Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1806-2563
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v28i2.26046

Principal components in the discrimination of outliers: A study in simulation sample data corrected by Pearson's and Yates´s chi-square distance

Manoel Vitor de Souza Veloso[1] and Marcelo Angelo Cirillo[2*]

*[1]Instituto de Ciências Sociais Aplicadas, Campus Avançado de Varginha, Universidade Federal de Alfenas, Varginha, Minas Gerais, Brazil. [2]Departamento de Ciências Exatas, Universidade Federal de Lavras, Cx. Postal 3037, 37200-000, Lavras, Minas Gerais, Brazil. *Author for correspondence. E-mail: macufla@dex.ufla.br*

ABSTRACT. Current study employs Monte Carlo simulation in the building of a significance test to indicate the principal components that best discriminate against outliers. Different sample sizes were generated by multivariate normal distribution with different numbers of variables and correlation structures. Corrections by chi-square distance of Pearson´s and Yates's were provided for each sample size. Pearson´s correlation test showed the best performance. By increasing the number of variables, significance probabilities in favor of hypothesis H_0 were reduced. So that the proposed method could be illustrated, a multivariate time series was applied with regard to sales volume rates in the state of Minas Gerais, obtained in different market segments.

Keywords: contaminated samples, Monte Carlo, significance test, p-value.

Componentes principais na discriminação de outliers: estudo de simulação em dados amostrais corrigidos pelas distâncias qui-quadrado de Pearson's and Yates.

RESUMO. Este trabalho tem por objetivo realizar um estudo, utilizando simulação Monte Carlo na construção de um teste de significância para indicar os componentes principais que melhor discriminam as discrepâncias. Neste contexto, diferentes tamanhos amostrais foram gerados pela distribuição normal multivariada com diferentes números de variáveis e estruturas de correlação. Para cada tamanho amostral, procedeu-se com as correções dadas pela distância qui-quadrado de Pearson e Yates. Concluiu-se ao considerar a correção de Pearson o teste apresentou melhor desempenho, entretanto, aumentando o número de variáveis as probabilidades de significância a favor a hipótese H_0 foram reduzidas. Por fim, para ilustrar a metodologia proposta realizou-se uma aplicação em uma série temporal multivariada referente a índices de volumes de vendas do estado de Minas Gerais obtidos em diferentes segmentos de mercados.

Palavras-chave: amostras contaminadas, Monte Carlo, teste de significância, p-value.

Introduction

A multivariate outlier is an observation that appears at great distance from the others on the p-dimensional space defined by all variables. Identification is usually performed with methods based on graph construction. One of the major contributions has been reported by Gnanadesikan and Kettenring (1972), Filzmoser (2005) and Filzmoser, Maronna and Werner (2008).

With regard to the principal component analysis, employed as an investigation method for outlier detection, Steiner, Neto, Braulio and Alves (2008) report that components are sensitive to outliers since component estimation may be influenced. An alternative to this problem is given by the implementation of robust statistical methods to outlier observations applied in the estimation of the principal components (Caroni, 2000; Peña & Prieto, 2001; Jackson & Chen, 2004; Bénasséni, 2005; Caroni & Billor, 2007; Silva, Moraes, & Cirillo, 2013). Enki, Trendafilov and Jollife (2013) argue that, as a rule, outlier identification via principal components method may be contradictory since it depends on which components are considered. In addition, the last components are most likely to provide additional information not available in the plot of the original variables. The authors also mention the possibility of outliers occurring at different direction from those detectable in a simple plot of the original variables in experiments with too many variables.

A single outlier may cause the distortion of the principal components to fit better the outlier,

leading to a bad interpretation of results. Outliers may also cause the so-called masking effect: due to their presence, the model is distorted in such a way that, based on the principal components, no outliers are detected (Serneels & Verdonck, 2008)

With respect to the method of time series, the effect of outliers is associated with structural changes that may be related to unexpected events, such as economic crisis, strikes or wars, measurement errors or inadequate recording of information.

