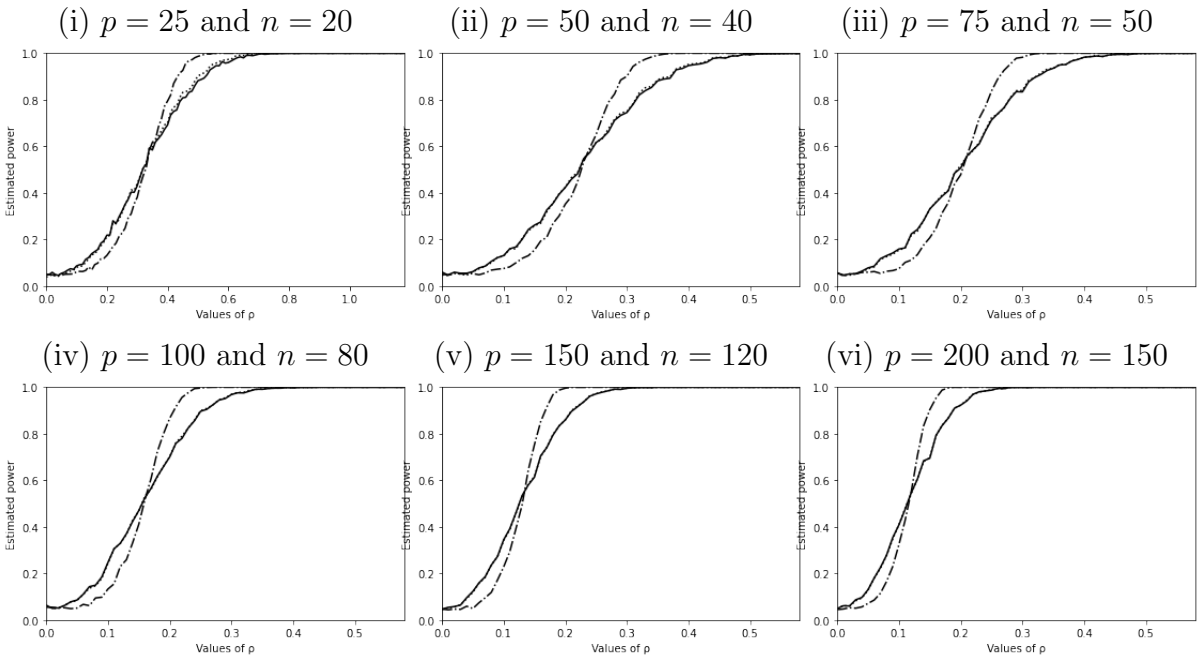

Figure 5: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively) and (ii) for the testing procedure presented in Schott (2005) (dotted-dashed line) as a function of ρ, considering random samples from a multivariate Normal population with $\Sigma = \Sigma_{AR}$ in (9).

