Revista Colombiana de Estadística - Applied Statistics
January 2021, volume 44, issue 1, pp. 43 a 64
http://dx.doi.org/10.15446/rce.v44n1.83987

Comparison of Correction Factors and Sample Size Required to Test the Equality of the Smallest Eigenvalues in Principal Component Analysis

Comparación de los factores de correción y tamaños de muestra requeridos para probar la igualdad de los valores propios más pequeños en el análisis de componentes principales

Eduard Gañan-Cardenas[1,a], Juan Carlos Correa-Morales[2,b]

[1]Departamento de Calidad y Producción, Facultad de Ciencias Económicas y Administrativas, Instituto Tecnológico Metropolitano, Medellín, Colombia

[2]Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Sede Medellín, Colombia

Abstract

In the inferential process of Principal Component Analysis (PCA), one of the main challenges for researchers is establishing the correct number of components to represent the sample. For that purpose, heuristic and statistical strategies have been proposed. One statistical approach consists in testing the hypothesis of the equality of the smallest eigenvalues in the covariance or correlation matrix using a Likelihood-Ratio Test (LRT) that follows a χ^2 limit distribution. Different correction factors have been proposed to improve the approximation of the sampling distribution of the statistic. We use simulation to study the significance level and power of the test under the use of these different factors and analyze the sample size required for an adequate approximation. The results indicate that for covariance matrix, the factor proposed by Bartlett offers the best balance between the objectives of low probability of Type I Error and high Power. If the correlation matrix is used, the factors W_B^* and $c\chi_d^2$ are the most recommended. Empirically, we can observe that most factors require sample sizes 10 or 20 times the number of variables if covariance or correlation matrices, respectively, are implemented.

***Key words*:** Chi-square distribution; Likelihood ratio test; Power comparisons; Principal components analysis; Sphericity test.

[a]M.Sc. E-mail: eduardganan@itm.edu.co
[b]Ph.D. E-mail: jccorrea@unal.edu.co

Resumen

Dentro del proceso inferencial del Análisis de Componentes Principales (PCA) uno de los interrogantes principales de los investigadores es sobre el número correcto de componentes para representar la muestra. Para este fin se han propuesto estrategias heurísticas y estadísticas. Un enfoque estadístico consiste en probar la hipótesis sobre la igualdad de los valores propios más pequeños de la matriz de covarianza o correlación a través de una prueba de razón de verosimilitud (LRT) que sigue una distribución límite χ^2. Diferentes factores de corrección han sido propuestos para mejorar la aproximación de la distribución muestral del estadístico. En este trabajo utilizamos simulación para estudiar el nivel de significancia y la potencia de la prueba bajo el uso de estos diferentes factores, así como una revisión del tamaño de muestra requerido para una adecuada aproximación. Los resultados para la matriz de covarianza indican que el factor propuesto por Bartlett ofrece el mejor equilibrio entre los objetivos de baja probabilidad de Error Tipo I y alta potencia. En caso de la matriz de correlación, los factores W_B^* y $c\chi_d^2$ son los más recomendados. Empíricamente se observa que la mayoría de los factores requieren tamaños de muestra 10 y 20 veces mayores al número de variables en caso de la matriz de covarianza o de correlación respectivamente.

***Palabras clave*:** Análisis de componentes principales; Comparación de potencias; Distribución Chi-cuadrado; Prueba de esfericidad; Prueba de razón de verosimilitud.

1. Introduction

Principal Component Analysis (PCA) is a multivariate technique used to reduce data dimensionality. During the inference process of PCA for a sample composed of $\boldsymbol{p}$ original variables, questions arise about the adequate $\boldsymbol{k}$ number of components to represent the data and the adequate sample size to produce the inference (Krazanowski, 1988). For instance, Chakraborty et al. (2020) used the Bartlett's test of sphericity in a correlation matrix for the construction of socioeconomic index based on PCA in the field of environmental justice. Similary Şahan et al. (2018) used the same test for the validation of a psychological questionnaire. PCA can also be used as an intermediate step in a prediction task. For example, Maté (2011) used PCA to generate combined forecasts by identifying the underlying structure within a set of prediction methods.

In the inferential context, principal components are no longer a strictly mathematical procedure to become a statistical method. The objective of obtaining a smaller dimension to represent the data is affected by the sample error. This sample error can lead to misrepresentation of the data. Such as the non-inclusion of components with relevant information (underestimation), or the inclusion of noise components (overestimation), causing a distortion in the analysis (Peres-Neto et al., 2005). As Björklund (2019) pointed out, when a study requires to extract a number of components, the differentiation between the eigenvalues must be previously tested before proceeding with the analysis since the patterns found may correspond to a simple sampling error of the correlations.

Several strategies have been adopted to define the k number of principal components that should be retained. In this regard, multiple works can be found comparing different methods with respect to their ability to identify the true number of non-trivial components (Ferré, 1995; Jackson, 1993; Peres-Neto, Jackson & Somers, 2005). For example, Jackson (1993) compared heuristic and statistical methodologies used to define the number of components. He compared Kaiser-Guttman, Bootstrapped Kaiser-Guttman, Scree Plot, Modified Scree Plot, Percentage of explained variation, and those based on hypothesis testing. Regarding the statistical approach, Jackson (1993) concluded that Bartletts test of sphericity, based on the hypothesis of the equality of the remaining $\boldsymbol{p-k}$ eigenvalues in the covariance matrix, correctly identified true dimensionality in many data sets. But it showed inconsistent results with matrices having a low observation-to-variable ratio (less than 3:1 ratio). Similarly, Peres-Neto et al. (2005) made a comparison of methods and proposed a two-stage selection strategy, using Bartlett's sphericity test to identify the significance of the first component. Later, different rules can be applied to validate the other components. However, it should be noted that in most of these works comparing methods or applying Bartlett's sphericity test, only one version of the test correction factors is used. Although, multiple correction factors have been proposed.

This study focuses on the analysis of a methodology based on a hypothesis testing process also known as isotropic test or equality of variance test of the $(p-k)$ last principal components. This is an important method in literature, which has even inspired graphic methods such as the scree-plot (Ferré, 1995). In this statistical method, the null hypothesis of interest is defined as $H_{0k} : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p = \lambda$, against H_{ak} : *some of them are different*. Where λ represents the unknown common value and λ_i is the population eigenvalue of the $i-th$ component obtained from the covariance or correlation matrix. The test tries to find evidence that the smallest $p-k$ last population eigenvalues are equal and could even be considered to be simple measurements of noise (Schott, 1988). Then, accepting H_{0k} means that, if more than k components are included, all the p components should be included because each one of the remaining components contains the same amount of information. H_{0k} is tested in a sequential manner starting with $k = 0$, and increasing k until the hypothesis is accepted (Mardia, Kent & Bibby, 1979; Krazanowski, 1988). To evaluate the hypothesis, it is used a Likelihood-Ratio Test (LRT), which, under H_{0k}, presents an χ^2 asymptotic distribution (Mardia, Kent & Bibby, 1979; Krazanowski, 1988). Alternatively, Schott (2012) proposed a new limiting distribution based on Saddlepoint approximations when Chi-square distribution is not adequate, but that scenario was not considered in this study.

To improve the approximation of the sampling distribution of the statistic to its distribution limit, several correction factors have been proposed; said factors change if the PCA is conducted based on the covariance matrix or the correlation matrix. For that reason, this work uses a simulation to compare different correction factors that have been proposed for the Likelihood-Ratio statistic when the test of equality of eigenvalues is used in PCA, whether with a covariance or correlation matrix. The comparison considers the number of variables, the number of components, and the sample size in order to recommend to PCA users which

factor to employ and under what conditions it would be adequate to do so. We also study the power of the test under the different factors, in order to obtain a complete view on the performance of the method.

This article is organized as follows: First, we present the test and different correction factors proposed for the covariance and correlation matrices. Afterward, we describe the simulation scheme and illustrate the process. Finally, we report the results of the simulation and draw some conclusions. In the conclusions, we highlight a series of recommendations regarding the test statistics to be used in PCA.