A significant contribution in identifying outliers in time series has been given by Fox (1972) who differentiated independent outliers affecting only a single observation from anomalous observations which influence successive observations. Thereby, the author introduced the concept of additive and innovative outliers. Thus, it was suggested that the two new time series models describing possible disruption in the series would accommodate the effect of such observations. Each model was defined according to the author´s classification. Subsequently, Chan (1992) showed that both models may be considered particular cases of the intervention analysis model. In general, outliers have been treated by intervention analysis in time series in many applications.

In the case of disturbances occurring in time series models, Tsay, Peña and Pankratz (2000) established four types of disturbance commonly used in univariate time series. The authors investigated the effects on the dynamic structure of a model in a multidimensional context, with real examples and proposals for further analysis.

In case of high dimensions, Filzmoser et al. (2008) presented a computationally fast procedure with the techniques of the principal components analysis. In summary, the method is formalized by constructing a procedure through which re-scaling data are calculated.

Regarding to the use of chi-square distance used in the improvement of multivariate methods, Knüsel (2008) proposed a new method of factorial rotation based on chi-square statistics, the Chisquaremax criterion. Pereira, Cirillo and Oliveira (2014) concluded that the efficiency of covariance matrix estimator provided by the factorial model using either Chisquaremax or Promax criteria was not affected by the presence of outliers.

Thus, a methodological contribution is proposed: improvement of the re-scaling procedure by Filzmoser et al. (2008) by incorporating chi-square corrections to data sample for building a significance test to evaluate the use of the first principal component. Monte Carlo simulation studies were presented for validating the methodology and applied in a multivariate time series regarding sales volume rates at different market segments in the state of Minas Gerais, Brazil.

Re-scaled data

Given a multivariate sample matrix, re-scaling operations started by setting the matrix $G^{(0)}$, whose element at ij position was represented by $g_{ij}^{(0)}$ referring to the i-th sample unit (i = 1, ..., N) and the j-th variable (j = 1, ..., r), in which N is the sample size and r is the number of observed variables. According to this notation, the vector of the i-th sample unit was rewritten as $G_i^{(0)}=\left[G_{1j}^{(0)},G_{2j}^{(0)},...,G_{ir}^{(0)}\right]$, and the vector of the j-th variable as $G_j^{(0)}=\left[G_{1j}^{(0)},G_{2j}^{(0)},...,G_{Nj}^{(0)}\right]$. With such specifications, the data re-scaling approach proposed by Filzmoser et al. (2008) was performed, considering the median of observations in each observed variable. Thus, re-scaling was based on Equations 1 and 3,

$$g_{ij}^{(1)} = \frac{g_{sij}^{(0)} - \text{median}\left(g_{1j}^{(0)},...,g_{Nj}^{(0)}\right)}{\text{MAD}\left(g_{1j}^{(0)},...,g_{Nj}^{(0)}\right)};\ j=1,...,r\,. \quad (1)$$

The median absolute deviation was obtained by Equation 2:

$$\text{MAD}\left(g_1^{(0)},...,g_N^{(0)}\right) = 1,4826\times \text{median}_j\left|g_j^{(0)} - \text{median}_i\left(g_i^{(0)}\right)\right|, \quad (2)$$

where:
1.4826 is the rate corresponding to quantile 75% of the standard univariate normal distribution suggested by Rousseeuw (1984). **A**ssuming elements given by $g_{ij}^{(1)}$ of the matrix $G^{(1)}$, a matrix eigenvectors in r-order defined by V was obtained, making it possible to obtain the components´ matrix according to the expression $G^{(2)}=G^{(1)}\times V$. Applying back rescaling, matrix $G^{(2)}$ was obtained, each element being obtained by Equation 3.

$$g_{ij}^{(2)} = \frac{g_{ij}^{(1)} - \text{median}\left(g_{ij}^{(1)},...,g_{Nj}^{(1)}\right)}{\text{MAD}\left(g_{ij}^{(1)},...,g_{Nj}^{(1)}\right)};\ j=1,...,r\,, \quad (3)$$

The median absolute deviation was obtained by Equation 4:

$$\text{MAD}\left(g_1^{(1)},...,g_N^{(1)}\right) = 1,4826\times \text{median}_j\left|g_j^{(1)} - \text{median}_i\left(g_i^{(1)}\right)\right|. \quad (4)$$

