Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A- Uygulamalı Bilimler ve Mühendislik
Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering

2017 - Volume: 18 Number: 3
Page: 754 - 762
DOI: 10.18038/aubtda.317765
Received: 31 May 2017 Revised: 07 August 2017 Accepted: 20 August 2017

ROBUST PRINCIPAL COMPONENT ANALYSIS BASED ON FUZZY CODED DATA

B. Barış ALKAN [1, *], Sevgi GANIK [2]

[1] Department of Statistics, Faculty of Science and Literature, Sinop University, Sinop, Turkey
[2] Department of Statistics, Faculty of Science and Literature, Sinop University, Sinop, Turkey

ABSTRACT

In the presence of outliers in the dataset, the principal component analysis method, like many of the classical statistical methods, is severely affected. For this reason, if there are outliers in dataset, researchers tend to use alternative methods. Use of fuzzy and robust approaches is the leading choice among these methods. In this study, a new approach to robust fuzzy principal component analysis is proposed. This approach combines the power of both robust and fuzzy methods at the same time and collects these two approaches under the framework of principal component analysis. The performance of proposed approach called robust principal component analysis based on fuzzy coded data is examined through a set of artificial dataset that are generated by considering three different scenarios and a real dataset to observe how it is affected by the increase in sample size and changes in the rate of outliers. In light of the study's findings, it is seen that the proposed approach gives better results than the ones in the classical and robust principal component analysis in the presence of outliers in dataset.

Keywords: Fuzzy coded data, Outliers, Robust principal component analysis

1. INTRODUCTION

Classical principal component analysis (CPCA) is thought to have vital importance for many research areas since it is widely used as a dimension reduction method in high-dimensional dataset and in the initial analysis of other multivariate statistical methods or as a solution step for problems such as multicollinearity. Because CPCA which has a very popular usage in the literature is negatively affected by the presence of outliers in dataset, scientists are increasingly interested in the development of its alternatives. Croux and Haesbroeck [1] show that robust PCA (RPCA) can be easily done by calculating the eigenvalues and eigenvectors of a robust estimator of the correlation or covariance matrix. This approach works well when the number of variables is small enough. A different approach to obtaining RPCA is proposed by Croux and Ruiz-Gazen [2]. This approach is expressed as a RPCA based on a projection pursuit (PP). It is useful in situations where the number of variables is greater than the number of observations and in the analysis of high-dimensional dataset. Among other proposals for RPCA are the orthogonal PCA method developed by Maronna [3] and spherical PCA method developed by Locantore et al. [4]. Alkan et al. [5] examined if the missing value imputation methods can be used as an alternative approach to the RPCA. Alkan [6] also adapted minimum covariance determinant (MCD) method using the jacknife resampling approach and he examined the impacts of the changes resulting from this adaptation on RPCA based on MCD.

The easiest way to obtain robust principal components is to replace their robust estimates with classical estimates of location and scale parameters. Devlin et al. [7] and Campbell [8] have used M estimators of location and scale parameters. However, the fact that M estimators have low breakdown points at high dimensions have reduced the use of these estimators in recent studies. MCD method proposed by Rousseeuw [9] gives robust estimators of multivariable position and scale parameters. It yields robust results up to outlier rate of 50%. Therefore, estimation of location and scale parameters for a multivariate dataset can be done using the MCD method, which provides a high breakdown

*Corresponding Author: bbalkan@sinop.edu.tr

point. The robust version of the PCA can be obtained by substituting the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ parameters with $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ robust estimates [10].

Other approaches used in the presence of outliers in the dataset are based on fuzzy logic. Fuzzy modeling and fuzzy statistical approaches are alternative approaches that can be used if problems such as uncertainty, missing data and outliers are encountered. In recent years, many fuzzy statistical approaches such as fuzzy regression, fuzzy clustering, and fuzzy PCA have been proposed. PCA is generally applied to crisp dataset. But, Lauro and Palumbo [11], Zimmermann [12], Taheri [13], Douzal-Chouakria et al. [14], Viertl [15], Calcagni et al. [16] have extended by a number of adaptations of the CPCA method for the analysis of fuzzy, interval, and symbolic dataset in their studies. In addition to these studies, it is possible to find many studies in the literature on fuzzy PCA based on fuzzy clustering. This issue was addressed by Bezdek et al. [17], Dumitrescu et al. [18], Pop et al. [19], Sarbu and Pop [20], Yang and Wang [21], Sarbu and Pop [22], Sarbu and Pop [23], Sarbu and Pop [24].