2. Tests of Equality of Eigenvalues

2.1. Test of Equality of Eigenvalues for the Covariance Matrix

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample of an n size taken from a normal $p-variate$ distribution with unknown population vector of μ means and Σ population covariance matrix. Let $\lambda_1 \geq \cdots \geq \lambda_p > 0$ be the population eigenvalues of Σ and $l_1 > l_2 > \cdots > l_p$ be the sample eigenvalues of sample covariance matrix S, with an n sample size. The test statistic to evaluate the hypothesis of equality of the smallest $p-k$ eigenvalues $H_{0k} : \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p = \lambda$, based on the sample covariance matrix, is given by (Mardia, Kent & Bibby, 1979; Krazanowski, 1988):

$$W = n' \left\{ (p-k) \log \left[\sum_{i=k+1}^{p} \frac{l_i}{(p-k)} \right] - \sum_{i=k+1}^{p} \log(l_i) \right\} \tag{1}$$

under true H_{0k}, W has approximately a χ^2 distribution with $\frac{1}{2}(p-k+2)(p-k-1)$ degrees of freedom, where n' is replaced by n or $n-1$, which are represented as W_n and W_{n-1}, respectively.

To achieve a better χ^2 approximation, n' is replaced with the next correction factor (Mardia, Kent & Bibby, 1979; Krazanowski, 1988), which is known as Bartletts Test of Sphericity:

$$FC2_{Bartlett} = n - \frac{2p+11}{6} \tag{2}$$

Lawley (1956) claims that a better χ^2 approximation is achieved if n' is replaced by the correction factor:

$$FC3_{Lawley} = n - k - \frac{1}{6}\left(2q + 1 + \frac{2}{q}\right) + \lambda^2 \sum_{i=1}^{k} \frac{1}{(\lambda_i - \lambda)^2} \tag{3}$$

where, for practical purposes, λ_i is replaced by l_i, $q = p-k$, and λ is estimated as $\widehat{\lambda} = \frac{\sum_{i=k+1}^{p} l_i}{p-k}$.

In Jackson (1993), n' is replaced by

$$FC4_{Jackson} = n - k \tag{4}$$

Likewise, in Ferré (1995), correction factor $FC2_{Bartlett}$ is written as:

$$FC5_{Ferre} = n - k - 1 - \frac{2p + 11}{6} \tag{5}$$

To test the hypothesis that all the variables are independent and have an equal variance (Jolliffe, 2002), that is, H_{0k} with $k = 0$, correction factor $FC3_{Lawley}$ becomes (Bartlett, 1954; Lawley, 1956):

$$n - \frac{1}{6}\left(2p + 1 + \frac{2}{p}\right) \tag{6}$$

Not rejecting hypothesis H_{00} would mean that it is not possible to reduce the dimensionality of the data at all.

2.2. Test of Equality of Eigenvalues for the Correlation Matrix

When PCA is based on variables that have been standardized, the hypothesis that all the eigenvalues in the population correlation matrix P are the same is equal to the hypothesis that $P = I$ (Mardia et al., 1979), that is, proving that H_{0k} with $k = 0$, which means that all the variables are independent without implying that the variances are the same (Jolliffe, 2002). To support this hypothesis, Mardia, Kent & Bibby (1979) introduced the following LRT in terms of sample correlation matrix R:

$$L_R = -n \cdot \log |R| \tag{7}$$

which, under H_{00}, has a χ^2 distribution with $\frac{1}{2}p(p-1)$ degrees of freedom. Box (1949) suggested a new correction factor to improve the χ^2 approximation, which is presented in Mardia, Kent & Bibby (1979), replacing n by:

$$FC1_{LR} = n - \frac{2p + 11}{6} \tag{8}$$

To conduct the same test, Bartlett (1954) present the following correction factor:

$$FC2_{LR} = n - \frac{2p + 5}{6} \tag{9}$$

Now, we are also interested in testing the hypothesis that the smallest $p - k$ eigenvalues of P are equal, where $0 < k < p - 1$. Mardia, Kent & Bibby (1979) examine the following statistic suggested by Bartlett (1951):

$$W^* = n' \left\{ (p-k) \log \left[\sum_{i=k+1}^{p} \frac{l_i}{(p-k)} \right] - \sum_{i=k+1}^{p} \log(l_i) \right\}, \tag{10}$$

where n' is replaced by $n-1$ or n (Lawley, 1956; Mardia, Kent & Bibby, 1979; Schott, 1988), which is represented as W_n^* and W_{n-1}^*, respectively. However, this statistic is not χ^2 asymptotically distributed, although it could be approximated if $\lambda_1, \lambda_2, \ldots, \lambda_k$ are big in relation to λ with a maximum number of degrees of freedom equal to $\frac{1}{2}(p-k+2)(p-k-1)$ (Bartlett, 1954; Lawley, 1956; Mardia et al., 1979).

To improve the approximation to the limit distribution, we calculate the W_B^* statistic, where n' is replaced by the following correction factor B (Bartlett, 1954; Jackson, 1991):

$$B = n - \frac{1}{6}(2p+5) - \frac{2}{3}k \tag{11}$$

Lawley (1956) improved the approximation presented by Bartlett (1954) under the same assumption of normality and calculating the effective number of degrees of freedom in a general case for the W^* statistic, taking $n' = n$:

$$\begin{aligned} \mu_{W^*} = {} & \frac{1}{2}(q-1)(q+2) \\ & - \frac{1}{q} \left[2(q-1)\lambda \sum_{i=1}^{p} c_{ii}^2 - q \sum_{i=1}^{p} \sum_{j=1}^{p} (c_{ij}^2 r_{ij}^2) + \sum_{i=1}^{p} \sum_{j=1}^{p} (c_{ii} c_{jj} r_{ij}^2) \right] \end{aligned} \tag{12}$$

where c_{ij} are the elements of $C = I - Q_1 Q_1^{'}$ (Q_1 can be estimated as the matrix of the eigenvectors of the corresponding $l_1, l_2, \ldots, l_k$ of R); r_{ij} denotes the correlation between x_i and x_j; $q = p - k$, and λ is estimated as $\widehat{\lambda}$. The W^* statistic, based on the degrees of freedom μ_{W^*}, is denoted as $\chi^2_{\mu_{W^*}}$.

Schott (1988) extended the study by Lawley (1956), which shows the way to obtain a new statistic of the form $c\chi_d^2$ proposed by Anderson (1963). From this, we obtain that $c = \frac{1}{2}\sigma_{W^*}^2/\mu_{W^*}$ and $d = 2\mu_{W^*}^2/\sigma_{W^*}^2$, where μ_{W^*} and $\sigma_{W^*}^2$ are the mean and variance of W^*. That is, μ_{W^*} is the result obtained by Lawley (1956) (see Equation 12), and the variance of W^* is:

$$\begin{aligned} \sigma_{W^*}^2 = {} & (q-1)(q+2) - 8\lambda \left(\frac{q-1}{q} \right) \sum_{i=1}^{p} c_{ii}^2 + 4 \sum_{i=1}^{p} \sum_{j=1}^{p} r_{ij}^2 \left(c_{ij}^2 - \frac{1}{q} c_{ii} c_{jj} \right) \\ & + 8\lambda^2 \sum_{i=1}^{p} \sum_{j=1}^{p} f_{ijij} - 8\lambda \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{\alpha=1}^{p} f_{i\alpha j\alpha} r_{ij}^2 + 2 \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{\alpha=1}^{p} \sum_{\beta=1}^{p} f_{i\beta j\alpha} r_{ij}^2 r_{\alpha\beta}^2, \end{aligned} \tag{13}$$

where $f_{i\beta j\alpha} = c_{i\beta}^2 c_{j\alpha}^2 - 2q^{-1} c_{j\alpha}^2 c_{ii} c_{\beta\beta} + q^{-2} c_{ii} c_{jj} c_{\alpha\alpha} c_{\beta\beta}$, $q = p - k$, c_{ij}, and r_{ij} are obtained as mentioned for Equation 12, and λ is estimated as $\widehat{\lambda}$.