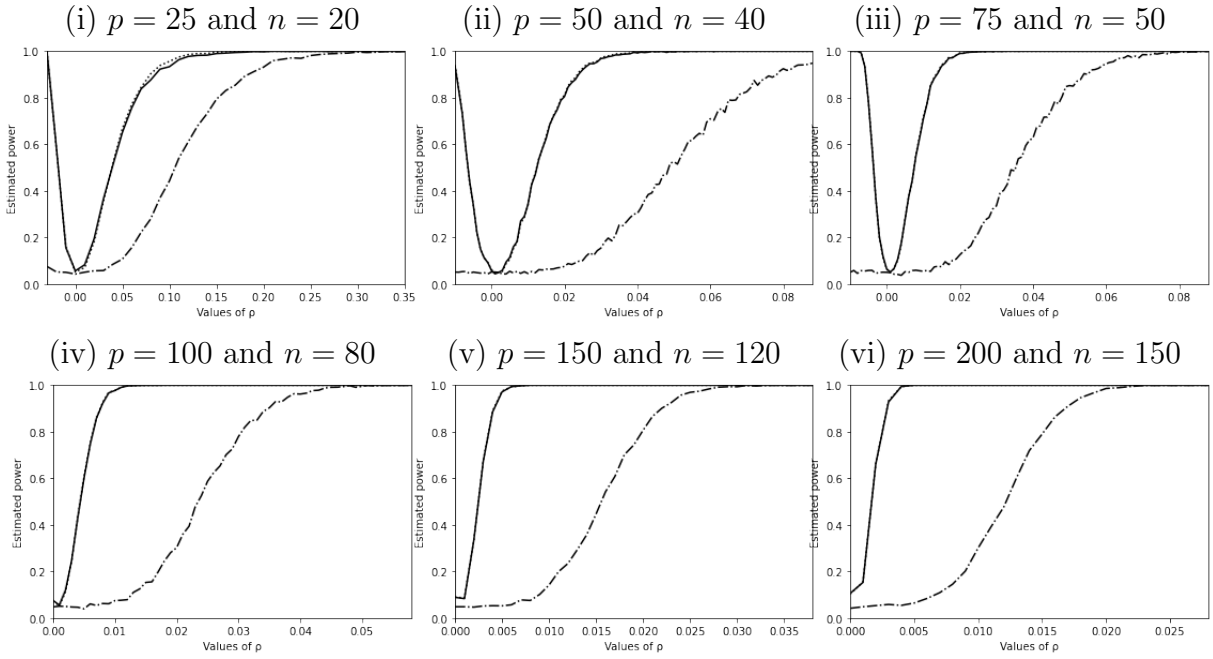

Figure 6: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively) and (ii) for the testing procedure presented in Schott (2005) (dotted-dashed line) as a function of ρ, considering random samples from a multivariate Normal population with $\Sigma = \Sigma_C$ in (10).

3.2 Significance levels and power estimates

From the plots presented in the previous subsection we may see that all the testing procedures considered in this work seem to be unbiased in the sense that under the null hypothesis, the estimated power is always in a narrow vicinity of the type I probability error. In this subsection, we analyze this property in more detail by comparing the estimated significance levels associated with the methodology described in Subsection 2.2 to the results presented in Tables 1 and 2 of Schott (2005). In Schott (2005), Table 1 gives the significance level results for his proposed methodology, whereas Table 2 displays the likelihood ratio test results. In this study, we replicate these tables using the univariate method described in Section 2.2 for $\alpha = 0.05$ in Table 1, and we also present results for $\alpha = 0.01$ in Table 2. Following Schott (2005), the estimated significance levels were derived from 5000 simulations of samples of size n randomly drawn from a multivariate Normal population with null mean vector and identity covariance matrix.

When we compare our Tables to the results of Schott (2005), we can see that:

- when p is small, the results of the new univariate method are slightly below the desired values of 0.05 or 0.01;
- the estimated significance levels converge to the desired values as n and p increase;
- in Table 1 in Schott (2005), mainly for small values of p, the estimated significance levels are slightly higher than 0.05, in contrast to what occurs with the proposed methodology in this study;
- for both $n < p$ and $n > p$, the outcomes of the new methodology are remarkably similar.

Table 1: Estimated type I error probabilities ($\alpha = 0.05$) for the univariate methodology outlined in Subsection 2.2 when the population variances are unknown.

p \ n	4	8	16	32	64	128	256
4	0.029	0.022	0.023	0.022	0.023	0.022	0.025
8	0.032	0.037	0.035	0.040	0.038	0.037	0.039
16	0.039	0.040	0.041	0.040	0.045	0.040	0.041
32	0.044	0.044	0.043	0.048	0.047	0.047	0.050
64	0.045	0.047	0.043	0.049	0.051	0.046	0.051
128	0.046	0.048	0.050	0.048	0.046	0.049	0.046
256	0.045	0.043	0.043	0.045	0.054	0.052	0.045

Table 2: Estimated type I error probabilities ($\alpha = 0.01$) for the univariate methodology outlined in Subsection 2.2 when the population variances are unknown.

p \ n	4	8	16	32	64	128	256
4	0.004	0.007	0.005	0.003	0.003	0.002	0.003
8	0.007	0.008	0.005	0.007	0.006	0.006	0.005
16	0.007	0.009	0.012	0.008	0.007	0.008	0.009
32	0.007	0.011	0.010	0.007	0.009	0.010	0.010
64	0.008	0.009	0.009	0.009	0.009	0.010	0.010
128	0.009	0.010	0.008	0.008	0.010	0.008	0.010
256	0.009	0.009	0.012	0.009	0.013	0.010	0.008

Schott (2005) estimated the power of the tests in Tables 3 and 4 for one specific choice of multivariate Normal distribution's parameters, more precisely, when the mean vector is null and the covariance matrix has the equivariance-equicorrelation structure in (8) for $\rho = 0.1$ and also $\sigma^2 = 1$. In Table 3 of this work, we compute the estimated power, for the same choice of parameters.

