A PROPOSAL FOR THE ANALYSIS OF MAIN COMPONENTS IN THE PRESENCE OF NON-RANDOM VARIABLES

Juliana Vieira GOMES[1]
Camila Rafaela Gomes DIAS[1]
José Ivo RIBEIRO JÚNIOR[1]

- ABSTRACT: For exploratory analysis of the principal components (CPs), the assumption of multivariate normality of the variables is not required, nor necessarily that they are random. This means that variables that do not behave randomly can also be included in this analysis. Thus, in order to carry out the analysis of the PCs with random variables or not, a correction of the matrix based on the coefficients of variation was proposed (Campana *et al.*, 2010) by applying the method of Lenth (1989), whose new array was named . To verify its feasibility, ten data sets of random variables Y_1, Y_2, Y_3 and Y_4 were simulated, with 10,000 values each and that followed multivariate normal distribution. After the simulation, 0%, 1%, 2%, 3% and 4% of the random values of Y_4 were replaced by the same and respective percentages of outliers, in order to break its randomness. Subsequently, response surface analyzes were performed for eight different absolute mean percentage errors obtained in relation to eight parameters related to the performance of the CP analysis, as a function of the replacement percentages by Y_4 outliers (0, 1, 2, 3 and 4) and the matrices used in the analysis of the PCs ($\boldsymbol{\Sigma}^*$and $\boldsymbol{\Sigma}_{\mathbf{C}}^*$). According to the results, it was concluded that, in the presence of only normal random variables, $\boldsymbol{\Sigma}^*$ it is the best matrix. On the other hand, when there are outliers, $\boldsymbol{\Sigma}_{\mathbf{C}}^*$ it is the most recommended.

- KEYWORDS: Coefficient of variation; Relative importance; Correlation.

1 Introduction

Principal component analysis (CPs) consists of decomposing the covariance matrix ($\boldsymbol{\Sigma}$) composed of n sample elements and p original variables, in order to generate p linear combinations of these, independent of each other and that maximize the total variance of this matrix. In general, it is preferable to work with standardized a variable, that is originais ($\boldsymbol{\rho}$). However, the standardization of variables provides different estimates of the CPs, since they are not invariant to the change of scale (Johnson; Wichern, 2002; Mingoti, 2007; Ferreira, 2009). When using the $\boldsymbol{\Sigma}$ matrix, the scale of the original variables interferes in the analysis of the PCs, which, in most cases, is not good. And when using the $\boldsymbol{\rho}$ matrix, all the original variables become equally important, which also, most of the time, is not good for the analysis.

[1] Universidade Federal de Viçosa - UFV, Departamento de Estatística, CEP: 36570-900, Viçosa, MG, Brazil, E-mail: juliana.v.vieira@ufv.br, camilarafaelagomesdias@gmail.com, jivo@ufv.com.br

Thus, as the different variables present, most of the time, different relative variability, they should necessarily imply different relative importance in the PC analysis. This suggests that the estimates of the PCs more related to the respective relative importance of the variables are the most adequate. In this sense, Campana *et al.* (2010) presented a matrix based on coefficients of variation ($\mathbf{\Sigma}^*$), for the analysis of PCs, in order to take into account the different relative variability of the original variables. However, in this study, all original variables followed multivariate normal distribution.

The analysis of CPs is also sensitive to the presence of outliers, that is, values that significantly distance themselves from the probability distribution of the others and, consequently, are capable of shifting the mean and increasing the variance. Therefore, as there is no requirement for the assumption of multivariate normality (Ferreira, 2009), different types of variables, that is, random variables that follow normal distribution or other probability distributions, as well as those that may not behave randomly due to the presence of outliers, can occur in the $n \times p$ matrix of individual observations obtained without considering experimental designs. This means that it is not possible to decompose the total variation of variable values into parts attributed to known (special) and unknown (random) causes.

Thus, using Lenth's (1989) method, we sought to obtain a new matrix of covariance ($\mathbf{\Sigma}_\mathbf{c}^*$) due, as much as possible, to the actions of random causes, in order to correct the matrix $\mathbf{\Sigma}^*$ proposed by Campana *et al.* (2010) and to use it in the analysis of PCs with the presence of outliers.

2 Literature review

2.1 Main components

The analysis of the CPs, initially conceived by Karl Pearson in 1901 and, based in 1933, by Hotelling (Hotelling, 1933), consists, in general, in obtaining the CPs through the decomposition of a matrix of covariance composed of n sample elements and p original variables. In fact, the CPs are linear combinations of the p original or standardized variables, independent of each other and estimated with the purpose of preserving, in order of estimation, the maximum amount of information in terms of the total variation. Thus, the information contained in the p variables can be replaced by the information contained in the k ($k \leq p$) CPs.

Ideally, the p original variables should be random and that they have p-varied normal distribution. Consequently, the k CPs have k-varied normal distribution. However, to use the CPs it is not necessary that the set of variables is random and even with normal p-varied distribution (Ferreira, 2009). However, the occurrence of random variables or not in the same set of data affects the estimates of the relationships of each variable with each PC, separately (Mingoti, 2007).

Algebraically, CPs represent linear combinations of p original variables Y_1, Y_2, ..., Y_p or p standardized variables Z_1, Z_2, ..., Z_p. Geometrically, these linear combinations represent the choice of new axes coordinated in the directions of maximum variability, which are obtained by system rotations of the original axes.

Consider the p original variables Y_1, Y_2, ..., Y_p with means vector μ (p × 1) and matrix $\mathbf{\Sigma}$ of covariance between Ys (p × p), being:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_p \end{bmatrix}; \text{ and}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}.$$

Denotes $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$ as the eigenvalues of the matrix $\mathbf{\Sigma}$ and with the respective normalized eigenvectors $\mathbf{e_1}$, $\mathbf{e_2}$, ..., $\mathbf{e_k}$, whose k CPs are defined by:

$$CP_1 = e_{11}Y_1 + e_{12}Y_2 + ... + e_{1p}Y_p;$$

$$CP_2 = e_{21}Y_1 + e_{22}Y_2 + ... + e_{2p}Y_p;$$

$$...$$

$$CP_k = e_{k1}Y_1 + e_{k2}Y_2 + ... + e_{kp}Y_p.$$

The method of obtaining the CPs through the $\mathbf{\Sigma}$ matrix is not useful when there are big differences in the scales of the original variables. One way out in this case is to standardize them, which provides means equal to zero and standard deviations equal to one, as follows:

$$Z_w = \frac{Y_w - \mu_w}{\sqrt{\sigma_w^2}}, \text{ to } w = 1, 2, ..., p.$$

In this second case, consider the p standardized variables Z_1, Z_2, ..., Z_p with vector of means $\boldsymbol{\mu_Z}$ (p × 1) and matrix $\boldsymbol{\rho}$ of covariance between Zs or of correlations between Ys (p × p), as follows:

$$\boldsymbol{\mu_Z} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}; \text{ and}$$

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}.$$

Similarly, we denote $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$ as the eigenvalues of the matrix $\boldsymbol{\rho}$ and with the corresponding normalized eigenvectors $\mathbf{e_1}$, $\mathbf{e_2}$, ..., $\mathbf{e_k}$. However, different from the previous ones.