Based on the re-scaled data represented in matrix $G^{(2)}$, the absolute rate of kurtosis defined in Equation 5 for each variable defined by ω_j was calculated:

$$\omega_j = \left| \frac{1}{N} \sum_{i=1}^{N} \frac{\left[g_{ij}^{(2)} - \text{median}\left(g_{1j}^{(2)}, ..., g_{Nj}^{(2)} \right) \right]^4}{\left[\text{MAD}\left(g_{1j}^{(2)}, ..., g_{Nj}^{(2)} \right) \right]^4} - 3 \right|; j = 1, ..., r. \quad (5)$$

So that these coefficients could be better interpreted, standardization (Equation 6) was undertaken so that the components associated with the highest and lowest rate of ξ_j may better detect outliers, according to Peña and Prieto (2001).

$$\xi_j = \frac{\omega_j}{\sum_{j=1}^{r} \omega_j}. \quad (6)$$

Methodology

The methodological contribution of current study is given in three stages: Incorporation of Pearson's chi-square and Yates´s correction to the re-scaled data; Construction of the significance test for kurtosis coefficient associated with the first principal component and Evaluation of the Monte Carlo simulation test.

Incorporation of Pearson's chi-square and Yates´s correction to re-scaled data

Incorporation of chi-square corrections to sample data was performed by replacing matrix $G^{(0)}$ for matrices Q_P and Q_Y. Q_P element q_{Pij} was calculated with Pearson's chi-square (Equation 7), whereas Q_Y element q_{Yij} was obtained by Yates´s correction (Equation 8).

$$q_{Pij} = \frac{g_{ij}^{(0)} - \sum_{i=1}^{N} g_{ij}^{(0)} \sum_{j=1}^{r} g_{ij}^{(0)}}{\sqrt{\sum_{i=1}^{N} g_{ij}^{(0)} \sum_{j=1}^{r} g_{ij}^{(0)}}} \quad (7)$$

$$q_{Yij} = \frac{\left(g_{ij}^{(0)} - \left| \sum_{i=1}^{N} g_{ij}^{(0)} \sum_{j=1}^{r} g_{ij}^{(0)} \right| - 1 \right)^2}{\sum_{i=1}^{N} g_{ij}^{(0)} + \sum_{j=1}^{r} g_{ij}^{(0)}} \quad (8)$$

Following these substitutions, the same matrix operations inherent to the re-scaling described in Equations (1)-(4), were maintained.

Constructing the significance test for kurtosis coefficient associated with the first principal component

Considering Pearson's chi-square (Equation 7) and Yates´s correction (Equation 8) applied to data, the significance test was built under the assumption defined in H_0: The first standardized coefficient of kurtosis is zero. It is worth noting that the hypothesis definition followed these considerations:

1st: Small and large rates of ξ_j, associated with the re-scaled components obtained in $G^{(2)}$, indicate that the components to be used in the score plot are more efficient at identifying outliers (Peña & Prieto, 2001);

2nd: According to Filzmoser et al. (2008), the number of outliers is associated with the size of ξ_j, so that kurtosis coefficient rates indicate a greater amount of outliers;

3rd: Proving statistically that the first standardized kurtosis coefficient is zero implies, in practical terms, elementary outlier identification, i.e. detection in a single dimension.

Finally, test statistic was computed in Equation 9 following the decision rule defined in Equation 10:

$$T = \max[\xi_j]; j = 1, ..., r \quad (9)$$

$$T > R = \frac{\lambda_1}{\sum_{j=1}^{p} \lambda_j}; j = 1, ..., r \quad (10)$$

where:

λ_j referred to the j-th eigenvalue of the sample data matrix, including outliers detected in sample data with and without chi-square correction.

Thus, the significance probability, i.e. the lower probability of rejecting the null hypothesis was computed by Equation 11:

$$p - value = \sum_{m=1} \frac{I(T \geq R)}{m}, \quad (11)$$

where:

I is an indicator function and m = 2000 Monte Carlo simulations. H_0 is rejected if p-value is lower than the significance level specified by the researcher.