3. Simulation Scheme

3.1. Simulation Scheme for the Covariance Matrix

The correction factors will be evaluated based on their distribution approximation to the corresponding limit distribution by comparing the level of nominal significance and the significance estimated for the specific statistic. For significance level simulation we use $p = 5, 10, 15, 30$ and $n = 10, 30, 50, 100, 200, 500$. We take $n \geq p + 1$ so that the sample eigenvalues of S are positive. We performed 100,000 simulations under the following sample generation process.

n-sized samples are taken from a distribution $N_p(0, \Sigma)$. By considering the covariance matrix $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_k, \lambda, \ldots, \lambda)$ a diagonal matrix of the population eigenvalues, generality is not lost because the eigenvalues of S are the same as those of $G^{'}SG$ for any orthogonal matrix G. Moreover, the W statistic is invariant to multiplications of S by a positive scalar. Therefore, under H_{0k}, we can assume $\Sigma = \text{diag}(\lambda_1/\lambda, \ldots, \lambda_k/\lambda, 1, \ldots, 1)$, which, for simplicity, is written as (Waternaux, 1984; Schott, 2006; Fujikoshi et al., 2007; Watanabe et al., 2008):

$$\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_k, 1, \ldots, 1) \tag{14}$$

The configuration of the population eigenvalues can be defined as follows (Schott, 2006; Fujikoshi et al., 2007; Watanabe et al., 2008):

i. If $k = 0$, then $\Sigma = I_p$. That is, all the components explain the same amount of variability.

ii. If $k = 2$, then $\lambda = 1$ and $\lambda_i = \frac{a_i(p-k)}{1-\sum_{j=1}^{k} a_j}$ with $a_1 = 0.56$, $a_2 = 0.24$. Thus, the first two components explain 80% of the total variation.

iii. If $k = 3$, then $\lambda = 1$ and $\lambda_i = \frac{a_i(p-k)}{1-\sum_{j=1}^{k} a_j}$ with $a_1 = 0.45$, $a_2 = 0.3$ $a_3 = 0.15$. In this case, the first three components explain 90% of the total variation.

The nominal significance level was set at $\alpha = 0.05$. As the estimated significance level approaches the nominal value, the sampling distribution of the statistic is considered to achieve a better approximation to its limit distribution (Waternaux, 1984; Schott, 2006; Fujikoshi et al., 2007; Watanabe et al., 2008). To calculate the estimated significance level, we first define the quantile of the limit distribution as $W_C = \chi^2_{\frac{1}{2}(p-k+2)(p-k-1),1-\alpha}$ with $\alpha = 0.05$. Afterward, we generate a sample based on a pre-established population configuration, calculate the test statistic, and check if $W \geq W_C$; that is, under a defined k, we check if H_{0k} is rejected being true. The same process is completed as many times as the number of simulations above; as a result, we obtain the number of times that the null hypothesis is rejected being true. That quantity is divided by the number of conducted simulations; thus, we obtain the percentage of times that Type I errors

were produced, which is defined as the estimated significance level. The simulation process was programmed in R Core Team (2019).

3.2. Simulation Scheme for the Correlation Matrix

The simulation scheme considered in this study is similar to that proposed above for the covariance matrix. The difference lies in the fact that we should define a covariance matrix where the variances of all the variables equal 1, and that $\lambda_1 + \cdots + \lambda_p = p$. For that purpose, we implemented the study by Arteaga & Ferrer (2010), in which they proposed an algorithm to obtain a covariance matrix with the eigenvalues and the specified variances. Likewise, the configurations of the population eigenvalues used in the covariance matrix are redefined so that they meet the previous constraint regarding the sum of the eigenvalues, but maintaining the same percentages of explained variation.

3.3. Simulation Scheme for Power of the Test

For the study of the power of the test, the following alternatives were designed with deviations from the null hypothesis given by $\delta = 0.5, 1, 1.5$ (Waternaux, 1984).

i. To prove $k = 0$, $H_{00} : \lambda_1 = \lambda_2 = \ldots = \lambda_p = 1$, when really $\lambda_1 = 1 + \delta$

ii. To prove $k = 2$, $H_{02} : \lambda_3 = \lambda_4 = \ldots = \lambda_p = 1$, when really $\lambda_3 = 1 + \delta$

The scenarios of $k = 0$ and $k = 2$ were evaluated with $p = 10$ and $p = 30$ and sample sizes $n = 30, 50, 100, 200, 500$. For the $k = 2$ scenario, we use the configuration of variance explained with $a_1 = 0.56, a_2 = 0.24$, for λ_1 and λ_2 respectively. This scheme leaves an unexplained 20% variance in the remaining components, seeking to make the identification of the different component more demanding. For scenario $k = 0$, similar to the power test performed by Knapp & Swoyer (1967), the eigenvalues of $\lambda_1 = 1.5, 2, 2.5$ represent overall correlation coefficients of 0.06, 0.11 and 0.17 (Friedman, 1981). To evaluate the factors of the correlation matrix, the eigenvalues were adjusted to the condition that all variances are equal to 1 and that $\lambda_1 + \cdots + \lambda_p = p$. The analysis focuses on the factors that show the best performance in significance analysis. The simulation process consisted in counting the number of times that H_{0k} is rejected considering that it is not true. To do this we define the critical value W_C with $\alpha = 0.05$.

3.4. Illustration

Figure 1 illustrates the sampling distribution W under different sample sizes, $p = 5$, $k = 0$, and replacing $n^{'} = n$, denoted as W_n. The solid line represents the limit distribution $\chi^2_{\frac{1}{2}(p-k+2)(p-k-1)}$, and the dark area denotes the estimated significance level, which is marked as $\hat{\alpha}$. Figure 1 shows that, as the sample size n increases, the sampling distribution of the W_n statistic moves closer to the limit distribution. This would lead us to conclude, in this case, that a sample size of

$n = 100$ would produce a good approximation of the statistic distribution and, therefore, an estimated significance level very close to the nominal value $\alpha = 0.05$.

In general, correction factors are considered to generate a good approximation or performance if the estimated significance level is close to the nominal level; and, among them, we select the factor that presents the lowest estimation. The purpose is to obtain the factor that produces the lowest error level.

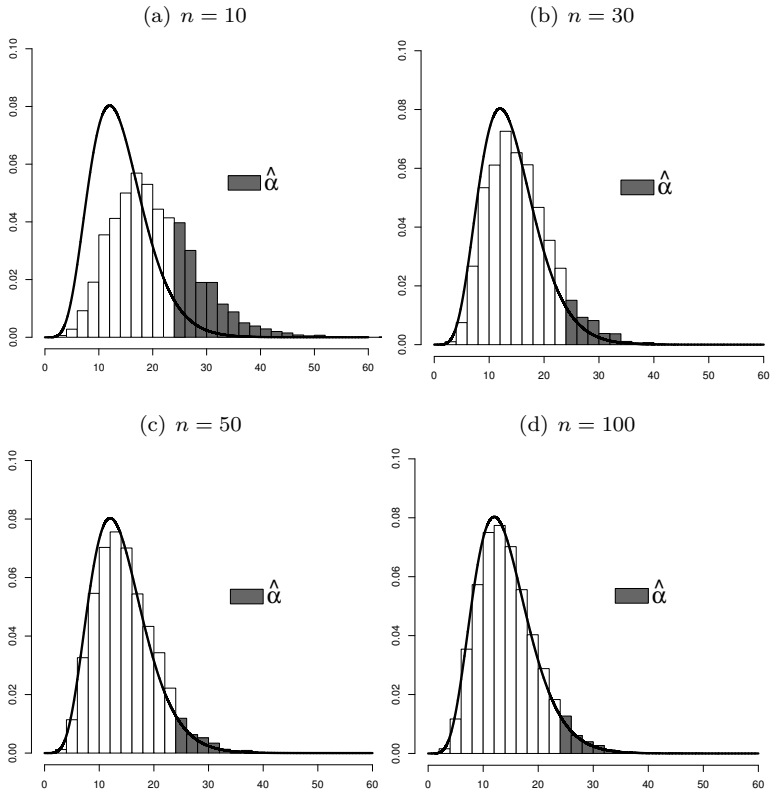

FIGURE 1: Comparison between the sampling distribution of the W_n statistic and its limit distribution for the case of $p = 5$ and $k = 0$. The dark area represented by $\hat{\alpha}$ indicates the estimated significance level, that is, the proportion of real rejection obtained with the W_n statistic under a scenario of p, n and k.

4. Simulation Results

4.1. Simulation Results in the Case of the Covariance Matrix

Tables 2, 3 and Figure 2 present the results of the simulation with the covariance matrix. To test $k = 0$ in Table 2, the factors are sorted from best to worst according to the quality of the approximation: $FC5_{Ferre}$, $FC2_{Bartlett}$, $FC3_{Lawley}$, W_{n-1}, and W_n. Additionally, factors $FC5_{Ferre}$ and $FC2_{Bartlett}$

present a good approximation even with sample sizes close to the number of variables. However, when $p = 30$, factor $FC2_{Bartlett}$ requires $n = 100$, while $n = 50$ is enough for $FC5_{Ferre}$. In turn, factors W_{n-1} and W_n exhibit a poor approximation, which quickly worsens as the number of variables increases.