On the other hand, when there is interest in circumventing the differences of the scales when working with the matrix $\mathbf{\Sigma}$ or the equalities of the variances when working with the matrix $\boldsymbol{\rho}$, Campana *et al.* (2010) proposed the following transformation:

Por outro lado, quando há o interesse em contornar as diferenças das escalas quando se trabalha com a matriz $\mathbf{\Sigma}$ ou as igualdades das variâncias quando se trabalha com a matriz $\boldsymbol{\rho}$, Campana *et al.* (2010) propuseram a seguinte transformação:

$$Z_w^* = Z_w \times CV_w = \frac{Y_w - \mu_w}{\mu_w}, \text{ to } w = 1, 2, ..., p, \text{ where:}$$

$$CV_w = \frac{\sqrt{\sigma_w^2}}{\mu_w}.$$

In this third case, all transformed variables Z*s will have mean equal to zero, but their standard deviations will be equal to the respective coefficients of variation of the original variables Ys. Therefore, the p transformed variables $Z_1^*, Z_2^*, \ldots, Z_p^*$ have mean vector $\boldsymbol{\mu}^*$ (p × 1) and matrix $\mathbf{\Sigma}^*$ among the Z*s (p × p), as follows:

$$\boldsymbol{\mu}^* = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{bmatrix}; \text{ and}$$

$$\mathbf{\Sigma}^* = \begin{bmatrix} CV_1^2 & \rho_{12}CV_1CV_2 & \ldots & \rho_{1p}CV_1CV_p \\ \rho_{12}CV_1CV_2 & CV_2^2 & \ldots & \rho_{2p}CV_2CV_p \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{1p}CV_1CV_p & \rho_{2p}CV_2CV_p & \ldots & CV_p^2 \end{bmatrix}.$$

Similarly, we denote $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ as the eigenvalues of the matrix $\mathbf{\Sigma}^*$ and with their respective normalized eigenvectors $\mathbf{e_1}$, $\mathbf{e_2}$, ..., $\mathbf{e_k}$. However and similarly different from the previous two cases.

2.2 Lenth's Method

For factorial experiments with two levels per factor and with individual observations per treatment, it is not possible to test the main effects and those of interactions of different orders using the F test of the analysis of variance, when all of them are considered in the model. In these cases, there are zero degrees of freedom for the residual, and therefore it is not possible to obtain the mean square of the residual (QMRes).

In this sense, Lenth (1989) presented a relatively simple methodology to obtain the estimate of a random standard deviation (σ_{random}), defined as pseudo standard error (PEP) and translated as the square root of QMRes. To do this, he considered the estimates of the effects as being values of a normal random variable. Thus, for m ≤ n, one has:

$$PEP = 1{,}5 \times Md\ |\hat{\mathbf{e}}_\mathbf{g}|$$

where $|\hat{\mathbf{e}}_\mathbf{g}| = |\hat{\mathbf{e}}_\mathbf{i}| < 2{,}5s_0$ = vector that contains only the absolute estimates of the effects to be used in the calculation of the PEP (g = 1, 2, ..., m); $s_0 = 1{,}5 \times Md\ |\hat{\mathbf{e}}_\mathbf{i}|$; $|\hat{\mathbf{e}}_\mathbf{i}|$ = vector containing all absolute estimates of the effects to be tested (i = 1, 2, ..., n); and Md = median.

Thus, given the lack of knowledge of the random part of a given variable, it is, in most cases, inappropriate to use all of its observed values to estimate the σ_{random}. As can be seen, the PEP is obtained based only on the absolute estimates of the effects that are

smaller than $2{,}5s_0$. According to Lenth (1989), the PEP is a consistent estimate of σ_{random} only when there are no significant effects. Otherwise, she overestimates him.

Using PEP to standardize effect estimates, you have:

$$\hat{e}_{p_i} = \frac{\hat{e}_i}{PEP}, \text{ to } i = 1, 2, ..., n.$$

3 Methodology

3.1 Data simulation

To conduct the study, a data set with four random variables (Y_1, Y_2, Y_3 and Y_4) and with 10,000 values each was simulated, according to the multivariate normal distribution. Thus, we have: Y_1, Y_2, Y_3 e $Y_4 \sim N_4$ ($\boldsymbol{\mu}$; $\boldsymbol{\Sigma}$).

In this work, the vector $\boldsymbol{\mu}$ was defined by:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \\ \boldsymbol{\mu_3} \\ \boldsymbol{\mu_4} \end{bmatrix} = \begin{bmatrix} \mathbf{10} \\ \mathbf{20} \\ \mathbf{30} \\ \mathbf{40} \end{bmatrix}.$$

Furthermore, the simulation aimed to provide, for the random variables Y_1, Y_2, Y_3 and Y_4, the same coefficient of variation, in non-percentage units, given by:

$$CV_1 = CV_2 = CV_3 = CV_4 = 0{,}10.$$

Consequently, the matrix $\boldsymbol{\Sigma}$, for $w \neq w'$ (w, w′ = 1, 2, 3 and 4), was obtained by:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma_1^2} & \boldsymbol{\sigma_{12}} & \boldsymbol{\sigma_{13}} & \boldsymbol{\sigma_{14}} \\ \boldsymbol{\sigma_{12}} & \boldsymbol{\sigma_2^2} & \boldsymbol{\sigma_{23}} & \boldsymbol{\sigma_{24}} \\ \boldsymbol{\sigma_{13}} & \boldsymbol{\sigma_{23}} & \boldsymbol{\sigma_3^2} & \boldsymbol{\sigma_{34}} \\ \boldsymbol{\sigma_{14}} & \boldsymbol{\sigma_{24}} & \boldsymbol{\sigma_{34}} & \boldsymbol{\sigma_4^2} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{2} & \mathbf{0} & \mathbf{0} \\ \mathbf{2} & \mathbf{4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{9} & \mathbf{12} \\ \mathbf{0} & \mathbf{0} & \mathbf{12} & \mathbf{16} \end{bmatrix}$$

where $\boldsymbol{\sigma_w^2} = (\mathbf{CV_w} \times \boldsymbol{\mu_w})^2$; $\boldsymbol{\sigma_{ww'}} = \boldsymbol{\rho_{ww'}}\sqrt{\boldsymbol{\sigma_w^2} \times \boldsymbol{\sigma_{w'}^2}}$; $\rho_{12} = \rho_{34} = 1$; e $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = 0$.