Test evaluation using Monte Carlo simulation

In order to evaluate test performance, significance probabilities were calculated (Equation 11) by Monte Carlo simulations with parametric rates assumed in the simulation. Vectors

of means μ_1 and μ_2 and dimensions (r×1) were defined as Equation 12:

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ e } \mu_2 = \begin{bmatrix} 100 \\ 100 \\ \vdots \\ 100 \end{bmatrix}, \quad (12)$$

where:
r is the number of variables involved. The specification of these parametric values was taken arbitrarily, without loss of generality, since only the covariance structure is used in the estimation of the principal components, either to the original or to modified data. Covariance matrices $\sum_1$ and $\sum_2$ of r-order were also taken into consideration and defined as Equation 13:

$$\sum_1 = \begin{bmatrix} 1 & \rho^{1-2} & \cdots & \rho^{1-r} \\ \rho^{2-1} & 1 & \cdots & \rho^{2-r} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{r-1} & \rho^{r-2} & \cdots & 1 \end{bmatrix} \text{ and}$$

$$\sum_2 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \quad (13)$$

where:
$\rho = 0.5$ is the rate of the assumed correlation coefficient.

It is worth noting that the choice of AR (1) correlation structure, denoted in Σ_1 was made for the sake of parsimony. This means that among the covariance structures applied to the modeling of time series, it is the most common and simplest to apply. So that results were not influenced by correlation degree, it was decided to evaluate an intermediate case specifying ρ = 0.5.

Different sample sizes (n = 20, 50, 100 and 150), different amounts of variables (r = 5, 10, 20 and 50) and different mixture rates (γ = 0.1, 0.2 and 0.3) were employed. Liu and Zumbo (2007) recommended the use of Monte Carlo simulation technique, in which contaminated samples may be generated from mixtures of distributions. Thus, the mixture of normal distributions was generated by alternating parametric values and selecting the rate u from a continuous uniform distribution between 0 and 1, with the following configuration: if $u \leq \gamma$, then $A \sim N_p(\mu_1, \sum_1)$; if $u \leq \gamma$, then $A \sim N_p(\mu_2, \sum_2)$.

It should be noted that, in practice, the use of methods with or without chi-square correction evidenced by the significance test proposed in current study faced an identifiable issue, i.e. to estimate the proportion of contaminated samples. Since this is an empirical study guided by objectives set out at the beginning of the report, estimation of γ was not discussed here. However, for more details on inference, see Chen, Tan and Zhang (2008) and Chen and Tan (2009). Finally, a program was built in software R version 2.14.0 (R Development Core Team, 2014) to obtain results.

Results and discussion

Empirical probabilities in favor of H0 for the proposed test of significance

In accordance with current method, the results described in Table 1 for small sample sizes n = 20 and n = 50 indicate that, given the application in unanalyzed samples, increase in outlier percentage (γ) reduced probability in favor of H_0. This fact is also confirmed by increase in the number of variables (r) including situations where samples were analyzed by Yates´s correction. However, samples analyzed by Pearson's chi-square showed mixed yet promising results, as increase in rates of outlier percentage (γ) leads towards greater statistical evidence in favor of the null hypothesis.

Table 1. p-values of the proposed significance test for assessing H_0: ξ_1 = 0 considering both samples analyzed using different chi-square distances and unanalyzed samples, given the configuration between sample sizes (n = 20, 50), number of variables (r) and different mixture probabilities (γ).

n	r	γ	Uncorrected sample	Pearson´s correction	Yates´s correction
20	5	0.1	0.5225	0.6790	0.1235
		0.2	0.4355	0.7585	0.0860
		0.3	0.2340	0.7740	0.0750
	10	0.1	0.1935	0.4390	0.0015
		0.2	0.0755	0.6260	0.0010
		0.3	0.0225	0.7000	0.0005
	20	0.1	0.1880	0.1315	0.0000
		0.2	0.0715	0.3080	0.0000
		0.3	0.0695	0.2905	0.0000
	50	0.1	0.0485	0.0100	0.0000
		0.2	0.0085	0.0220	0.0000
		0.3	0.0015	0.0365	0.0000
50	5	0.1	0.8860	0.7240	0.1355
		0.2	0.5100	0.7580	0.1030
		0.3	0.2455	0.8245	0.0775
	10	0.1	0.4095	0.6405	0.0100
		0.2	0.1550	0.7310	0.0040
		0.3	0.0300	0.8275	0.0010
	20	0.1	0.0105	0.1760	0.0005
		0.2	0.0025	0.3310	0.0000
		0.3	0.0000	0.4565	0.0000
	50	0.1	0.0405	0.0000	0.0000
		0.2	0.0145	0.0265	0.0000
		0.3	0.0080	0.0445	0.0000