Based on Table 3, to test $k = 2$ the order of the factors is $FC5_{Ferre}$, $FC2_{Bartlett}$, $FC3_{Lawley}$, $FC4_{Jackson}$, W_{n-1}, and W_n. Factors $FC2_{Bartlett}$ and $FC3_{Lawley}$ exhibit a very similar behavior, although $FC2_{Bartlett}$ still presents significance levels slightly lower than those of $FC3_{Lawley}$. Factors W_{n-1} and W_n still show the worst approximation to their limit distribution, as can be seen in Figure 2(d). When $k = 3$ the order is $FC5_{Ferre}$, $FC3_{Lawley}$, $FC2_{Bartlett}$, $FC4_{Jackson}$, W_{n-1}, and W_n. Factors $FC2_{Bartlett}$ and $FC3_{Lawley}$ still present a similar behavior; however, in this case, factor $FC3_{Lawley}$ produces slightly lower significance levels. Factors $FC4_{Jackson}$, W_{n-1}, and W_n still show the worst approximation.

In order to get a closer view of the performance of the correction factors with respect to the sample size, we estimate the significance levels for each change in the sample size, in units, from $n = p + 1$ to $n = 1000$. Figure 2 presents the behaviors we obtained. All the charts enable us to conclude that, with sample sizes very close to the number of variables, most statistics exhibit a poor behavior, except for $FC5_{Ferre}$, which presented an acceptable behavior with a low number of variables ($p = 5,\ 10$). Furthermore, as p increases, the performance of all the statistics is reduced, which results in the need for larger sample sizes to obtain good approximations. Factors W_n, W_{n-1}, and $FC4_{Jackson}$ are most affected by such increase, while $FC5_{Ferre}$ continues presenting the lowest estimations.

4.2. Simulation Results in the Case of the Correlation Matrix

If the PCA is carried out using the correlation matrix and $k = 0$ is tested, Table 1 shows that the correction factor with the best approximation, even with small sample sizes, is $FC1_{LR}$. That factor, for example, for $p = 10$, requires a sample size of $n = 30$ to obtain an estimated significance level close to the nominal one. $FC2_{LR}$ presented the second best performance, although it requires sample sizes that are sometimes much larger than those of $FC1_{LR}$ to achieve a good approximation. Finally, the L_R statistic, simply multiplied by n, presents a poor approximation.

Tables 1 and 4 present the results of the simulation with the correlation matrix. The quantity $n - \frac{2p+11}{6}$ of factor $FC1_{LR}$ is less than $n - \frac{2p+5}{6}$ of factor $FC2_{LR}$. This enables factor $FC1_{LR}$ to produce more contraction of the basic form of the statistic, that is, of $\log|R|$. As a consequence, the sampling distribution of $FC2_{LR}$ will be more displaced to the left and, therefore, it will exhibit lower estimated significance levels.

TABLE 1: Comparison of the significance levels estimated based on the statistics of the correlation matrix; L_R, $FC1_{LR}$ and $FC2_{LR}$; nominal significance level $\alpha = 0.05$ and $k = 0$.

p		$n = 10$	$n = 30$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
5	L_R	0.3129	0.0939	0.0737	0.0614	0.0534	0.0492
	$FC1_{LR}$	0.0622	0.0521	0.0496	0.0501	0.0484	0.0467
	$FC2_{LR}$	0.1187	0.0624	0.0560	0.0534	0.0497	0.0476
10	L_R		0.2637	0.1403	0.0886	0.0650	0.0583
	$FC1_{LR}$		0.0512	0.0532	0.0521	0.0492	0.0529
	$FC2_{LR}$		0.0776	0.0653	0.0586	0.0525	0.0536
15	L_R		0.6737	0.3158	0.1380	0.0840	0.0617
	$FC1_{LR}$		0.0773	0.0584	0.0504	0.0518	0.0493
	$FC2_{LR}$		0.1340	0.0799	0.0602	0.0560	0.0506
30	L_R			0.9971	0.6420	0.2340	0.1002
	$FC1_{LR}$			0.1493	0.0640	0.0506	0.0494
	$FC2_{LR}$			0.2497	0.0856	0.0607	0.0521

TABLE 2: Comparison of the significance levels estimated based on the statistics of the covariance matrix; W_n, W_{n-1}, $FC2_{Bartlett}$, $FC3_{Lawley}$, and $FC5_{Ferre}$; nominal significance level $\alpha = 0.05$ and $k = 0$.

p		$n = 10$	$n = 30$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
5	W_n	0.2947	0.0931	0.0733	0.0604	0.0549	0.0523
	W_{n-1}	0.2021	0.0771	0.0649	0.0567	0.0532	0.0515
	$FC2_{Bartlett}$	0.0362	0.0437	0.0467	0.0481	0.0491	0.0498
	$FC3_{Lawley}$	0.1287	0.0634	0.0580	0.0534	0.0518	0.0509
	$FC5_{Ferre}$	0.0098	0.0333	0.0398	0.0447	0.0476	0.0492
10	W_n		0.2471	0.1382	0.0843	0.0652	0.0561
	W_{n-1}		0.1928	0.1147	0.0761	0.0617	0.0546
	$FC2_{Bartlett}$		0.0413	0.0438	0.0474	0.0484	0.0491
	$FC3_{Lawley}$		0.0848	0.0667	0.0575	0.0535	0.0512
	$FC5_{Ferre}$		0.0241	0.0325	0.0416	0.0453	0.0477
15	W_n		0.6479	0.3027	0.1342	0.0834	0.0619
	W_{n-1}		0.5502	0.2491	0.1172	0.0775	0.0599
	$FC2_{Bartlett}$		0.0515	0.0442	0.0466	0.0479	0.0492
	$FC3_{Lawley}$		0.1417	0.0816	0.0617	0.0553	0.0522
	$FC5_{Ferre}$		0.0239	0.0293	0.0383	0.0436	0.0474
30	W_n			0.9967	0.6161	0.2366	0.0990
	W_{n-1}			0.9926	0.5573	0.2131	0.0944
	$FC2_{Bartlett}$			0.1058	0.0495	0.0487	0.0489
	$FC3_{Lawley}$			0.2751	0.0870	0.0638	0.0545
	$FC5_{Ferre}$			0.0496	0.0344	0.0410	0.0457

In a test to reduce p variables to $k = 2$ components based on the correlation matrix (see Table 4), correction factor B, that is, W_B^*, produces the best approximation with estimated significance levels, in most cases, much lower than those obtained with other statistics or correction factors. However, the increase in the number of variables has a negative effect on the performance of such factor, which generates situations that require sample sizes above 500 for 30 variables. Nevertheless, such factor offers the best approximation. Factors W_n^* and W_{n-1}^* present the same behavior: poor performance. Statistics $\chi^2_{\mu_{W^*}}$ and $c\chi^2_d$ exhibited a similar behavior, although the latter with slightly lower significance levels and a better behavior than factor W_B^* in cases of large sample sizes. Statistics $\chi^2_{\mu_{W^*}}$ and $c\chi^2_d$ did not exhibit a good performance with small sample sizes.