And according to the matrix $\boldsymbol{\Sigma}$, was obtained the matrix $\boldsymbol{\Sigma}^*$ (Campana *et al.*, 2010), as follows:

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} CV_1^2 & \rho_{12}CV_1CV_2 & \rho_{13}CV_1CV_3 & \rho_{14}CV_1CV_4 \\ \rho_{12}CV_1CV_2 & CV_2^2 & \rho_{23}CV_2CV_3 & \rho_{24}CV_2CV_4 \\ \rho_{13}CV_1CV_3 & \rho_{23}CV_2CV_3 & CV_3^2 & \rho_{34}CV_3CV_4 \\ \rho_{14}CV_1CV_4 & \rho_{24}CV_2CV_4 & \rho_{34}CV_3CV_4 & CV_4^2 \end{bmatrix} = \begin{bmatrix} 0{,}01 & 0{,}01 & 0 & 0 \\ 0{,}01 & 0{,}01 & 0 & 0 \\ 0 & 0 & 0{,}01 & 0{,}01 \\ 0 & 0 & 0{,}01 & 0{,}01 \end{bmatrix}.$$

Similarly, the matrix $\boldsymbol{\Sigma}^*$ (Campana *et al.*, 2010) can be obtained by:

$$\boldsymbol{\Sigma}^{*} = \begin{bmatrix} \frac{\sigma_1^2}{\mu_1^2} & \frac{\sigma_{12}}{\mu_1\mu_2} & \frac{\sigma_{13}}{\mu_1\mu_3} & \frac{\sigma_{14}}{\mu_1\mu_4} \\ \frac{\sigma_{12}}{\mu_1\mu_2} & \frac{\sigma_2^2}{\mu_2^2} & \frac{\sigma_{23}}{\mu_2\mu_3} & \frac{\sigma_{24}}{\mu_2\mu_4} \\ \frac{\sigma_{13}}{\mu_1\mu_3} & \frac{\sigma_{23}}{\mu_2\mu_3} & \frac{\sigma_3^2}{\mu_3^2} & \frac{\sigma_{34}}{\mu_3\mu_4} \\ \frac{\sigma_{14}}{\mu_1\mu_4} & \frac{\sigma_{24}}{\mu_2\mu_4} & \frac{\sigma_{34}}{\mu_3\mu_4} & \frac{\sigma_4^2}{\mu_4^2} \end{bmatrix}.$$

This meant in the same relative importance of the random variables Y_1, Y_2, Y_3 and Y_4, according to the same magnitude of the relative variability defined by the CV.

Thus, the following estimate of $\boldsymbol{\Sigma}^{*}$ (Campana *et al.*, 2010), for $w \neq w'$ (w, w′ = 1, 2, 3 and 4):

$$\mathbf{S}^{*} = \begin{bmatrix} \frac{s_1^2}{\bar{y}_1^2} & \frac{s_{12}}{\bar{y}_1\bar{y}_2} & \frac{s_{13}}{\bar{y}_1\bar{y}_3} & \frac{s_{14}}{\bar{y}_1\bar{y}_4} \\ \frac{s_{12}}{\bar{y}_1\bar{y}_2} & \frac{s_2^2}{\bar{y}_2^2} & \frac{s_{23}}{\bar{y}_2\bar{y}_3} & \frac{s_{24}}{\bar{y}_2\bar{y}_4} \\ \frac{s_{13}}{\bar{y}_1\bar{y}_3} & \frac{s_{23}}{\bar{y}_2\bar{y}_3} & \frac{s_3^2}{\bar{y}_3^2} & \frac{s_{34}}{\bar{y}_3\bar{y}_4} \\ \frac{s_{14}}{\bar{y}_1\bar{y}_4} & \frac{s_{24}}{\bar{y}_2\bar{y}_4} & \frac{s_{34}}{\bar{y}_3\bar{y}_4} & \frac{s_4^2}{\bar{y}_4^2} \end{bmatrix}, \text{ where:}$$

$$\bar{y}_w = \frac{\sum_{i=1}^{10.000} y_{w_i}}{10.000};$$

$$s_w^2 = \frac{\sum_{i=1}^{10.000} (y_{w_i} - \bar{y}_w)^2}{10.000-1}; \text{ and}$$

$$s_{ww'} = \frac{\sum_{i=1}^{10.000} (y_{w_i} - \bar{y}_w)(y_{w'_i} - \bar{y}_{w'})}{10.000-1}.$$

3.2 Substitution for outliers

After simulating the 10,000 random values Y_1, Y_2. Y_3 and Y_4, have been replaced 100 (1%), 200 (2%), 300 (3%) and 400 (4%) values of the random variable Y_4 were replaced by outliers in order to break its randomness. The outliers of Y_4 were added one-sidedly to the right, considering them as all those greater than:

$$q_3 + 1{,}5a_{iq}$$

where: q_3 = quartil 3; and a_{iq} = interquartile range.

For this, four other simulations of 100, 200, 300, and 400 random values were performed according to a new multivariate normal distribution, that is, Y_1, Y_2, Y_3 and $Y_4 \sim N_4$ ($\boldsymbol{\mu}_{out}$; $\boldsymbol{\Sigma}_{out}$), being:

$$\boldsymbol{\mu}_{out} = \begin{bmatrix} \mathbf{10} \\ \mathbf{20} \\ \mathbf{30} \\ \mathbf{64} \end{bmatrix}; \text{ e}$$

$$\boldsymbol{\Sigma}_{\mathbf{out}} = \begin{bmatrix} \mathbf{1} & \mathbf{2} & \mathbf{0} & \mathbf{0} \\ \mathbf{2} & \mathbf{4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{9} & \mathbf{12} \\ \mathbf{0} & \mathbf{0} & \mathbf{12} & \mathbf{16} \end{bmatrix}.$$

Therefore, five data sets were generated, one without outliers and four with random values of Y_4 replaced by outliers. In the present work, each simulation was repeated ten times, and thus 50 data sets were obtained with 0%, 1%, 2%, 3% and 4% of the random values of Y_4 replaced by outliers.