Table 3: Estimated power for the univariate methodology in Subsection 2.2 when the population variances are unknown and for samples extrated from a multivariate Normal distribution with null mean vector and covariance matrix with the equivariance-equicorrelation structure in (8) for $\rho = 0.1$ and $\sigma^2 = 1$.

p \ n	4	8	16	32	64	128	256
4	0.002	0.001	0.005	0.023	0.102	0.300	0.724
8	0.005	0.028	0.132	0.346	0.734	0.976	1.000
16	0.078	0.259	0.572	0.894	0.997	1.000	1.000
32	0.335	0.661	0.941	0.998	1.000	1.000	1.000
64	0.647	0.922	0.997	1.000	1.000	1.000	1.000
128	0.870	0.987	1.000	1.000	1.000	1.000	1.000
256	0.961	0.999	1.000	1.000	1.000	1.000	1.000

Comparing Table 3 with the findings of Schott (2005) results, the following can be observed:

- the estimated power of the testing procedure proposed in Subsection 2.2 of this paper is almost always higher than the values presented in Tables 3 and 4 of Schott (2005);
- the estimated power increases with the sample size and the number of variables;
- these results are consistent with the behavior depicted in Figures 1 and 4.

3.3 Departures from normality

In this section, we look at how well the tests discussed in this work handle deviations from normality. We consider samples from a multivariate t distribution with null mean vector, covariance matrix define in Eq.(8) and varying degrees of freedom, denoted by df. For each combination of p, n, and df, 2000 simulated samples were generated, and the estimated power was calculated. To keep the number of figures to a reasonable level, we ignore the structures defined in Eqs.(9) and (10). In Figure 7, we consider $n = 50$ and $p = 15$, and in Figure 8, we consider $n = 40$ and $p = 50$ in a high dimensional case. In these figures, we can see that the testing procedure proposed in this study appears to be more resistant to deviations from normality, mainly for small values of df. As expected, as df increases the estimated power approximates that obtained when the multivariate normal distribution was assumed in the previous subsections.