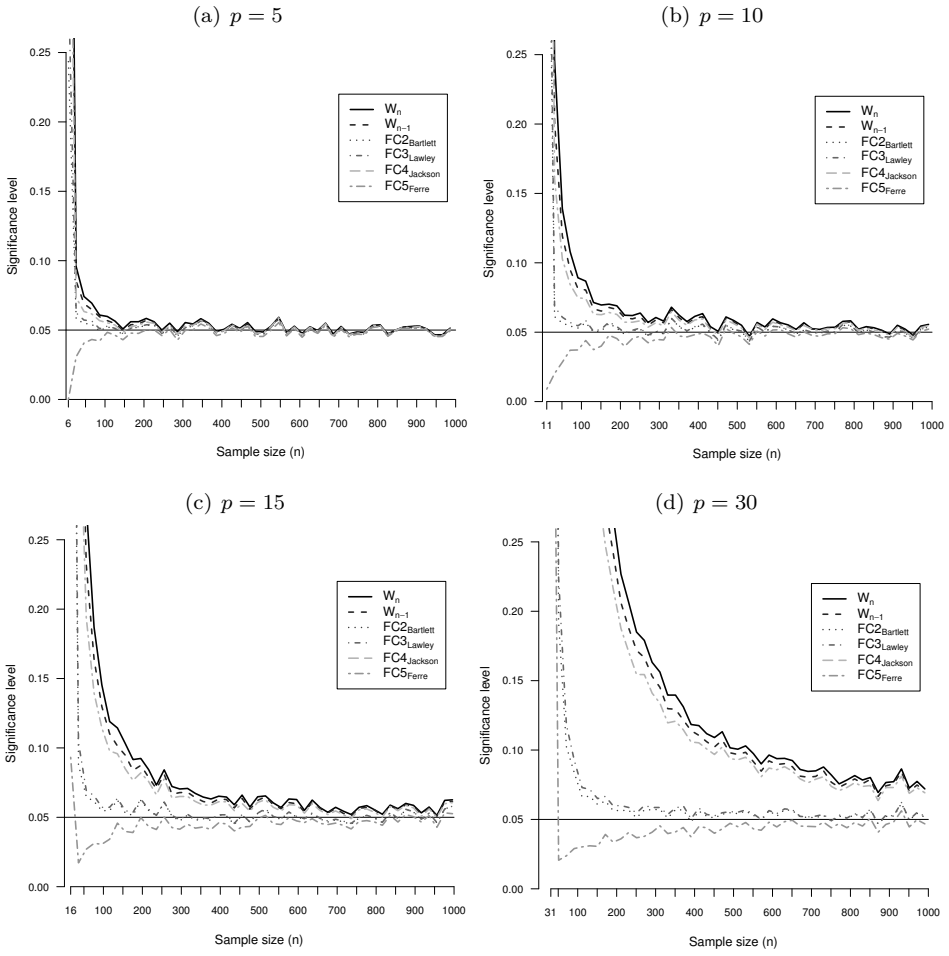

FIGURE 2: Detailed comparison by sample size n of correction factors for the covariance matrix: W_n, W_{n-1}, $FC2_{Bartlett}$, $FC3_{Lawley}$, $FC4_{Jackson}$, and $FC5_{Ferre}$, with $\alpha = 0.05$ y $k = 2$.

TABLE 3: Comparison of the significance levels estimated based on the statistics of the covariance matrix; W_n, W_{n-1}, $FC2_{Bartlett}$, $FC3_{Lawley}$, $FC4_{Jackson}$ and $FC5_{Ferre}$; nominal significance level $\alpha = 0.05$; $k = 2$ and $k = 3$.

		$k=2$						$k=3$					
p		$n=10$	$n=30$	$n=50$	$n=100$	$n=200$	$n=500$	$n=10$	$n=30$	$n=50$	$n=100$	$n=200$	$n=500$
5	W_n	0.2616	0.0879	0.0716	0.0602	0.0549	0.0514	0.2175	0.0826	0.0672	0.0580	0.0531	0.0505
	W_{n-1}	0.2059	0.0773	0.0664	0.0576	0.0537	0.0510	0.1834	0.0756	0.0634	0.0561	0.0523	0.0502
	$FC2_{Bartlett}$	0.0784	0.0540	0.0533	0.0516	0.0509	0.0501	0.0958	0.0588	0.0548	0.0521	0.0502	0.0495
	$FC3_{Lawley}$	0.0948	0.0576	0.0553	0.0525	0.0513	0.0503	0.0815	0.0564	0.0533	0.0513	0.0499	0.0493
	$FC4_{Jackson}$	0.1521	0.0673	0.0609	0.0551	0.0527	0.0507	0.1127	0.0621	0.0564	0.0529	0.0507	0.0496
	$FC5_{Ferre}$	0.0028	0.0314	0.0396	0.0451	0.0474	0.0490	0.0021	0.0350	0.0414	0.0457	0.0472	0.0482
10	W_n		0.2666	0.1425	0.0858	0.0666	0.0551		0.2645	0.1441	0.0868	0.0659	0.0555
	W_{n-1}		0.2206	0.1225	0.0786	0.0636	0.0541		0.2233	0.1267	0.0806	0.0630	0.0547
	$FC2_{Bartlett}$		0.0712	0.0585	0.0530	0.0522	0.0498		0.0838	0.0661	0.0580	0.0525	0.0510
	$FC3_{Lawley}$		0.0792	0.0625	0.0547	0.0529	0.0501		0.0744	0.0617	0.0562	0.0517	0.0507
	$FC4_{Jackson}$		0.1775	0.1048	0.0721	0.0607	0.0531		0.1486	0.0948	0.0689	0.0578	0.0532
	$FC5_{Ferre}$		0.0196	0.0295	0.0390	0.0449	0.0467		0.0175	0.0276	0.0398	0.0437	0.0476
15	W_n		0.6865	0.3257	0.1413	0.0859	0.0616		0.6927	0.3301	0.1419	0.0858	0.0618
	W_{n-1}		0.6040	0.2779	0.1261	0.0805	0.0601		0.6180	0.2849	0.1275	0.0808	0.0603
	$FC2_{Bartlett}$		0.1043	0.0692	0.0560	0.0528	0.0502		0.1369	0.0817	0.0624	0.0551	0.0515
	$FC3_{Lawley}$		0.1224	0.0769	0.0594	0.0542	0.0508		0.1158	0.0740	0.0591	0.0537	0.0510
	$FC4_{Jackson}$		0.5136	0.2330	0.1116	0.0754	0.0583		0.4461	0.1994	0.1022	0.0711	0.0573
	$FC5_{Ferre}$		0.0122	0.0228	0.0335	0.0415	0.0454		0.0088	0.0204	0.0328	0.0409	0.0462
30	W_n			0.9979	0.6436	0.2497	0.1013			0.9984	0.6562	0.2519	0.1033
	W_{n-1}			0.9950	0.5893	0.2266	0.0962			0.9959	0.6053	0.2301	0.0983
	$FC2_{Bartlett}$			0.2069	0.0746	0.0577	0.0521			0.2652	0.0911	0.0628	0.0545
	$FC3_{Lawley}$			0.2411	0.0829	0.0609	0.0532			0.2264	0.0821	0.0598	0.0534
	$FC4_{Jackson}$			0.9900	0.5337	0.2045	0.0916			0.9834	0.4943	0.1888	0.0891
	$FC5_{Ferre}$			0.0237	0.0264	0.0355	0.0432			0.0159	0.0228	0.0334	0.0426

TABLE 4: Comparison of the significance levels estimated based on the statistics of the correlation matrix; W_n^*, W_{n-1}^*, W_B^*, $\chi^2_{\mu_{W^*}}$, and $c\chi^2_d$; nominal significance level $\alpha = 0.05$; $k = 2$ and $k = 3$.

		$k=2$						$k=3$					
p		$n=10$	$n=30$	$n=50$	$n=100$	$n=200$	$n=500$	$n=10$	$n=30$	$n=50$	$n=100$	$n=200$	$n=500$
5	W_n^*	0.2835	0.0975	0.0810	0.0600	0.0600	0.0645	0.2570	0.0980	0.0895	0.0760	0.0825	0.0760
	W_{n-1}^*	0.2175	0.0845	0.0775	0.0575	0.0590	0.0635	0.2235	0.0940	0.0835	0.0735	0.0815	0.0755
	W_B^*	0.0670	0.0560	0.0625	0.0520	0.0550	0.0625	0.0795	0.0670	0.0720	0.0645	0.0765	0.0745
	$\chi^2_{\mu_{W^*}}$	0.2725	0.0975	0.0810	0.0600	0.0600	0.0645	0.2415	0.0980	0.0895	0.0760	0.0825	0.0760
	$c\chi^2_d$	0.2425	0.0900	0.0765	0.0510	0.0530	0.0505	0.1955	0.0780	0.0615	0.0480	0.0495	0.0535
10	W_n^*		0.2955	0.1920	0.1005	0.0825	0.0735		0.3885	0.2385	0.1775	0.1285	0.1280
	W_{n-1}^*		0.2465	0.1670	0.0955	0.0795	0.0730		0.3435	0.2185	0.1645	0.1230	0.1275
	W_B^*		0.0765	0.0810	0.0680	0.0695	0.0670		0.1245	0.1145	0.1190	0.1025	0.1175
	$\chi^2_{\mu_{W^*}}$		0.2585	0.1585	0.0870	0.0720	0.0605		0.2810	0.1575	0.0990	0.0695	0.0695
	$c\chi^2_d$		0.2475	0.1445	0.0815	0.0620	0.0555		0.2415	0.1440	0.0900	0.0655	0.0645
15	W_n^*		0.7215	0.3920	0.1745	0.1165	0.0960		0.7785	0.4805	0.2500	0.1835	0.1430
	W_{n-1}^*		0.6445	0.3320	0.1610	0.1095	0.0940		0.7190	0.4375	0.2320	0.1750	0.1390
	W_B^*		0.1135	0.0860	0.0785	0.0760	0.0770		0.1640	0.1315	0.1195	0.1255	0.1195
	$\chi^2_{\mu_{W^*}}$		0.6690	0.3235	0.1420	0.0950	0.0690		0.6650	0.3490	0.1555	0.1025	0.0740
	$c\chi^2_d$		0.6525	0.3045	0.1320	0.0865	0.0585		0.6310	0.3010	0.1250	0.0810	0.0550
30	W_n^*			0.9985	0.6980	0.3005	0.1510			0.9995	0.7875	0.4190	0.2235
	W_{n-1}^*			0.9965	0.6505	0.2835	0.1450			0.9990	0.7490	0.4020	0.2160
	W_B^*			0.2245	0.1050	0.0925	0.0880			0.2585	0.1515	0.1515	0.1320
	$\chi^2_{\mu_{W^*}}$			0.9970	0.6295	0.2450	0.1150			0.9980	0.6405	0.2610	0.1140
	$c\chi^2_d$			0.9965	0.6100	0.2330	0.1075			0.9955	0.5885	0.2335	0.1005