Consequently, the first multivariate normal distribution in which Y_1, Y_2, Y_3 and $Y_4 \sim N_4$ ($\boldsymbol{\mu}$; $\boldsymbol{\Sigma}$) was slightly modified by the second one in which Y_1, Y_2, Y_3 and $Y_4 \sim N_4$ ($\boldsymbol{\mu}_{\mathbf{out}}$; $\boldsymbol{\Sigma}_{\mathbf{out}}$). Therefore, in data sets with the presence of outliers of Y_4, Therefore, in data sets with the presence of.

For an approximate verification of multivariate normality, the Kolmogorov-Smirnov test was applied separately to each variable Y_w, at 5% significance level, as recommended by Ferreira (2009). In this case, the objective was to confirm the normal randomness of Y_1, Y_2, Y_3 and Y_4 in the data sets without outliers, and the normal randomness of variables Y_1, Y_2 and Y_3 and the non-randomness of variable Y_4, in those with 1%, 2%, 3% and 4% replacement of the random values of Y_4 by outliers. The choice of the Kolmogorov-Smirnov test was due to the data set containing an excessive number of 10,000 values for each variable separately.

3.3 Lenth Method

For this study, the $Y_w - \mu_w$ variances were considered to be the values of a normal random variable with zero mean. Thus, we have:

$$\hat{e}_{w_i} = y_{w_i} - \bar{y}_w, \text{ to } w = 1, 2, 3 \text{ e } 4, \text{ where:}$$

$$\bar{y}_w = \frac{\sum_{i=1}^{10.000} y_{w_i}}{10.000}.$$

Therefore, the random standard deviation of Y_w was estimated by:

$$PEP_w = 1{,}5 \times Md\,|\hat{e}_{w_g}|, \text{ to } w = 1, 2, 3 \text{ and } 4 \text{ and } g = 1, 2, ..., m\ (m \leq 10.000)$$

where: $|\hat{e}_{w_g}| = |\hat{e}_{w_i}| < 2{,}5s_{0_w}$ ($i = 1, 2, ..., 10.000$); and $s_{0_w} = 1{,}5 \times Md\,|\hat{e}_{w_i}|$.

The random covariance between Y_w and $Y_{w'}$ was estimated only based on the respective values of the estimates $\hat{e}_{w_g}$ and $\hat{e}_{w'_g}$, as follows:

$$PEP_{ww'} = \frac{\sum_{g=1}^{m}(y_{w_g} - \bar{y}^*_w)(y_{w'_g} - \bar{y}^*_{w'})}{m-1}, \text{ to } w \neq w' \text{ (w, w' = 1, 2, 3 and 4) and } m \leq 10.000.$$

$$\bar{y}^*_w = \frac{\sum_{g=1}^{m} y_{w_g}}{m}$$

Thus, the matrix $\boldsymbol{\Sigma}^*_{\mathbf{c}}$ replacing matrix $\boldsymbol{\Sigma}^*$ (Campana *et al.*, 2010) was proposed for the analysis of CPs in the presence of outliers, whose estimate was obtained by:

$$\mathbf{S_c^*} = \begin{bmatrix} \frac{PEP_1^2}{\bar{y}_1^{*2}} & \frac{PEP_{12}}{\bar{y}_1^*\bar{y}_2^*} & \frac{PEP_{13}}{\bar{y}_1^*\bar{y}_3^*} & \frac{PEP_{14}}{\bar{y}_1^*\bar{y}_4^*} \\ \frac{PEP_{12}}{\bar{y}_1^*\bar{y}_2^*} & \frac{PEP_2^2}{\bar{y}_2^{*2}} & \frac{PEP_{23}}{\bar{y}_2^*\bar{y}_3^*} & \frac{PEP_{24}}{\bar{y}_2^*\bar{y}_4^*} \\ \frac{PEP_{13}}{\bar{y}_1^*\bar{y}_3^*} & \frac{PEP_{23}}{\bar{y}_2^*\bar{y}_3^*} & \frac{PEP_3^2}{\bar{y}_3^{*2}} & \frac{PEP_{34}}{\bar{y}_3^*\bar{y}_4^*} \\ \frac{PEP_{14}}{\bar{y}_1^*\bar{y}_4^*} & \frac{PEP_{24}}{\bar{y}_2^*\bar{y}_4^*} & \frac{PEP_{34}}{\bar{y}_3^*\bar{y}_4^*} & \frac{PEP_4^2}{\bar{y}_4^{*2}} \end{bmatrix}$$

3.4 Main components

After generating the 50 data sets, two analyzes of the CPs, were performed, according to the estimates of the matrices $\mathbf{\Sigma}^*$ (Campana *et al.*, 2010) and $\mathbf{\Sigma_c^*}$,, respectively:

$$\left|\mathbf{S}^* - \hat{\lambda}_j\mathbf{I}\right| = 0 \text{ and } \left[\mathbf{S}^* - \hat{\lambda}_j\mathbf{I}\right]\hat{\mathbf{a}}_\mathbf{j} = \mathbf{0}, \text{ to j = 1 e 2; and}$$

$$\left|\mathbf{S_c^*} - \hat{\lambda}_j\mathbf{I}\right| = 0 \text{ and } \left[\mathbf{S_c^*} - \hat{\lambda}_j\mathbf{I}\right]\hat{\mathbf{a}}_\mathbf{j} = \mathbf{0}, \text{ to j = 1 and 2, where:}$$

$$\hat{\mathbf{e}}_\mathbf{j} = \frac{1}{\|\hat{\mathbf{a}}_\mathbf{j}\|}\hat{\mathbf{a}}_\mathbf{j};$$

$\hat{\mathbf{e}}_\mathbf{1}' = [\hat{e}_{11} \quad \hat{e}_{12} \quad \hat{e}_{13} \quad \hat{e}_{14}]$ (estimation of the normalized eigenvector of the CP1); and

$\hat{\mathbf{e}}_\mathbf{2}' = [\hat{e}_{21} \quad \hat{e}_{22} \quad \hat{e}_{23} \quad \hat{e}_{24}]$ (estimation of the normalized eigenvector of the CP2).