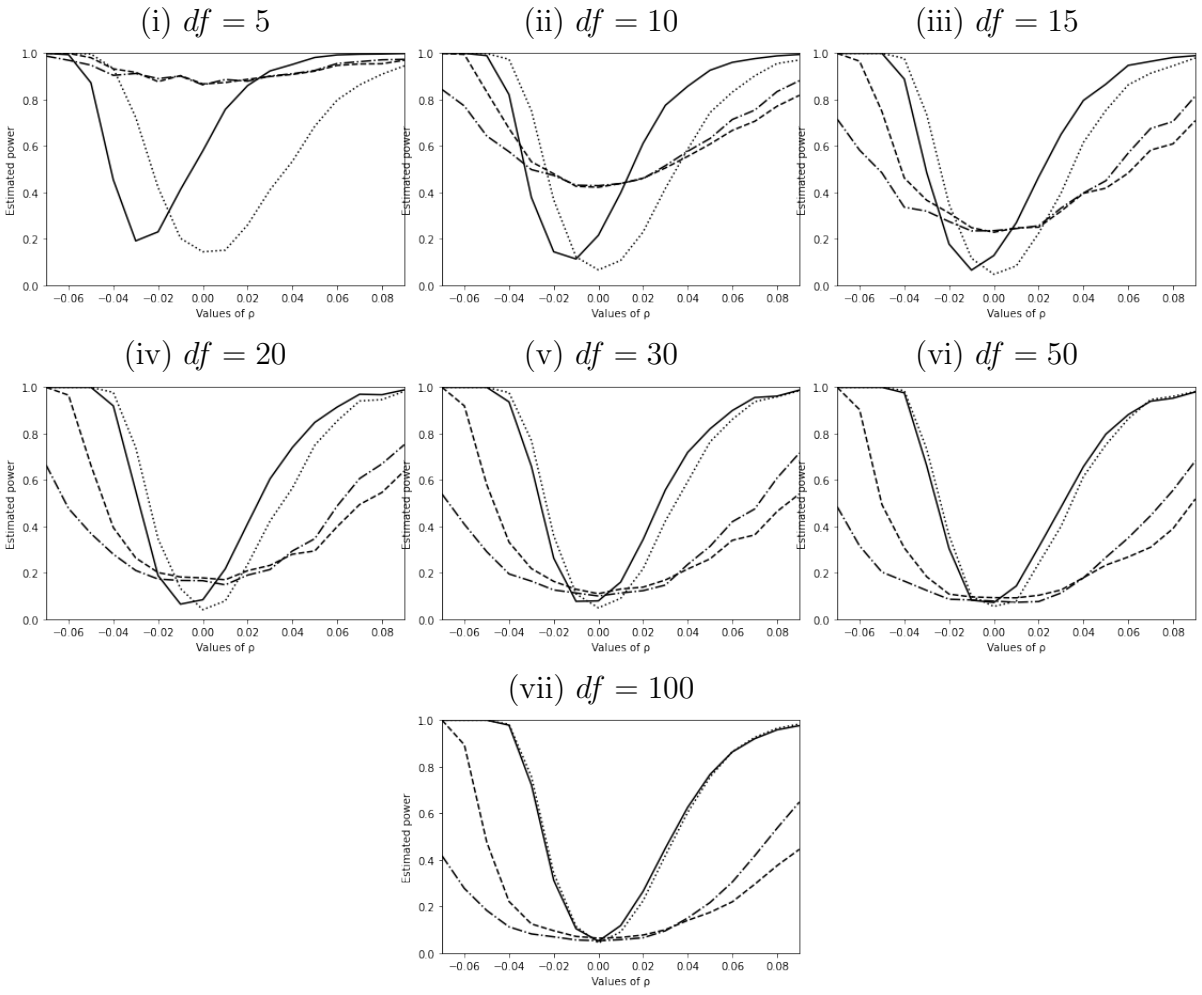

Figure 7: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively) and (ii) for the testing procedure presented in Schott (2005) (dotted-dashed line), as a function of ρ for samples of size $n = 50$, from a multivariate t distribution, with dimension $p = 15$, null mean vector, $\Sigma = \Sigma_{CS}$, defined in Eq. (8), and df.