In this study, a new approach to robust Fuzzy PCA is proposed, which combines the power of both robust and fuzzy methods at the same time and collects these two approaches under the framework of PCA. The proposed method, Robust Principal Component Analysis based Fuzzy Coded Data (RPCA-FCD), uses a robust covariance matrix based on MCD instead of the classical data covariance matrix and fuzzy coded data. For the proposed approach, we evaluate a real dataset and three artificial dataset with different outlier rates in terms of changes in the outlier rates and increasing in the sample size. According to the proposed approach, the dataset is re-coded in a fuzzy way. And then the original dataset is weighted with these obtained fuzzy codes. After this process, RPCA based on MCD method is applied to the modified version of dataset. This approach is called as RPCA-FCD in this study.

In the second part of the work, basic theoretical concepts about fuzzy coded data are mentioned. In the following chapters, basic concepts and required mathematical theory are given for CPCA and RPCA based on MCD, respectively. Methods are examined in detail in the previous sections are applied to Daudin's milk composition dataset as the real dataset and the three artificial dataset obtained in the context of the scenarios based on different outlier ratios in the applications of artificial and real dataset section of the study. Finally, the seventh section concludes this article yielding some findings and suggestions for future extensions of proposed approach.

2. FUZZY CODED DATA

Fuzzy coding was proposed by Guitonneau and Roux [25]. Guitonneau and Roux [25] conducted studies on correspondence analysis. The idea of fuzzy coding is also used in the application areas of multivariate statistical methods in the following years [26, 27]. In this study, a triangular membership function is used to obtain fuzzy coded data. This function is defined as

$$f(x;a,b,c)=\begin{cases} 0, x\le a \\ \dfrac{x-a}{b-a}, a<x\le b \\ \dfrac{c-x}{c-b}, b<x<\text{c} \\ 0, c\le x \end{cases}$$

with the parameters a, b and c correspond to the minimum, average, and maximum values of x, respectively. With this function, the dataset is converted into fuzzy codes ranging from 0 to 1. In the literature there are membership functions such as Trapezoidal, Gaussian and Cauchy which can be an alternative to triangular membership function [26, 28, 29].

3. CLASSICAL PRINCIPAL COMPONENT ANALYSIS

CPCA is a multivariate method that aims to reduce the dimension by finding k linear combination of the original variable p, k<p which allows better summarization and interpretation of the dataset. The principal components correspond to the vectors in directions that maximize the variance of the projected data on this *k* linear combination [30].

We firstly define CPCA for a data matrix, $\boldsymbol{X} = \boldsymbol{X}_{n,p} \in \mathbb{R}^{n\times p}$. The p-dimensional observations in $\boldsymbol{X}$ are shown with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The loadings of PCs are located in the columns of the estimated orthogonal *loadings matrix* $\boldsymbol{P}$. Loadings matrix $\boldsymbol{P}$ and mean $\widehat{\boldsymbol{\mu}}$, projecting the centered $\boldsymbol{X}$ on the new directions give the *scores matrix* $\boldsymbol{T} = (\boldsymbol{X} - \boldsymbol{1}_n\widehat{\boldsymbol{\mu}}')\boldsymbol{P}$, with $\boldsymbol{1}_n$, a *n×1*-dimensional column vector consisting of n ones. CPCA can be defined as a finding $\widehat{\boldsymbol{\mu}}$ and $\boldsymbol{P}$ providing that the scores have the maximum variance and unrelated. While CPCA directions correspond to the eigenvectors of the classical covariance matrix $\boldsymbol{S}$ of $\boldsymbol{X}$, the variance of the data projected on each of the eigenvectors correspond to the eigenvalues of $\boldsymbol{S}$. If the variances of the original variables show big differences, the dataset should be standardized. Usually, $k < p$ dimensions are required to describe the information in the dataset. There are several approaches to choosing the number of principal components, k. One of the simplest and most popular approaches is to use a scree plot. This chart gives decreasing eigenvalues versus their index. The number of PCs corresponds to the broken points in the graph. The number of important PCs is determined by this way. And then, the first k columns of $\boldsymbol{P}$ are used and indicated as

$\boldsymbol{P}_{p,k} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k]$ [31].

4. ROBUST PRINCIPAL COMPONENT ANALYSIS BASED ON MCD

As the CPCA is based on the sample covariance matrix, observations that take place in the dataset and move very far in the general structure of the dataset may lead to totally biased and unreliable results. Even a single outlier can disrupt an entire process. In the event of disrupt, the first principal component with the greatest variance explanatory rate changes direction towards outliers. This may cause a more swollen variability than it is actually exist. In other words, it can lead to over-optimistic eigenvalues and a high total variance explained proportions that cannot actually exist. These problems can be overcome by using robust methods for PCA [32].