TABLE 5: Power comparison for statistics based on the covariance matrix; for H_{00} and H_{02} testing when really $\lambda_1 = (1+\delta)$ and $\lambda_3 = (1+\delta)$ respectively, with $\alpha = 0.05$

	$1+\delta$		$n=30$	$n=50$	$n=100$	$n=200$	$n=500$
$p=10, k=0$	1.5	$FC2_{Bartlett}$	0.0650	0.0923	0.1796	0.4074	0.9095
		$FC3_{Lawley}$	0.1238	0.1303	0.2053	0.4244	0.9118
		$FC5_{Ferre}$	0.0409	0.0731	0.1646	0.3973	0.9079
	2.0	$FC2_{Bartlett}$	0.1418	0.2725	0.6298	0.9611	1.0000
		$FC3_{Lawley}$	0.2316	0.3356	0.6608	0.9643	1.0000
		$FC5_{Ferre}$	0.0973	0.2349	0.6101	0.9591	1.0000
	2.5	$FC2_{Bartlett}$	0.2753	0.5441	0.9282	0.9997	1.0000
		$FC3_{Lawley}$	0.3932	0.6110	0.9380	0.9997	1.0000
		$FC5_{Ferre}$	0.2088	0.4992	0.9221	0.9997	1.0000
$p=30, k=0$	1.5	$FC2_{Bartlett}$		0.1287	0.0834	0.1353	0.4017
		$FC3_{Lawley}$		0.3110	0.1381	0.1668	0.4217
		$FC5_{Ferre}$		0.0619	0.0596	0.1178	0.3900
	2.0	$FC2_{Bartlett}$		0.1860	0.2141	0.5202	0.9872
		$FC3_{Lawley}$		0.4022	0.3037	0.5697	0.9886
		$FC5_{Ferre}$		0.0977	0.1661	0.4897	0.9862
	2.5	$FC2_{Bartlett}$		0.2768	0.4541	0.9004	1.0000
		$FC3_{Lawley}$		0.5143	0.5615	0.9195	1.0000
		$FC5_{Ferre}$		0.1620	0.3901	0.8865	1.0000
$p=10, k=2$	1.5	$FC2_{Bartlett}$	0.1075	0.1301	0.2335	0.4916	0.9485
		$FC3_{Lawley}$	0.1186	0.1364	0.2377	0.4941	0.9487
		$FC5_{Ferre}$	0.0335	0.0758	0.1931	0.4663	0.9460
	2.0	$FC2_{Bartlett}$	0.2170	0.3614	0.7202	0.9805	1.0000
		$FC3_{Lawley}$	0.2336	0.3727	0.7244	0.9808	1.0000
		$FC5_{Ferre}$	0.0879	0.2585	0.6763	0.9774	1.0000
	2.5	$FC2_{Bartlett}$	0.3782	0.6409	0.9579	1.0000	1.0000
		$FC3_{Lawley}$	0.3990	0.6505	0.9589	1.0000	1.0000
		$FC5_{Ferre}$	0.1983	0.5350	0.9463	0.9999	1.0000
$p=30, k=2$	1.5	$FC2_{Bartlett}$		0.2439	0.1258	0.1678	0.4515
		$FC3_{Lawley}$		0.2831	0.1367	0.1741	0.4553
		$FC5_{Ferre}$		0.0304	0.0501	0.1149	0.4177
	2.0	$FC2_{Bartlett}$		0.3277	0.2893	0.5897	0.9926
		$FC3_{Lawley}$		0.3712	0.3065	0.5989	0.9928
		$FC5_{Ferre}$		0.0539	0.1464	0.5034	0.9906
	2.5	$FC2_{Bartlett}$		0.4430	0.5580	0.9318	1.0000
		$FC3_{Lawley}$		0.4890	0.5775	0.9348	1.0000
		$FC5_{Ferre}$		0.0993	0.3718	0.9012	1.0000

TABLE 6: Power comparison for statistics based on the correlation matrix; for H_{00} and H_{02} testing when really $\lambda_1 = (1+\delta)$ and $\lambda_3 = (1+\delta)$ respectively, with $\alpha = 0.05$.

	$1+\delta$		$n=30$	$n=50$	$n=100$	$n=200$	$n=500$
$p=10, k=0$	1.5	$FC1_{LR}$	0.1304	0.2080	0.4474	0.8360	0.9998
		$FC2_{LR}$	0.1786	0.2402	0.4660	0.8416	0.9998
	2.0	$FC1_{LR}$	0.3354	0.6126	0.9532	0.9998	1.0000
		$FC2_{LR}$	0.4036	0.6468	0.9566	0.9998	1.0000
	2.5	$FC1_{LR}$	0.6224	0.8996	0.9992	1.0000	1.0000
		$FC2_{LR}$	0.6834	0.9136	0.9992	1.0000	1.0000
$p=30, k=0$	1.5	$FC1_{LR}$		0.1858	0.1204	0.1952	0.5764
		$FC2_{LR}$		0.3042	0.1544	0.2198	0.5886
	2.0	$FC1_{LR}$		0.2806	0.3152	0.6872	0.9984
		$FC2_{LR}$		0.4186	0.3710	0.7126	0.9986
	2.5	$FC1_{LR}$		0.4264	0.6264	0.9702	1.0000
		$FC2_{LR}$		0.5654	0.6822	0.9738	1.0000
$p=10, k=2$	1.5	W_B^*	0.1203	0.1477	0.2638	0.5266	0.9486
		$\chi^2_{\mu_{W^*}}$	0.3471	0.2613	0.3111	0.5336	0.9431
		$c\chi^2_d$	0.3341	0.2496	0.3039	0.5213	0.9336
	2.0	W_B^*	0.2233	0.3815	0.7381	0.9845	1.0000
		$\chi^2_{\mu_{W^*}}$	0.4994	0.5354	0.7782	0.9851	1.0000
		$c\chi^2_d$	0.4798	0.5258	0.7655	0.9850	1.0000
	2.5	W_B^*	0.3945	0.6602	0.9671	0.9999	1.0000
		$\chi^2_{\mu_{W^*}}$	0.6796	0.7812	0.9753	0.9999	1.0000
		$c\chi^2_d$	0.6658	0.7710	0.9754	0.9999	1.0000
$p=30, k=2$	1.5	W_B^*		0.2517	0.1570	0.2079	0.5126
		$\chi^2_{\mu_{W^*}}$		0.9980	0.7265	0.4522	0.5790
		$c\chi^2_d$		0.9976	0.7107	0.4325	0.5588
	2.0	W_B^*		0.3310	0.3138	0.6361	0.9920
		$\chi^2_{\mu_{W^*}}$		0.9989	0.8722	0.8450	0.9942
		$c\chi^2_d$		0.9986	0.8614	0.8335	0.9937
	2.5	W_B^*		0.4480	0.5862	0.9400	1.0000
		$\chi^2_{\mu_{W^*}}$		0.9994	0.9630	0.9868	1.0000
		$c\chi^2_d$		0.9992	0.9598	0.9850	1.0000