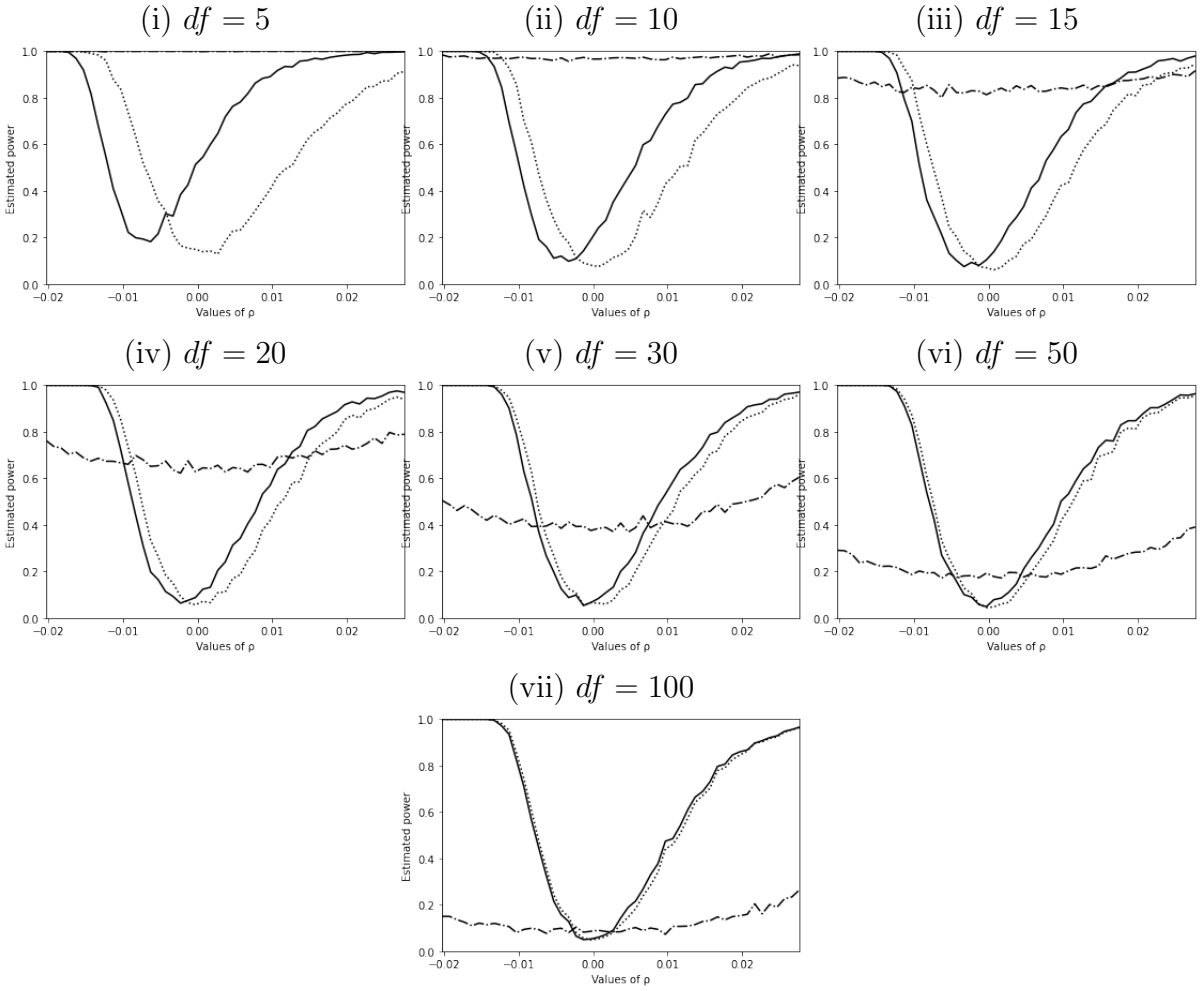

Figure 8: Estimated powers of the (i) univariate testing procedure, assuming known variances and unknown variances (solid and dotted lines, respectively) and (ii) for the testing procedure presented in Schott (2005) (dotted-dashed line), as a function of ρ for samples of size $n = 40$, from a multivariate t distribution, with dimension $p = 50$, null mean vector, $\Sigma = \Sigma_{CS}$, defined in Eq. (8), and df.

4 Likelihood ratio tests of complete independence for specific covariance structures

It is possible to develop likelihood ratio tests assuming specific covariance structures for the null and alternative hypotheses. The choice of covariance structures can not be done randomly. In fact, the structure considered under the null hypothesis must be a special case of the structure assumed for the alternative hypothesis. When a specific covariance structure is assumed, the choice of structures can be made so that the obtained likelihood ratio test is equivalent to a test of independence of a set of variables. Following are two examples of these tests, along with a comparison of the estimated powers to those obtained using the methodology described in Subsection 2.2.

4.1 The sphericity versus equivariance-equicorrelation test

The sphericity test (Mauchly, 1940) is a well-known test in multivariate analysis used to examine the equality of variances and the independence between all the variables of a multivariate Normal random vector. The equivariance-equicorrelation test, introduced in Wilks (1946), also known as the compound symmetry test, is useful for inferring whether the random variables of a multivariate Normal vector have the same variances and covariances. In this section, we consider the test of sphericity versus equivariance-equicorrelation whose hypotheses, for a p-variate Normal population, $\underline{X} \sim N_p(\underline{\mu}\,, \Sigma)$, are represented as

$$H_0 : \Sigma = \sigma^2 I_p = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \quad \text{vs.} \quad H_1 : \Sigma = \Sigma_{CS} = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \cdots & \sigma^2 \end{bmatrix}. \tag{11}$$

We should note that $\Sigma_{CS} = \sigma^2((1-\rho)I_p + \rho J_p)$, with $-1/(p-1) < \rho < 1$, I_p is the identity matrix of order p and J_p denotes a matrix of order p with all elements equal to one. Under the alternative hypothesis, the considered structure is the one in Eq.(8). This test was initially discussed by Marques and Coelho (2015). In this work, the authors developed the likelihood ratio test statistic, the expression of its null moments, and the characteristic function of the negative logarithm of the test statistic (see expressions (2) and (3) in Marques and Coelho (2015) for details). In addition, the authors also developed sharp and manageable asymptotic approximations that facilitate the use of this test and guarantee the accuracy of the results.

The testing procedure presented in this study's Subsection 2.2 can be used to test the null hypotheses in Eq.(11). In fact, if we assume that we are working in the space of the covariance matrices with the structure assumed in Eq.(11) under alternative hypothesis, we can always use the testing procedure proposed in Subsection 2.2 because all covariances have the same sign. Additionally, failing to reject the sphericity structure of Σ for the test of null hypotheses in (11) is equivalent to failing to reject the independence of the variables.

Simulations will be used to compare the estimated power of the test developed by Marques and Coelho (2015) and the one proposed in this study. Two additional notes: the test for complete independence described in Subsection 2.2 can be used in a high dimensional setting, unlike the usual likelihood ratio tests, and it can be applied to hypotheses that are more general than those underlying alternative hypothesis in Eq.(11).