While keeping the same specifications described in Table 1, albeit considering larger samples (n = 100 and 150), results in Table 2 are also promising with regard to the application of Pearson's chi-square in the sample data, with the exception of situations involving r = 50, where significance probabilities were practically null.

Table 2. p-values of the proposed significance test for assessing H_0: $\xi_1 = 0$ considering both samples analyzed using different chi-square distances and unanalyzed samples, given the configuration between sample sizes (n = 100, 150), number of variables (r) and different mixture probabilities (γ).

n	r	γ	Uncorrected sample	Pearson´s correction	Yates´s correction
100	5	0.1	0.9270	0.8255	0.1515
		0.2	0.6735	0.8910	0.1030
		0.3	0.2705	0.9005	0.0925
	10	0.1	0.5640	0.7235	0.0055
		0.2	0.0620	0.7970	0.0040
		0.3	0.0090	0.8210	0.0025
	20	0.1	0.0015	0.2660	0.0000
		0.2	0.0000	0.4265	0.0000
		0.3	0.0000	0.5910	0.0000
	50	0.1	0.0000	0.0065	0.0000
		0.2	0.0000	0.0030	0.0000
		0.3	0.0000	0.1330	0.0000
150	5	0.1	0.9940	0.7205	0.0875
		0.2	0.6485	0.8685	0.0865
		0.3	0.1715	0.9045	0.1145
	10	0.1	0.5970	0.7035	0.0025
		0.2	0.2680	0.7015	0.0060
		0.3	0.0040	0.8795	0.0010
	20	0.1	0.0005	0.4115	0.0015
		0.2	0.0000	0.4505	0.0000
		0.3	0.0000	0.5245	0.0005
	50	0.1	0.0000	0.0000	0.0000
		0.2	0.0000	0.0185	0.0000
		0.3	0.0000	0.0360	0.0000

In the discussion of descriptive results (Table 1 and 2), it may be observed that, as a rule, Yates´s correction showed reduced significance probabilities due to the increase in sample size and number of variables. Result may be explained by the fact that Yates´s correction is treated as a continuity correction; consequently, it produces more conservative results. For smaller samples in particular the probability of rejecting the hypothesis H_0 is even greater when compared to other sample sizes; in fact, larger samples tend to produce smaller significance probabilities (Hubbard, 2011).

Application to sales volumes rates in the state of Minas Gerais in different market segments

Exploratory analysis

The method was performed on a data set comprising independent variables named 'retail trade indicators' in the state of Minas Gerais, Brazil, between January 2007 and December 2009, obtained from the Brazilian Institute of Geography and Statistics website, with 36 observations. Each variable was named according to segments numbered as follows: 1-Fuels and Lubricants; 2-Hypermarkets, supermarkets, food products, beverages and tobacco; 3-Hypermarkets and Supermarkets; 4-Textiles, apparel and footwear; 5-Furniture and appliances; 6-pharmaceutical, medical, orthopedic, perfumery and cosmetics; 7-books, newspapers, magazines and stationery; 8-office equipment and supplies, computer and communication; 9-Other articles of personal and domestic use. Figure 1 shows an exploratory analysis for each segment based on these variables.

Observing the time series for each segment in Figure 1, expected seasonal peaks during the period of 12 months were identified, since retail sales rise in December.

Note that in the case of this application, the identification of the observation preliminarily classified as an outlier and that corresponding to seasonal peaks did not show a behavior that featured an intervention or change of level, usually caused by the lack of observations at a given period. For such situations, it would be recommended to the researcher to infer through intervention models in univariate or multivariate approach, since, to infer the structural relationship between variables, there are chances to estimate biased models resulting in a reduction in the accuracy and prediction of forecasts. This reduction may be more attenuating when considering simultaneous observations generated by two or more processes, featuring a multivariate temporal modeling.