In the analysis of the CPs, only the first two were estimated. And as a reference CP_1 was related to Y_3 and Y_4 ($\rho_{34} = 1$) and, CP_2, to Y_1 and Y_2 ($\rho_{12} = 1$). Thus, in parametric terms, we have:

$$CP_1 = e_{13}Z_3^* + e_{14}Z_4^*;$$

$$CP_2 = e_{21}Z_1^* + e_{22}Z_2^*;$$

$$Z_1^* = \frac{Y_1-\mu_1}{\mu_1} = \frac{Y_1-10}{10};$$

$$Z_2^* = \frac{Y_2-\mu_2}{\mu_2} = \frac{Y_2-20}{20};$$

$$Z_3^* = \frac{Y_3-\mu_3}{\mu_3} = \frac{Y_3-30}{30};$$

$$Z_4^* = \frac{Y_4-\mu_4}{\mu_4} = \frac{Y_4-40}{40};$$

$$e_{13} = e_{14} = e_{21} = e_{22} = 0{,}7071; \text{ and}$$

$$e_{11} = e_{12} = e_{23} = e_{24} = 0.$$

Therefore, the parameters of eigenvalues, relative importance and correlations in the analysis of the CPs, were defined by:

$$\lambda_1 = \lambda_2 = 0{,}02 \text{ and } \lambda_3 = \lambda_4 = 0;$$

$$IR_{CP_1}(\%) = IR_{CP_2}(\%) = 50 \text{ and } IR_{CP_3}(\%) = IR_{CP_4}(\%) = 0; \text{ and}$$

$$\rho_{CP_1Y_3} = \rho_{CP_1Y_4} = \rho_{CP_2Y_1} = \rho_{CP_2Y_2} = 1 \text{ and } \rho_{CP_1Y_1} = \rho_{CP_1Y_2} = \rho_{CP_2Y_3} = \rho_{CP_2Y_4} = 0.$$

On the other hand, it could have adopted CP_1 related to Y_1 and Y_2 and, CP_2, to Y_3 and Y_4, given the same relative importance of the four random variables and the correlations of Y_1 with Y_2 and Y_3 with Y_4, to be equal to one . However, it was preferred to adopt CP_1 with Y_3 and Y_4 and CP_2 with Y_1 and Y_2, because in data sets with Y_4 outliers, this variable, with increased relative variability, will increase the CP variance with the highest eigenvalue. In this case, the CP_1.

3.5 Statistical Analysis

To evaluate the estimation quality of the CPs based on the matrices $\mathbf{S}^*$ and $\mathbf{S}_\mathbf{c}^*$ and the five percent substitution of the random values of Y_4 for outliers, the following mean absolute percentage errors (EPMAs) were estimated:

$$EPMA_{CV_{123}} = 100 \times \frac{1}{10}\sum_{r=1}^{10} a_r \text{, to } a_r = \frac{1}{3}\sum_{w=1}^{3}\left|\frac{CV_w - \widehat{CV}_w}{CV_w}\right| = \frac{1}{3}\sum_{w=1}^{3}\left|\frac{0{,}10 - \widehat{CV}_w}{0{,}10}\right|;$$

$$EPMA_{CV_4} = 100 \times \frac{1}{10}\sum_{r=1}^{10} b_r, \text{to } b_r = \left|\frac{CV_4 - \widehat{CV}_4}{CV_4}\right| = \left|\frac{0{,}10 - \widehat{CV}_4}{0{,}10}\right|;$$

$$EPMA_{\lambda_1} = 100 \times \frac{1}{10}\sum_{r=1}^{10} d_r, \text{to } d_r = \left|\frac{\lambda_1 - \hat{\lambda}_1}{\lambda_1}\right| = \left|\frac{0{,}02 - \hat{\lambda}_1}{0{,}02}\right|;$$

$$EPMA_{\lambda_2} = 100 \times \frac{1}{10}\sum_{r=1}^{10} e_r, \text{to } e_r = \left|\frac{\lambda_2 - \hat{\lambda}_2}{\lambda_2}\right| = \left|\frac{0{,}02 - \hat{\lambda}_2}{0{,}02}\right|;$$

$$EPMA_{e_{13}} = 100 \times \frac{1}{10}\sum_{r=1}^{10} f_r, \text{to } f_r = \left|\frac{e_{13} - \hat{e}_{13}}{e_{13}}\right| = \left|\frac{0{,}7071 - \hat{e}_{13}}{0{,}7071}\right|;$$

$$EPMA_{e_{14}} = 100 \times \frac{1}{10}\sum_{r=1}^{10} g_r, \text{to } g_r = \left|\frac{e_{14} - \hat{e}_{14}}{e_{14}}\right| = \left|\frac{0{,}7071 - \hat{e}_{14}}{0{,}7071}\right|;$$

$$EPMA_{e_{21}} = 100 \times \frac{1}{10}\sum_{r=1}^{10} h_r, \text{to } h_r = \left|\frac{e_{21} - \hat{e}_{21}}{e_{21}}\right| = \left|\frac{0{,}7071 - \hat{e}_{21}}{0{,}7071}\right|; \text{ and}$$

$$EPMA_{e_{22}} = 100 \times \frac{1}{10}\sum_{r=1}^{10} i_r, \text{to } i_r = \left|\frac{e_{22} - \hat{e}_{22}}{e_{22}}\right| = \left|\frac{0{,}7071 - \hat{e}_{22}}{0{,}7071}\right|.$$

The EPMAs show the absolute differences between the parameters and the estimates obtained from the two matrices. Therefore, for a good analysis of the CPs, it is expected that all eight differences are equal to zero.

Subsequently, for each EPMA evaluated, a response surface analysis was performed as a function of the matrices and the percentages of substitution of the random values of Y_4 for outliers, whose regression coefficients were tested by Student's t-test at 5% significance and whose largest response surface model adopted was given by:

$$ep_{st} = \beta_0 + \beta_1 p_s + \beta_2 p_s^2 + \beta_3 m_t + \beta_4 p_s m_t + \varepsilon_{st}$$

where ep_{st} = observed value of EPMA in the combination between levels s (0, 1, 2, 3 and 4) e t (0 and 1); p_s = percentage of replacement of random values of Y_4 by outliers at level s (0, 1, 2, 3 and 4); m_t = matrix used in the analysis of CPs at level t ($\mathbf{S}^* = 0$ and $\mathbf{S}_\mathbf{c}^* = 1$);

β_0 = regression constant; β_1, β_2, β_3 and β_4 = regression coefficients; and ε_{st} = regression error associated with the observed value ep_{st}, where $\varepsilon \sim N(0; \sigma_\varepsilon^2)$.