In Figure 9, we present the estimated powers of the likelihood ratio test and the univariate testing procedure described in Subsection 2.2, with known and unknown variances. This figure shows that the estimated powers are very similar and exhibit the same pattern, mainly for $p > 5$. Despite this similarity, it is important to note that: (i) the new procedure is based on a simple χ^2 distribution, making it easy to implement, and (ii) unlike the likelihood ratio test, the new procedure can be used in high dimensional setting.

4.2 The sphericity versus circularity test

In this subsection, we present a test that also considers the sphericity structure under the null hypothesis, but for the alternative hypothesis, we assume the circular covariance structure

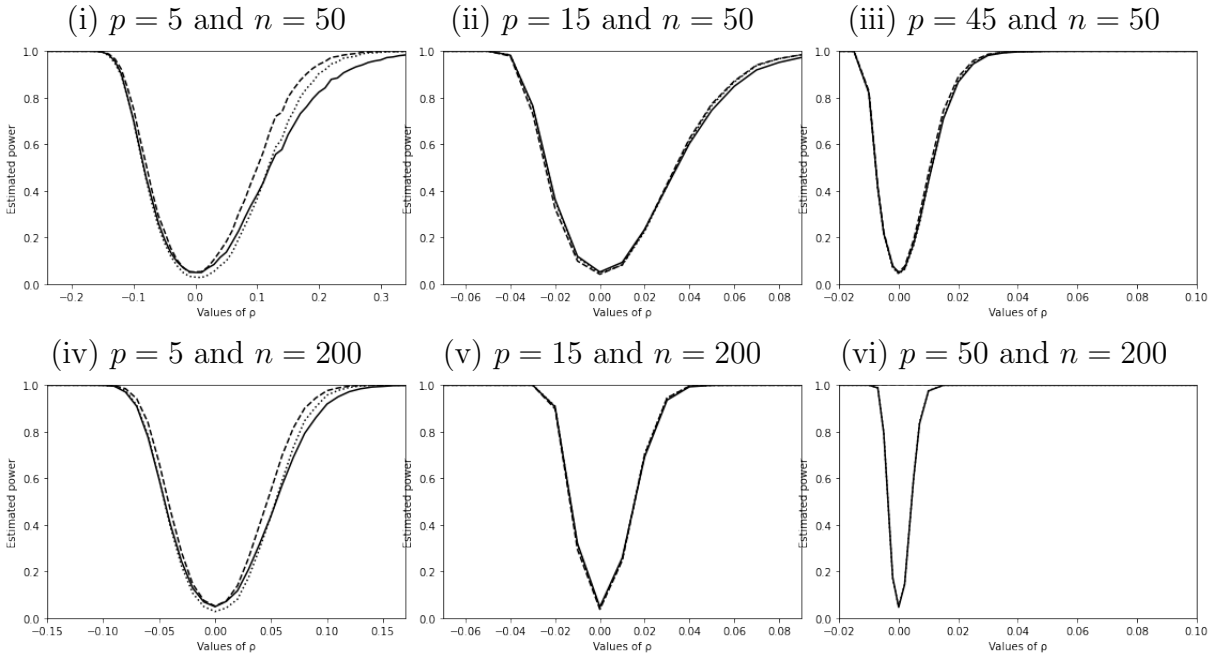

Figure 9: Samples from a multivariate Normal population with $\Sigma = \Sigma_{CS}$; plots of the estimated powers of the univariate testing procedures, with known and unknown variances (solid and dotted lines, respectively), likelihood ratio test in Marques and Coelho (2015) (dashed).

defined in (10). These types of structures may be important in different fields of research such as biological sciences, psychometry, quality control, signal detection, spatial statistics, time series analysis and engineering (Khattree, 1996; Gray, 2006). Thus, we are interested in testing:

$$H_0 : \Sigma = \sigma^2 I_p = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \quad \text{vs.} \quad H_1 : \Sigma = \Sigma_C = \begin{bmatrix} \sigma^2 & b_1 & \ldots & b_{p-1} \\ b_{p-1} & \sigma^2 & \ldots & b_{p-2} \\ \vdots & \vdots & & \vdots \\ b_1 & b_2 & \ldots & \sigma^2 \end{bmatrix}. \tag{12}$$

If the decision is not to reject the null hypothesis, then the independence of the variables is also not rejected when the underlying structure is circular. This test was addressed and introduced in Olkin and Press (1969). The authors show that the likelihood ratio test statistic raised to the power of $2/N$ is given by

$$\Lambda^{2/N} = p^p 2^{-2(p-m-1)} \prod_{j=1}^{p} v_j \left(\sum_{i=1}^{m+1} v_i \right)^{-p}$$

where $m = \lfloor p/2 \rfloor$ and v_j are given by (2.5a) and (2.5b) in Olkin and Press (1969). Additionally, the authors show that the exact distribution of Λ may be represented as the product of independent random variables with Beta distributions and provide an approximation based on a mixture of χ^2 distributions.