Application of the significance test for kurtosis coefficient associated with the first principal component and outlier identification using the principal components

Keeping the procedure described in Equations (1)-(11), 5000 resampling trials were performed and the probability of significance under the hypothesis H_0: $\xi_1 = 0$ was calculated. Results are described in Table 3.

According to the results in Table 3, there is statistical evidence that Pearson's correction in sample data suggests that the first principal component (PCA-1) is recommended for outlier detection. The other component for constructing the two-dimensional graph was determined by the coefficients of kurtosis (6) found in Table 4.

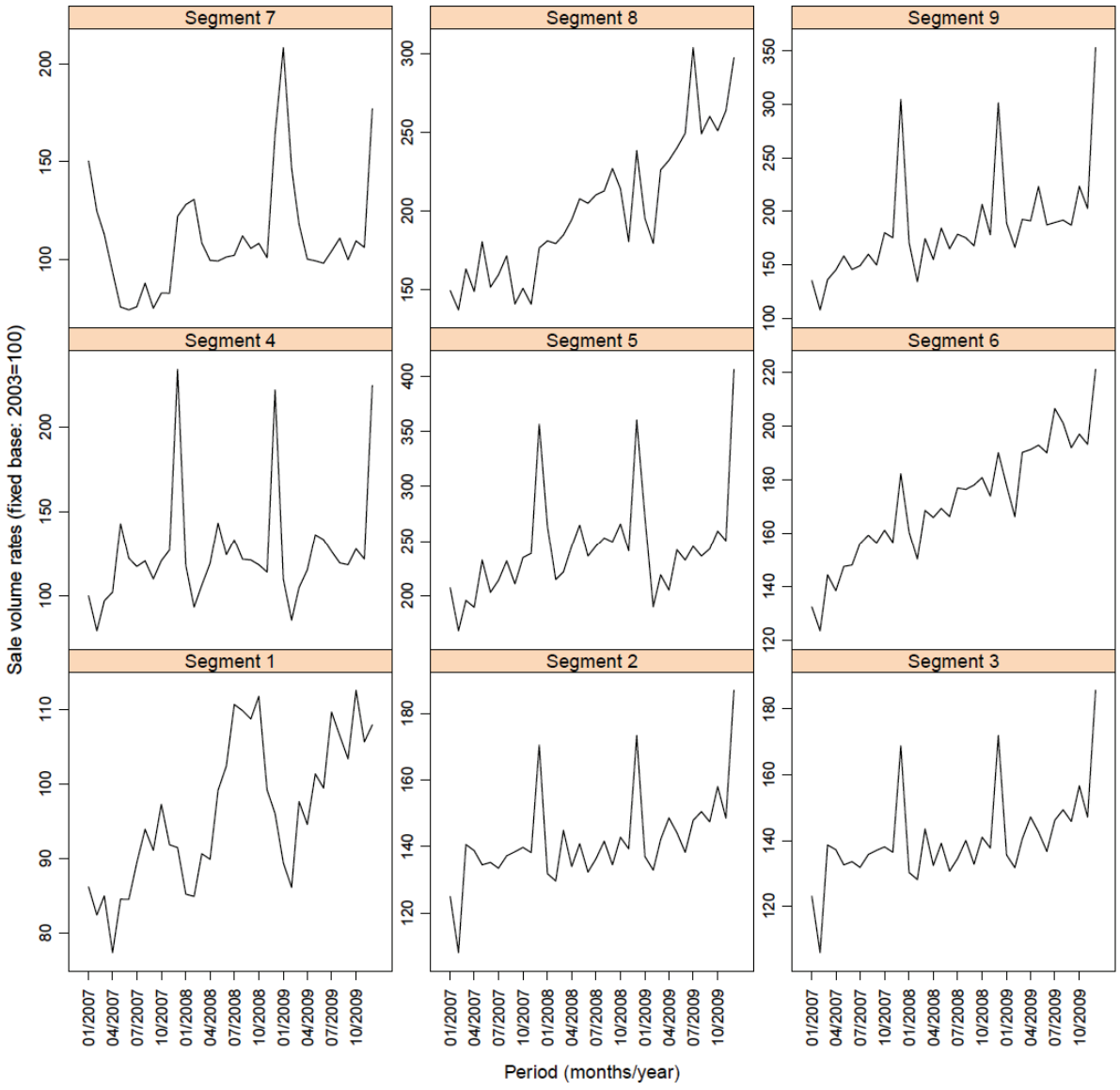

Figure 1. Plot of ratio of sales volume collected monthly for each market segment.