R version 4.0.2 (R CORE TEAM, 2020) was used for the statistical analyses.

4 Results and Discussion

As expected, the original variables Y_1, Y_2 e Y_3 were confirmed to be normally distributed ($P > 0.05$) in all 50 data sets. Variable Y_4 on the other hand, was randomly normal ($P > 0.05$) only in the 10 data sets without outliers. In the remaining sets with 1%, 2%, 3%, and 4% replacement of the random values of Y_4 with outliers, it was concluded that it did not ($P < 0.05$) behave normally. This suggested that in these 40 data sets with outliers of Y_4, its standard deviation was not exclusively constituted by random causes.

The increased ($P < 0.05$) as a function of increasing percent substitution of random values of Y_4 for outliers and was higher ($P < 0.05$) according to the matrix, as shown in the following fitted response surface:

$$\widehat{ep}_{st} = 0{,}4671 + 0{,}0602^{*}p_s + 0{,}4013^{*}m_t \ (R^2 = 0{,}92)$$

where $\widehat{ep}_s = 0{,}4671 + 0{,}0602p_s$, for the matrix $\mathbf{S}^*$;$\widehat{ep}_t = 0{,}8684 + 0{,}0602p_s$, for the matrix $\mathbf{S}_\mathbf{c}^*$; and *: significance by t-test of Student ($P < 0{,}05$).

According to the matrices $\mathbf{S}^*$ and $\mathbf{S}_\mathbf{c}^*$, the largest estimates of the largest estimates of $EPMA_{CV_{123}}$ (0.71% and 1.11%) occurred for the data set with 4% outliers of Y_4, respectively. Thus, it was concluded that the increase in the percentage of substitution of the random values of Y_4 by outliers and the matrix $\mathbf{S}_\mathbf{c}^*$, impaired the estimates of the coefficients of variation of the normal random variables Y_1, Y_2 and Y_3. This means that in an analysis of CPs with variables without or with outliers, the relative variabilities of those variables that are free of outliers will also be inflated, even if in small proportions.

On the other hand, $EPMA_{CV_4}$ increased ($P < 0.05$) as a function of increasing percentage of substitution of random values of Y_4 for outliers, but much more pronouncedly, and showed higher ($P < 0.05$) according to the matrix $\mathbf{S}^*$, as per the following fitted response surface:

$$\widehat{ep}_{st} = 2{,}8111 + 12{,}5421^{*}p_s - 2{,}2695m_t - 11{,}9751^{*}p_s m_t \ (R^2 = 1{,}00)$$

where $\widehat{ep}_s = 2{,}8111 + 12{,}5421p_s$, for the matrix $\mathbf{S}^*$;
$\widehat{ep}_s = 0{,}5416 + 0{,}5670p_s$, for the matrix $\mathbf{S}_\mathbf{c}^*$; and *: significance by t-test of Student ($P < 0{,}05$).

The highest $EPMA_{CV_4}$ estimates (52.98% and 2.81%) also occurred for the dataset with 4% estimates (52.98% and 2.81%) also occurred for the dataset with 4% Y_4 outliers, according to the and matrices $\mathbf{S}^*$ and $\mathbf{S}_\mathbf{c}^*$, respectively. However, only the second estimate of the $EPMA_{CV_4}$ was considered low. This means that in an analysis of the CPs with variables without or with outliers the relative variability of those that present them will be better estimated when the matrix $\mathbf{S}_\mathbf{c}^*$ is used. Consequently, the matrix $\mathbf{S}^*$ was not able to avoid the effect of non-randomness, here represented by the presence of outliers, on the estimate of the coefficient of random variation.

Likewise, $EPMA_{\lambda_1}$ and the $EPMA_{\lambda_2}$ increased ($P < 0.05$) as a function of the increase in the percentage of replacement of random Y4 values by outliers, but only according to the matrix $\mathbf{S}^*$, as the response surfaces adjusted shown in Table 1.

Table 1 - Adjusted response surfaces for $EPMA_{\lambda_1}$ and the $EPMA_{\lambda_2}$

EPMA	Response Surface	R^2
λ_1	$1{,}1688 + 9{,}9928^*p_s + 1{,}8306^*m_t - 10{,}2671^*p_s m_t$	1,00
	$1{,}1688 + 9{,}9928p_s$, to the matrix $\mathbf{S}^*$	
	$2{,}9994 - 0{,}2743p_s$, to the matrix $\mathbf{S}_\mathbf{c}^*$	
λ_2	$0{,}5974 + 0{,}1055^*p_s + 1{,}8021^*m_t - 0{,}2793^*p_s m_t$	0,99
	$0{,}5974 + 0{,}1055p_s$ to the matrix $\mathbf{S}^*$	
	$2{,}3995 - 0{,}1738^*p_s$, to the matrix $\mathbf{S}_\mathbf{c}^*$	

*: significant by Student's t-test ($P < 0.05$).

The highest estimates of $EPMA_{\lambda_1}$ (41.14% and 3.00%) occurred for datasets with 4% and 0% of Y_4, outliers, according to the and matrices $\mathbf{S}^*$ and $\mathbf{S}_\mathbf{c}^*$, respectively. However, only the first estimate of $EPMA_{\lambda_1}$ was considered high. According to the matrix $\mathbf{S}^*$, the increase in the estimate of $EPMA_{\lambda_1}$ was probably due to the increase λ_1 in as a function of the increase in CV_4 which, in this way, implied a greater relative importance of Y_4 in CP_1. In this case, incorrectly. Thus, when using the matrix $\mathbf{S}^*$, the estimation of relative variability interferes with the estimation of the eigenvalue (CAMPANA *et al.*, 2010).

Consequently, in an analysis of the CPs with variables without and with outliers, the estimate of the will be overestimated λ_1 mainly due to the relative variability, also overestimated, of the variables with outliers. Therefore, the variables with outliers will probably have a much higher relative importance than those that do not, which will harm the quality of the analysis performed with the matrix $\mathbf{S}^*$.