In Figure 10, we present the results for the estimated powers of the likelihood ratio test in Olkin and Press (1969) and for the univariate testing procedure in Subsection 2.2, with known and unknown variances.

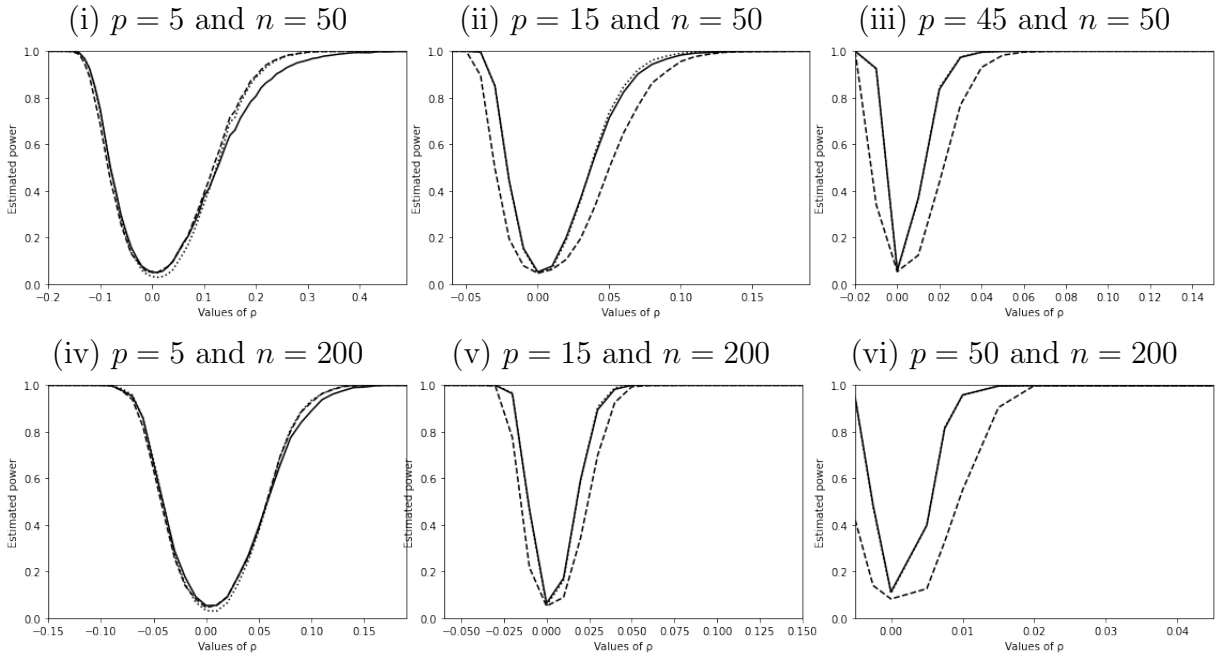

Figure 10: Samples from a multivariate Normal population with $\Sigma = \Sigma_C$; plots of the estimated powers of the univariate testing procedures, with known and unknown variances (solid and dotted lines, respectively), likelihood ratio test in Olkin and Press (1969) (dashed).

Figure 10 shows that for $p > 5$, the estimated power values for the new procedure appear to be greater than those obtained for the likelihood ratio test by Olkin and Press (1969). In addition, as already mentioned, the new procedure is based on a simple χ^2 distribution and is applicable in the high dimensional setting.

5 Conclusions

In this paper, we propose a new method for testing the independence of a set of variables that can be applied to specific covariance structures. More precisely, the methodology can be employed whenever the covariances are all non-positive or non-negative.

A comprehensive simulation study was conducted to evaluate the properties of the new testing methodology in comparison to other different likelihood ratio tests and the test proposed in Schott (2005).

The simulations show that the new testing methodology is more effective than either the likelihood ratio test in Subsection 2.1 or the test proposed in Schott (2005). Moreover, the methodology proposed in Subsection 2.2 yields excellent results for high dimensional scenarios. In comparison to the likelihood ratio tests in Section 4, the new methodology yields comparable results for the estimated power, with the added benefit of being applicable

to high dimensional scenarios. The proposed method can also be used when the data has missing values and seems more robust for violations of the normality assumption. It is shown that the new test is slightly biased, especially for small values n and p.