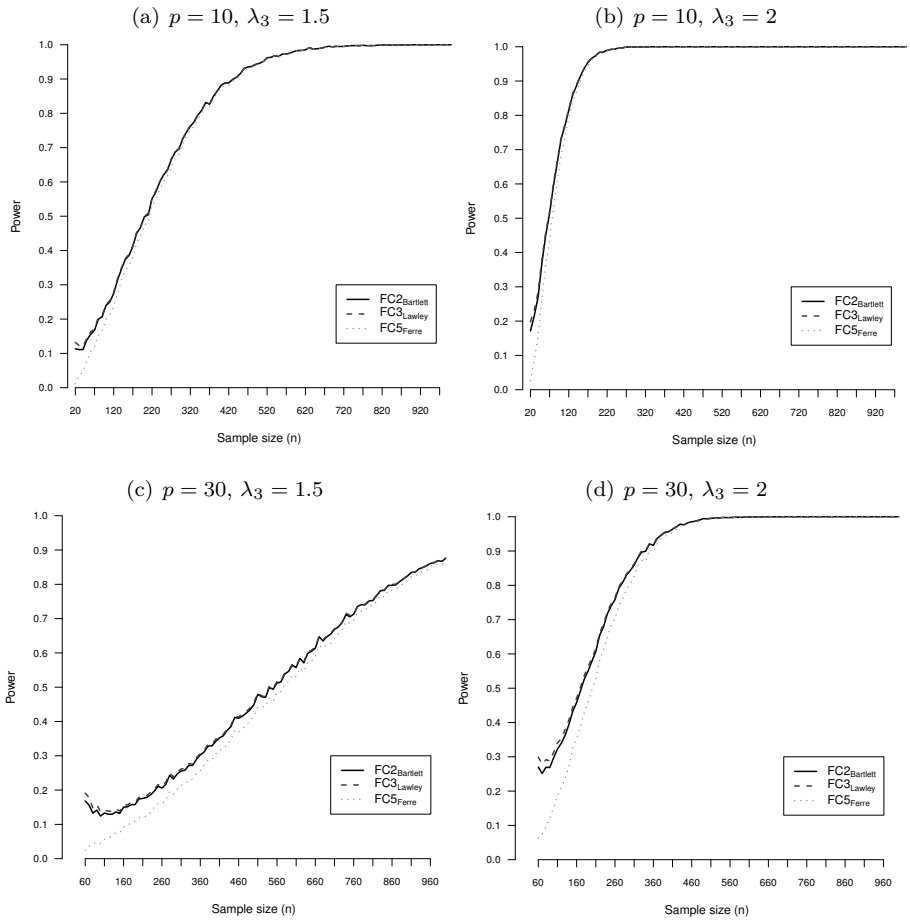

FIGURE 3: Detailed power performance by sample size of the covariance matrix correction factors for H_{02} testing when $\lambda_3 = (1 + \delta)$ with $\alpha = 0.05$.

With the increase in components to $k = 3$ it is observed a deterioration of the approximation of factors W_n^*, W_{n-1}^*, and W_B^*. In turn, statistics $\chi^2_{\mu_{W^*}}$ and $c\chi^2_d$ still present estimated significance levels similar to those obtained with $k = 2$.

4.3. Simulation Results for the Power of the Test

Regarded to the power of the test, for both scenarios k=0 and k=2, when the covariance matrix is used, the factor $FC3_{Lawley}$ presents the best performance, followed by $FC2_{Bartlett}$, and with the lowest performance we have $FC5_{Ferre}$ (Table 5). This means that $FC3_{Lawley}$ generates the lowest probability of Type II Error in all scenarios. As expected, the power of the test for the different factors increases with the sample size, although the growth rate is subject to the level of deviation of the eigenvalue from the common value λ (Figure 3). In general terms

the three factors show a better performance in the k=2 scenario than in the k=0 scenario. For cases with a low differentiation ($\delta = 0.5$) and $p = 10$, sample sizes of 500 are needed to obtain powers close to 1. Although when $p = 30$, in the same sample size of $n = 500$, probabilities of Type II Error greater than 50% are obtained and only for sample sizes greater than 1000, powers greater than 90% are observed (see Figure 3).

In order to evaluate the relevance of the first component using the correlation matrix, the best result is obtained with $FC2_{LR}$ (Table 6). This performance difference from $FC1_{LR}$ is most noticeable when $p = 30$. To evaluate $k = 2$, the best result is obtained with $\chi^2_{\mu_{W^*}}$ followed by $c\chi^2_d$ with a close performance. However, both factors show less consistent behavior when n is close to p. Unlike W^*_B which shows a more consistent growth of power with increasing sample size. On the other hand, similar to the covariance matrix, estimated powers of around 0.51 and 0.57 are observed when there is a low differentiation of the component, with the number of variables of $p = 30$ and a sample size of $n = 500$.

It can also be seen that in general the powers obtained with the correlation matrix are greater than those obtained with the covariance matrix under the same scenarios. For example, for $k = 0$, $p = 10$, $1 + \delta = 1.5$ and $n = 500$ we get power estimates of 0.99, while in the covariance matrix the values are between 0.90 and 0.91. Similarly, for $k = 2$ and $p = 30$, $1+\delta = 1.5$ and $n = 500$, the estimated power ranges from 0.51 to 0.57 for the three factors in the correlation matrix, while in the covariance matrix, the values are between 0.41 and 0.45.

5. Conclusions

Assuming a normal distribution, we compared the different correction factors that have been proposed for likelihood-ratio statistic to define the number of components that should be retained in PCA. Using the test of the hypothesis of equality of the smallest last $(p - k)$ eigenvalues of the covariance or correlation matrix. For large sample sizes in the order of $n = 500$, most factors generated estimated significance levels close to their nominal counterparts and even below them, which indicates that a good approximation was achieved.

In order to conduct a PCA based on the covariance matrix, factors $FC5_{Ferre}$ and $FC2_{Bartlett}$ present the best approximation with $k = 0, 2, 3$, even with sample sizes close to the number of variables p. With $k = 2$, factors $FC2_{Bartlett}$ and $FC3_{Lawley}$ exhibit a very similar behavior, although $FC2_{Bartlett}$ requires larger sample sizes to obtain good approximations. With $k = 3$, factors $FC2_{Bartlett}$ and $FC3_{Lawley}$ still present a similar behavior, although $FC2_{Bartlett}$ exhibits slightly higher levels than $FC3_{Lawley}$. Furthermore, factors $FC4_{Jackson}$, W_{n-1}, and W_n offer the worst performance, which is even more critical when there is a big difference between the number of variables and the sample size. This produces situations in which sample sizes above 800 are required for 30 variables (see Figure 2(d)). Hence, these factors would not be recommended to determine the number of components. Finally, we can observe that, as the number k of components grows,

the estimated significance levels of factors $FC2_{Bartlett}$ and $FC3_{Lawley}$ exhibit patterns of slight increase and decrease, respectively.

In this work we also investigated the power of the test in relation to the use of the different factors. To do this, we focused on the factors that performed best at the significance level. As it was mentioned above, the factors $FC2_{Bartlett}$, $FC3_{Lawley}$ and $FC5_{Ferre}$ generated the best approximation when using the covariance matrix. Among these factors, the best power for the different scenarios was obtained with the factor $FC3_{Lawley}$, followed by the factor $FC2_{Bartlett}$ and finally $FC5_{Ferre}$ with the lowest performance. As it can be seen, the single recommendation of a factor that provides the lowest probability of Type I Error and in turn the highest power is a difficult problem. However, given the scenarios that have been studied, we dare to consider the factor $FC2_{Bartlett}$ as the most balanced between these two objectives. This considering that although it does not generate the least level of significance or the greatest power, it was close to the factors with the best performance in each case. The simulation results also showed that if the component differentiation is low ($\delta = 0.5$), sample sizes of 30 to 50 times the number of variables are required to obtain powers greater than 90%. If the deviation is higher ($\delta = 1.5$), only 6 or 10 larger sample sizes are required.