Table 3. p-values obtained in samples analyzed using Pearson's chi-square and Yates´s correction and in uncorrected samples.

Situation	p-value
Uncorrected	0.3814
Pearson´s correction	0.4936
Yates´s correction	0.3908

Table 4. Standardized coefficient of kurtosis (ξ_j) for j = 1,..., r = 9 calculated using sample analyzed with Pearson's chi-square.

Coefficient of kurtosis (ξ_j)	Principal Components
$\xi_1 = 0.1126$	PCA-1
$\xi_2 = 0.7052$	PCA-2
$\xi_3 = 0.0101$	PCA-3
$\xi_4 = 0.0306$	PCA-4
$\xi_5 = 0.0141$	PCA-5
$\xi_6 = 0.0112$	PCA-6
$\xi_7 = 0.0937$	PCA-7
$\xi_8 = 0.0131$	PCA-8
$\xi_9 = 0.0095$	PCA-9

Based on the results described in Table 4, the component score plot was established with components selected in response to kurtosis rates according to Peña and Prieto (2001), who state that components with higher and lower kurtosis rates are most appropriate for outlier detection. Thus, the graphs in Figure 2 show the following situations: (A) refers to score plots for the first and second principal components, considering non re-scaled data analyzed with Pearson's chi-square; (B) refers to score plots for the second and first components, considering re-scaled data also analyzed with Pearson's chi-square.

Figure 2 (A) demonstrates that observations classified as outliers were identified in December 2007 (obs.12); December 2008 (obs.24) and

December 2009 (obs. 36). The same observations were identified in Figure 2B, but including January 2009 (obs. 25), as retail sales also rise in January. Thus, December and January are special months according to the indicators evaluated in current study. Therefore, Pearson's chi-square analysis of re-scaled sample data provided more informative results.

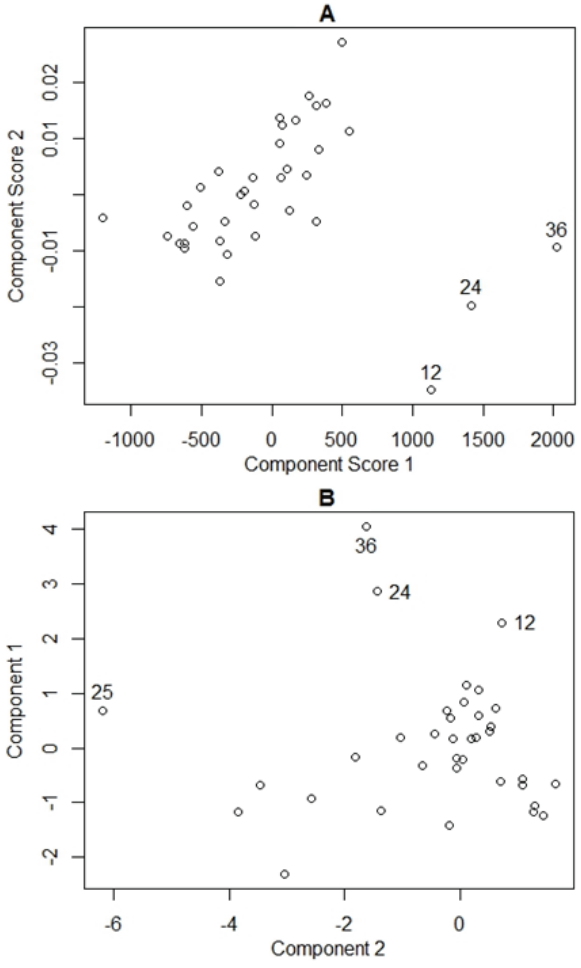

Figure 2. Outlier detection considering Pearson's chi-square analysis of data sample: A) non re-scaling, and B) re-scaling.