It should be noted that the power studies for the (AR) structure show different results and patterns than the ones for the (CS) or (C) structures. The justification may be related to the fact that, opposite to the other two structures, for the (AR) structure when ρ increases the values of $\rho^{|p-j|}$, $j = 1, \ldots, p-2$ decrease. Nevertheless, more studies are required to better understand these results.

This work conclusions are based on numerical studies. The theoretical proofs requires results for the distribution of the statistics under the alternative hypothesis which are not available. These proofs will be addressed in future works.

As a final comment, it seems that the new test presented in Subsection 2.2 has excellent asymptotic properties, being unbiased for large values of n and p and exhibiting large values of the estimated power for small deviations from the null hypothesis. It is a very simple procedure can be applied to real-world problems when it is reasonable to assume that the all covariances have the same sign. This assumption is easily verifiable with some theoretical understanding of the problem and the variables involved, which can also be confirmed by computing the sample covariance matrix.

Acknowledgements

This work is funded by national funds through the FCT – Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications). This work does not have any conflicts of interest.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd edn. Wiley, New York.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria, *Biometrika*, **36**, 317-346.

Butler, R.W., Huzurbazar, S., and Booth, J. G.. (1993) Saddlepoint approximations for tests of block independence, sphericity and equal variances and covariances, *Journal of the Royal Statistical Society: Series B*, **55**, 171–183.

Chan, J. and Choy, B. Analysis of covariance structures in time series. *J Data Sci.*, **6**, 573-589.

Coelho, C. A. (2004). The Generalized Near-Integer Gamma Distribution: A Basis for 'Near-Exact' Approximations to the Distribution of Statistics which are the Product of an Odd Number of Independent Beta Random Variables. *Journal of Multivariate Analysis*, **89**, 191-218.

Coelho, J. T. and Marques, F. J. (2010). Near-exact distributions for the independence and sphericity likelihood ratio test statistics, *Journal of Multivariate Analysis*, **101**, 583-593.

Gray, R. M. (2006) Toeplitz and Circulant Matrices: A Review. Foundations and Trends, *Communications and Information Theory*, **2**, 155–239.

Khattree, R. (1996) Multivariate statistical inference involving circulant matrices: a review, in: Gupta, A. K., Girko, V. L. (Eds.) Multidimensional Statistical Analysis and Theory of Random Matrices 101–110, VSP, Netherlands.

Ledoit, O. and Wolf, M. (2002). Hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, **30**, 4, 1081–1102.

Li, W. and Liu, Z. (2016). A test for the complete independence of high-dimensional random vectors, *Journal of Statistical Computation and Simulation*, **86**(16), 3135-3140.

Littell, R.C., Pendergast, J., Natarajan, R. Modelling covariance structure in the analysis of repeated measures data. *Stat Med.*, **19**, 1793-1819.

Marques F. J. and Coelho, C. A. (2015). The sphericity versus equivariance-equicorrelation test. *AIP Conference Proceedings*, **1648**, 540009.

Marques, F. J. and Coelho, C. A. (2020). Testing simultaneously different covariance block diagonal structures – the multi-sample case. *Journal of Applied Statistics*, **47**, 2765–2784.

Marques, F. J., Coelho, C. A., and Arnold, B. C. (2010). A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis. *Test* **20**, 180-203.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, **11**, 204–209.

Mudholkar, G.S., Trivedi, M.C., and Lin, C.T. (1982). An approximation to the distribution of the likelihood ratio statistic for testing the complete independence. *Technometrics*, **24**, 139–143.

Olkin, I. and Press, S. J. (1969). Testing and estimation for a circular stationary model. *The Annals of Mathematical Statistics*, **40**, 1358-1373.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis*, Third Edition, John Wiley & Sons, Inc.

Schott, J. R. (2005). Testing for Complete Independence in High Dimensions. *Biometrika*, **92**, 4, 951-956.

Srivastava, M. S. (2005). Some Tests Concerning the Covariance Matrix in High Dimensional Data, *Journal of the Japan Statistical Society*, **35**(2), 251-272.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, *The Annals of Mathematical Statistics*, **9**, 60-62.

Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a Normal multivariate distribution. *The Annals of Mathematical Statistics*, **17**, 257–281.