If the correlation matrix is used with $k = 0$, factor $FC1_{LR}$ presents the lowest estimated significance levels, close to the nominal value, even with small sample sizes. Moreover, this factor is the most consistent as the number of variables increases, as opposite to the L_R statistic, which becomes more erroneous. Now, in terms of power, the factor $FC2_{LR}$ presents the best performance in the different scenarios studied. Knapp & Swoyer (1967) using this same factor pointed out the sensitivity of the test in identifying the first component. This is consistent with the current study, where we observed a high power of 0.9136 in a case where the first component represents a global correlation between all variables of only 0.17 ($\lambda_1 = 2.5$). This means that the test is highly powerful even in scenarios with low global correlation.

Regarding the test with $k = 2$ and $k = 3$, Table 4 shows that W_B^* is the factor with the best approximation and consistency. Nevertheless, factor $c\chi_d^2$, with large sample sizes, presents even better results than W_B^*. In addition, as the number of components grows, most statistics are deteriorated. With $k = 2$, the statistics $\chi^2_{\mu_{W^*}}$ and $c\chi_d^2$ are similar; however, with $k = 3$, a greater difference can be observed, and the $c\chi_d^2$ statistic presents a better approximation. This is in line with Schott (1988) regarding the superior performance of $c\chi_d^2$ with respect to $\chi^2_{\mu_{W^*}}$. With respect to power, $\chi^2_{\mu_{W^*}}$ shows the best results followed by $c\chi_d^2$. The factor W_B^* only begins to show approximately comparable results after sample sizes between 10 and 20 times larger than the number of variables. Thus, if large sample sizes are available, it is recommended to use the factor W_B^* for $k > 0$. And in the case of testing $k = 0$ it is recommended to use the factor $FC2_{LR}$

Finally, as p increases, the behavior of all the statistics worsens. Therefore, it would be interesting to precisely study the n/p ratio under which the sampling distribution of the statistic would exhibit a good approximation in general. We can empirically establish that, to achieve adequate approximations, we require sample

sizes 10 times the number of variables in the case of the covariance matrix, and 20 times if a correlation matrix is used. Which are more demanding sizes than the minimum of $2p$ indicated by Schott (2012) for covariance matrix using the factors $FC2_{Bartlett}$ and $FC3_{Lawley}$. Although the values proposed in this work are in consideration of the significance level of all factors, the sample requirements can be increased up to more than 30 times if a high power is desired.

As indicated at the beginning of this study, Table 7 provides a series of recommendations about correction factors that should be used according to a specific configuration, p number of variables, n sample size, and k number of main components being tested. The objective is to obtain the best performance in the test. Furthermore, the results are discriminated depending on the type of matrix (covariance or correlation) on which the PCA was based.

TABLE 7: Recommended correction factors according to a specific configuration of the p number of variables, n sample size, and k number of components considering the type of matrix (covariance or correlation) on which the PCA was based.

Association Matrix	k	p	n	Recommended Factor
Covariance	$k=0$	5-9	10-50	$FC5_{Ferre}$, $FC2_{Bartlett}$
			50-200	$FC2_{Bartlett}$, $FC5_{Ferre}$
			200+	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
		10-29	30-50	$FC2_{Bartlett}$, $FC5_{Ferre}$
			50-200	$FC2_{Bartlett}$, $FC5_{Ferre}$
			200+	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
		30+	50-200	$FC5_{Ferre}$
			200+	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
	$k>0$	5-9	10-50	$FC2_{Bartlett}$, $FC5_{Ferre}$
			50-200	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
			200+	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
		10-29	30-50	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
			50-200	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
			200+	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
		30+	50-200	$FC5_{Ferre}$
			200+	$FC2_{Bartlett}$, $FC5_{Ferre}$, $FC3_{Lawley}$
Correlation	$k=0$	5-9	10-50	$FC1_{LR}$
			50-200	$FC2_{LR}$, $FC1_{LR}$
			200+	$FC2_{LR}$, $FC1_{LR}$, L_R
		10-29	30-50	$FC1_{LR}$
			50-200	$FC2_{LR}$, $FC1_{LR}$
			200+	$FC2_{LR}$, $FC1_{LR}$
		30+	50-200	$FC1_{LR}$
			200+	$FC2_{LR}$, $FC1_{LR}$
	$k>0$	5-9	10-50	W_B^*
			50-200	W_B^*, $c\chi_d^2$
			200+	W_B^*, $c\chi_d^2$
		10-29	30-50	W_B^*
			50-200	W_B^*
			200+	W_B^*, $c\chi_d^2$, $\chi^2_{\mu_{W^*}}$
		30+	50-200	W_B^*
			200+	W_B^*, $c\chi_d^2$

[Received: December 2019 — Accepted: September 2020]

References

Anderson, T. (1963), 'Asymptotic theory for principal component analysis', *The Annals of Mathematical Statistics* **34**(1), 122–148.

Arteaga, F. & Ferrer, A. (2010), 'How to simulate normal data sets with the desired correlation structure', *Chemometrics and Intelligent Laboratory Systems* **101**, 38–42.

Bartlett, M. (1951), 'The effect of standardization on a χ^2 approximation in factor analysis', *Biometrika* **38**(3/4), 337–344.

Bartlett, M. (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.

Björklund, M. (2019), 'Be careful with your principal components', *Evolution* **73**(10), 2151–2158.

Box, G. E. P. (1949), 'A general distribution theory for a class of likelihood criteria', *Biometrika* **36**(3/4), 317–346.

Chakraborty, L., Rus, H., Henstra, D., Thistlethwaite, J. & Scott, D. (2020), 'A place-based socioeconomic status index: Measuring social vulnerability to flood hazards in the context of environmental justice', *International Journal of Disaster Risk Reduction* **43**.

Ferré, L. (1995), 'Selection of components in principal component analysis: a comparison of methods', *Computational Statistics & Data Analysis* **19**, 669–689.

Friedman, S. (1981), 'Interpreting the first eigenvalue of a correlation matrix', *Educational and Psychological Measurement* **41**, 11–21.

Fujikoshi, Y., Yamada, T., Watanabe, D. & Sugiyama, T. (2007), 'Asymptotic distribution of the LR statistic for equality of the smallest eigenvalues in high-dimensional principal component analyisis', *Journal of Multivariate Analysis* **98**, 2002–2008.

Jackson, D. (1993), 'Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches', *Ecological Society of America* **74**(8), 2204–2214.

Jackson, J. E. (1991), *A User's Guide To Principal Components*, John Wiley & Sons, Inc.

Jolliffe, I. (2002), *Principal Component Analysis*, 2 edn, Springer.

Knapp, T. R. & Swoyer, V. H. (1967), 'Some empirical results concerning the power of Bartlett's Test of the significance of a correlation matrix', *American Educational Research Association* **4**(1), 13–17.

Krazanowski, W. J. (1988), *Principles of Multivariate Analysis, A User's Perspective*, Oxford Statistical Science.

Lawley, D. (1956), 'Test of significance for latent roots of covariance and correlations', *Biometrika* **43**(1/2), 128–136.

Mardia, K., Kent, J. & Bibby, J. (1979), *Multivariate Analysis*, 6 edn, Academic Press, San Diego.

Maté, C. G. (2011), 'A multivariate analysis approach to forecasts combination. application to foreign exchange (FX) markets', *Revista Colombiana de Estadistica* **34**(2), 347–375.

Peres-Neto, P. R., Jackson, D. A. & Somers, K. M. (2005), 'How many principal components? stopping rules for determining the number of non-trivial axes revisited', *Computational Statistics and Data Analysis* **49**(4), 974–997.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Şahan, C., Baydur, H. & Demiral, Y. (2018), 'A novel version of copenhagen psychosocial questionnaire-3: Turkish validation study', *Archives of Environmental & Occupational Health* **74**(6), 297–309.

Schott, J. R. (1988), 'Testing the equality of the smallest latent roots of a correlation matrix', *Biometrika* **75**(4), 794–796.

Schott, J. R. (2006), 'A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix', *Journal of Multivariate Analysis* **97**, 827–843.

Schott, J. R. (2012), 'An Approximation for the Test of the Equality of the Smallest Eigenvalues of a Covariance Matrix', *Communications in Statistics-Theory and Methods* **41**, 4439–4443.

Watanabe, D., Okada, S., Fujikoshi, Y. & Sugiyama, T. (2008), 'Large sample approximations for LR statistic for equality of the smallest eigenvalues of a covariance matrix under elliptical population', *Computational Statistics & Data Analysis* **52**, 2714–2724.

Waternaux, C. (1984), 'Principal components in the nonnormal case: the test of equality of Q roots', *Journal of Multivariate Analysis* **14**, 323–335.