For confirmation and validation of the method employed, identified outliers were confronted with individual graph results of time series for each market segment, as shown in Figure 1. This comparison showed that outliers corresponded to seasonal peaks in all segments assessed in the study. It should be emphasized that the results obtained in the analysis do not identify which segment showed greater importance in retail. A study of greater impact should add a more sophisticated statistical analysis which includes external factors, for instance, the reduction of consumption among households, defaults on retail, etc.

Conclusion

Owing to the agreement between the simulation study and application, there is statistical evidence to assert that the selection of principal components using the proposed significance test as the most appropriate is the chi-square Pearson. It has been observed that in the application the use of the proposed methodology allowed the identification, by exploratory analysis, of the outliers of all segments in the retail, as seasonal peaks. In terms of the significance test performance, the effect of the number of variables resulted in a smaller reduction in the probability of significance for Pearson´s chi-square correction for all sample sizes evaluated.

Acknowledgements

The authors would like to thank CNPq and Fapemig for the funding of current research.

References

Bénasséni, J. (2005). A concentration study of principal components. *Journal of Applied Statistics, 32*(9), 947-957.

Caroni, C. (2000). Outlier detection by robust principal components analysis. *Communications Statistics Data Simulation, 29*(1), 139-151.

Caroni, C., & Billor, N. (2007). Robust detection of multiple outliers in grouped multivariate data. *Journal of Applied Statistics, 34*(10), 1241-1250.

Chan, W. S. (1992). A note on time series model specification in the presence of outliers. *Journal of Applied Statistics, 19*(1), 17-124.

Chen, J., & Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis, 100*(7), 1367-1383.

Chen, J., Tan, X., & Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statististics Sinica, 18*(2), 443-465.

Enki, D. G., Trendafilov, N. T. L, & Jolliffe, I. T. (2013). A clustering approach to interpretable principal components. *Journal of Applied Statistics, 40*(3), 583-599.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics, 34*(2), 127-138.

Filzmoser, P., Maronna, R., & Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis, 52*(3), 1694-1711.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society, 34*(3), 350-363.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics, 28*(1), 81-124.

Hubbard, R. (2011). The widespread misinterpretation of p-values as error probabilities. *Journal of Applied Statistics, 38*(11), 2617-2626.

Jackson, D. A., & Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics, 15*(2), 129-139.

Knüsel, L. (2008). Chisquare as a rotation criterion in factor analysis. *Computational Statistics and Data Analysis, 52*(9), 4243-4252.

Liu, Y., & Zumbo, B. D. (2007). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: visual analogue scales. *Educational and Psychological Measurement, 67*(4), 620-634.

Peña, D., & Prieto, F. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics, 43*(3), 286-310.

Pereira, T. M., Cirillo, M. A., & Oliveira, L. F. P. (2014). Chisquaremax rotation criterion in factor analysis: A Monte Carlo assessment of the effect of outliers. *Acta Scientiarum Technology, 26*(4), 643-649.

R Development Core Team (2014). *R: A language and environment for statistical computing*. Vienna, AT: R Foundation for Statistical Computing.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871-880.

Serneels, S., & Verdonck, J. (2008). Principal component analysis for data containing outliers and missing elements. *Computational Statistics & Data Analysis, 52*(11), 1712-1727.

Silva, A. M., Moraes, A. R., & Cirillo, M. A. (2013). Efeito de diferentes estruturas de correlação nos ângulos formados entre componentes principais e interpretáveis em amostras com presença de pontos discrepantes. *Ciência e Natura, 35*(2), 95-104.

Steiner, M. T. A., Neto, A. C., Braulio, S. N., & Alves, V. (2008). Métodos estatísticos multivariados aplicados à engenharia de avaliações. *Gestão Produção, 15*(1), 23-32.

Tsay, R. S., Peña, D., & Pankratz, A. (2000). Outliers in multivariate time series. *Biometrika, 87*(4), 789-804.

Received on December 13, 2014.
Accepted on November 23, 2015.