On the other hand, the $EPMA_{\lambda_2}$ estimates (1.02% and 2.40%) that occurred for the data sets with 4% and 0% outliers of Y_4, proved to be very low for both matrices $\mathbf{S}^*$ and $\mathbf{S}_\mathbf{c}^*$, respectively.

The $EPMA_{e_{13}}$ and the $EPMA_{e_{14}}$ increased ($P < 0{,}05$) as a function of the increase in the percentage of replacement of random Y_4 values by outliers and were smaller ($P < 0{,}05$) according to the matrix $\mathbf{S}_\mathbf{c}^*$, according to the adjusted response surfaces presented in Table 2.

Table 2 - Response surfaces fitted for the $EPMA_{e_{13}}$ and the $EPMA_{e_{14}}$

EPMA	Superfície de resposta	R^2
e_{13}	$5{,}5714 + 5{,}7735^*p_s - 6{,}9286^*m_t$	0,88
	$5{,}5714 + 5{,}7735^*p_s$ to the matrix $\mathbf{S}^*$	
	$-1{,}3572 + 5{,}7735^*p_s$ to the matrix $\mathbf{S}_\mathbf{c}^*$	
e_{14}	$4{,}9054 + 4{,}4676^*p_s - 4{,}6345^*m_t$	0,87
	$4{,}9054 + 4{,}4676^*p_s$ to the matrix $\mathbf{S}^*$	
	$0{,}2709 + 4{,}4676^*p_s$ to the matrix $\mathbf{S}_\mathbf{c}^*$	

*: significant by Student's t-test ($P < 0{,}05$).

The highest estimates of $EPMA_{e_{13}}$ (28,67% and 21,74%) and of $EPMA_{e_{14}}$ (22,78% and 18,14%) occurred for the dataset with 4% of Y_4, outliers, according to the matrices $\mathbf{S}^*$and $\mathbf{S}_\mathbf{c}^*$, respectively. Although both matrices failed to avoid the effect of outliers on the coefficient estimates of the two normalized eigenvectors, especially for the higher percentages of substitution of random values of Y_4 by outliers, the matrix $\mathbf{S}_\mathbf{c}^*$ was more efficient because it provided lower error estimates. Again, it was noticed the greatest increase in the estimation error due to the variable Y4 with outliers, when the matrix $\mathbf{S}^*$ was used.

The $EPMA_{e_{21}}$and $EPMA_{e_{22}}$ suffered ($P < 0,05$) only the influence of the interaction of the percentages of substitution of the random values of Y_4 for outliers with the matrices, according to the fitted response surfaces presented in Table 3.

Table 3 - Response surfaces fitted for the $EPMA_{e_{21}}$ and the $EPMA_{e_{22}}$

EPMA	Superfície de resposta	R^2
e_{21}	$1{,}7664 - 0{,}5861p_s + 0{,}3161m_t + 4{,}3893^*p_sm_t$ $1{,}7664 - 0{,}5861p_s$, to the matrix$\mathbf{S}^*$ $2{,}0825 + 3{,}8032p_s$ to the matrix $\mathbf{S}_\mathbf{c}^*$	0,88
e_{22}	$1{,}7664 - 0{,}5860p_s + 0{,}3373m_t + 4{,}3857^*p_sm_t$ $1{,}7652 - 0{,}5860p_s$, to the matrix $\mathbf{S}^*$ $2{,}1025 + 3{,}7997p_s$ to the matrix $\mathbf{S}_\mathbf{c}^*$	0,88

*: significant by Student's t-test ($P < 0,05$).

For $EPMA_{e_{21}}$ and $EPMA_{e_{22}}$, the matrix $\mathbf{S}^*$ provided the best results. Furthermore, the highest estimates of $EPMA_{e_{21}}$ and of $EPMA_{e_{22}}$ (1.77% and 17.30%) occurred for the dataset with 4% of Y_4 outliers, according to the and matrices $\mathbf{S}^*$ and $\mathbf{S}_\mathbf{c}^*$, respectively. Again, it was noticed that the matrix $\mathbf{S}^*$ does not provide estimation errors arising from the random variables Y_1, Y_2 and Y_3.

According to Lawson (2008), outliers can be the main impediment to validating the interpretation of experimental results. Thus, the use of a matrix that minimizes their effect on the data set will provide an estimate of the CPs more correctly.

As observed in the results of $EPMA_{\lambda_1}$, of $EPMA_{e_{13}}$ and of $EPMA_{e_{14}}$, the matrix $\mathbf{S}^*$ was greatly affected by the direct or indirect ($\rho_{13} = 1$) presence of the Y_4 variable with outliers. This clearly showed that it should be avoided when there are outliers in the CPs analysis. On the other hand, as it was not harmed by the direct and indirect presence of random variables Y_1, Y_2 and Y_3, as observed in the results of $EPMA_{\lambda_2}$, of $EPMA_{e_{21}}$ and of $EPMA_{e_{22}}$, it is concluded that it should be used when only random variables occur in the analysis of the CPs. This conclusion was due to the damage suffered by the matrix $\mathbf{S}_\mathbf{c}^*$, with the presence of variable Y_4 with outliers on the estimates of the parameters associated with the random variables Y_1, Y_2 and Y_3.

However, as the results associated with the matrix $\mathbf{S}_\mathbf{c}^*$ were, in general, better when there is the presence of the Y_4 variable with outliers, it recommends it for the analysis of the CPs when these outliers are present.

According to the construction of the scatter diagrams of the scores from CP_2 by those from CP_1, the matrices $\mathbf{S}^*$ (Figure 1a) and $\mathbf{S}_\mathbf{c}^*$ (Figure 1c) provided the same results

only when considering 0% of outliers. On the other hand, when 4% of Y_4, outliers were considered, the dispersion provided by the matrix $\mathbf{S}_{\mathbf{c}}^{*}$ (Figure 1d) was better, given that in this case, the Y_4 outliers were less prominent than in the dispersion provided by the matrix $\mathbf{S}^{*}$ (Figure 1c). According to the matrix $\mathbf{S}_{\mathbf{c}}^{*}$, the configurations provided with 0% and 4% of Y_4 outliers were more similar to each other (Figure 1c, d). By the matrix $\mathbf{S}^{*}$, these configurations were more different from each other (Figure 1a, b). According to the last two, CP_1 showed greater variance and Y_4 outliers occurred, with great emphasis, along its direction (Figure 1b).

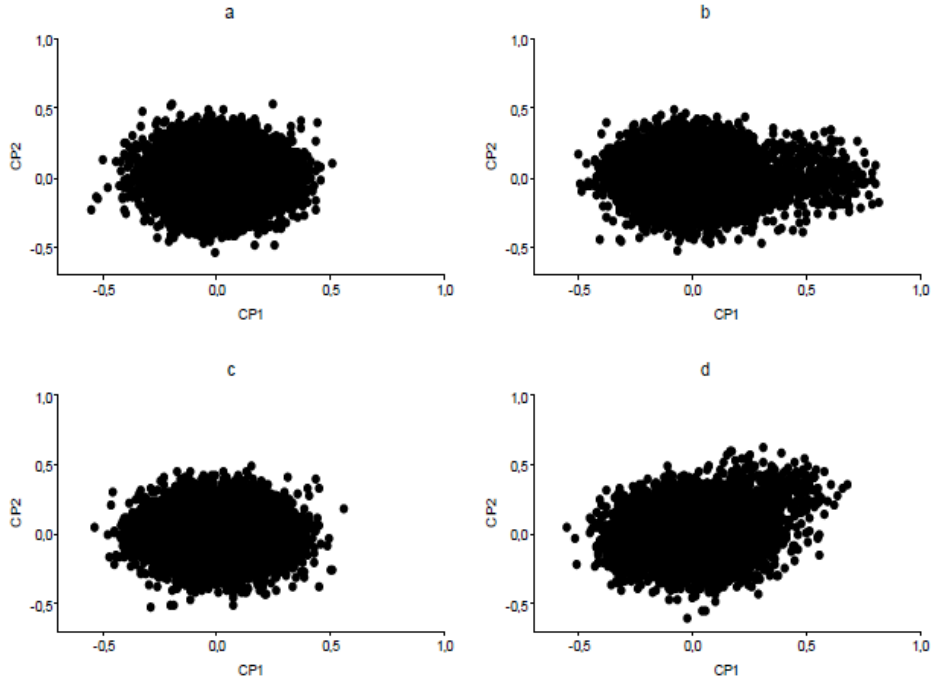

Figure 1 - Scatter diagrams of CP_2 by CP_1 according to the scores obtained based on matrices $\mathbf{S}^{*}$ with 0% outliers Y_4 (a), $\mathbf{S}^{*}$ with 4% outliers Y_4 (b), $\mathbf{S}_{\mathbf{c}}^{*}$ with 0% outliers Y_4 (c) and $\mathbf{S}_{\mathbf{c}}^{*}$ with 4% outliers of Y_4 (d).

5 Conclusions

When only variables without outliers occur in the analysis of the principal components, it is recommended to use the matrix based on coefficients of variation.

When there is at least one variable with outliers in the analysis of the principal components, it is recommended to correct the matrix based on the coefficients of variation by applying the method of Lenth (1989).

GOMES, J. V., JÚNIOR, J. I. R., DIAS, C. R. G. uma proposta para análise de componentes principais na presença de *outliers*. *Braz. J. Biom*. Lavras, v.40, n.3, p.242-255, 2022.

- *RESUMO: Para análise exploratória dos componentes principais (CPs), não é exigida a pressuposição de normalidade multivariada das variáveis e, nem necessariamente, que elas sejam aleatórias. Isso significa que variáveis que não se comportam aleatoriamente também podem ser incluídas nessa análise. Desse modo, a fim de realizar a análise dos CPs com variáveis aleatórias ou não, foi proposta uma correção da matriz* $\boldsymbol{\Sigma}^*$ *baseada nos coeficientes de variação (Campana et al., 2010) por meio da aplicação do método de Lenth (1989), cuja nova matriz foi denominada* $\boldsymbol{\Sigma}_{\boldsymbol{c}}^*$*. Para verificar a sua viabilidade, foram simulados dez conjuntos de dados das variáveis aleatórias* Y_1*,* Y_2*,* Y_3 *e* Y_4*, com 10.000 valores cada e que seguiram distribuição normal multivariada. Após a simulação, foram substituídos 0%, 1%, 2%, 3% e 4% dos valores aleatórios de* Y_4 *pelos mesmos e respectivos percentuais de outliers, com o objetivo de quebrar a aleatoriedade da mesma. Posteriormente, foram realizadas análises de superfícies de respostas para oito diferentes erros percentuais médios absolutos obtidos em relação a oito parâmetros relacionados ao desempenho da análise dos CPs, em função dos percentuais de substituição por outliers de* Y_4 *(0, 1, 2, 3 e 4) e das matrizes utilizadas na análise dos CPs (*$\boldsymbol{\Sigma}^* = 0$ *e* $\boldsymbol{\Sigma}_{\boldsymbol{c}}^* = 1$*). De acordo com os resultados, concluiu-se que, na presença de apenas variáveis aleatórias normais,* $\boldsymbol{\Sigma}^*$ *é a melhor matriz. Por outro lado, quando há a presença de outliers,* $\boldsymbol{\Sigma}_{\boldsymbol{c}}^*$ *é a mais recomendada.*

- *PALAVRAS-CHAVE: Coeficiente de variação; Importância relativa; Correlação.*

References

CAMPANA, A. C. M.; RIBEIRO JÚNIOR, J. I.; NASCIMENTO, M. Uma proposta de transformação de dados para a análise de componentes principais. *Revista Brasileira de Biometria*, v. 28, p. 1-15, 2010.

FERREIRA, D. F. M *Estatística multivariada*. 2. ed. Lavras: Editora UFLA, 2009. 676 p.

HOTELLING, H. Review of the triumph of mediocrity in business. *Journal of the American Statistical Association*. v. 28, p. 463-465, 1933.

JOHNSON, R. A; WICHERN, D. W. *Applied multivariate statistical analysis*. 5. ed. New Jersey: Prentice Hall, 2002. 767 p.

LAWSON, J. SAS macros for analysis of unreplicated 2^k and 2^{k-p} designs with a possible outlier. *Journal of Statistical Software*, v. 25, p. 1-17, 2008.

LENTH, R. V. Quick and easy analysis of unreplicated factorials. *Technometrics*, v. 31, p. 469-473, 1989.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada – uma abordagem aplicada*. Belo Horizonte: Editora UFMG, 2007. 297 p.

R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2020. URL https://www.r-project.org.

Received in 21.01.2021

Approved after revised in 12.08.2021