MCD method is used for finding robust estimate of covariance matrix when the number of variables, p is less than the number of observations, n [9, 33, 34]. This method is very popular due to its high degree of robustness against outliers and it is also the fastest algorithm developed recently in terms of computation [35].We consider the h-dimensional sub-clusters of the entire dataset consisting of n observations to describe the MCD estimator. h-value determines the robustness of the estimator and at least [((n + p + 1)) / 2] should be taken as a lower bound. The MCD estimator tries to find the optimal h-subset which have minimum covariance determinant of these subclusters. The estimate of location parameter $\widehat{\boldsymbol{\mu}}_{\boldsymbol{MCD}}$ is given by the mean of the optimal h-subset, and the estimate of scale parameter $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{MCD}}$ is given by its covariance matrix. The MCD estimator has $(n - h + 1)/n$ breakdown point value [34].

5. ROBUST PRINCIPAL COMPONENT ANALYSIS BASED FUZZY CODED DATA

RPCA-FCD is described as robust PCA based on MCD of new dataset weighted with fuzzy coded data obtained by using the triangular membership function. The proposed algorithm in this context is given below.

Proposed RPCA-FCD Algorithm

Step1. With the aid of the triangular membership function, the dataset is re-coded

Step 2. The original dataset is weighted with the fuzzy codes obtained by re-encoding.

Step 3. For weighted new dataset,

Step 3.1. Find the combination (n, h).

Step 3.2. For sub-clusters with each h example,

Step 3.2.1. Calculate the sample covariance matrix

Step 3.2.2. Calculate sample covariance matrix determinant

Step 4. Select the sub-cluster with the smallest determinant. Find the sample mean vector and sample covariance matrix of this subset.

From this algorithm, $\widehat{\boldsymbol{\mu}}_{MCD}$ and $\widehat{\boldsymbol{\Sigma}}_{MCD}$ multivariate location and scale estimates are obtained. The approach based on these estimates is called as RPCA-FCD.

6. APPLICATIONS OF ARTIFICIAL AND REAL DATASET

In this study, Daudin's milk composition dataset was used as a real dataset to compare the performance of CPCA, RPCA based on MCD and our proposed RPCA-FCD in the presence of outliers in the dataset. Also, these methods are examined on three artificial dataset obtained in the direction of scenarios created according to rates of different outliers. In artificial dataset, we examine the actions of the methods in a low dimension (p=5, n=30, 100) for two scenarios and in a high dimension (p=30, n=1000) situation for one scenario. The robustbase, rrcov and rrcovHD libraries included in the R statistical software were used for analysis [36, 37, 38]. A program written in R is used to evaluate for proposed RPCA-FCD. For the evaluation of proposed RPCA-FCD approach, we used the program that we wrote in R.

Artificial dataset 1- low dimensional (p=5, n=30, 100) *situation, scenario1*

In artificial dataset 1 we examine the actions of the methods in a low dimensional (p=5, n=30, 100) situation. In dataset 1, we generated data points from the p-variate multivariate normal distribution $N_5([5\ 10\ 7\ 2\ 1], \text{diag}[100\ 20\ 8\ 4\ 2])$ and outliers in different proportions (10%, 20%, 30%) were generated from multivariate t distribution with 5 degrees of freedom (T_5).

Table 1. Artificial dataset 1- low dimensional (p=5, n=30, 100) situation, scenario 1

		%10 Outliers			**%20 Outliers**			**%30 Outliers**		
		PC1	**PC2**	**CVEP***	**PC1**	**PC2**	**CVEP***	**PC1**	**PC2**	**CVEP***
n=30	**CPCA**	0.3706	0.2934	**0.6640**	0.4330	0.24227	**0.6753**	0.4577	0.1974	**0.6552**
	RPCA	0.5656	0.2919	**0.8575**	0.7085	0.2133	**0.9217**	0.5792	0.2801	**0.8592**
	RPCA-FCD	0.4346	0.4077	**0.8423**	0.8708	0.07149	**0.9422**	0.6916	0.2212	**0.9128**
n=100	**CPCA**	0.3337	0.2203	**0.5541**	0.3616	0.21496	**0.5766**	0.4587	0.1861	**0.6449**
	RPCA	0.7640	0.1495	**0.9135**	0.6612	0.2327	**0.8938**	0.7996	0.1041	**0.9036**
	RPCA-FCD	0.5970	0.3490	**0.9460**	0.7729	0.1797	**0.9527**	0.7452	0.1857	**0.9309**

* Cumulative Variance Explained Proportion

When the artificial dataset 1 is analyzed by CPCA, RPCA and RFPCA methods, the findings are presented in Table 1. According to Table 1, in the presence of 10% outliers and in the case of n = 30, according to the cumulative total variance explained proportion of the first two principal components, RPCA-FCD shows no improvement when compared to RPCA. However, if the outlier rates increase to 20% and 30%, it is shown that the proposed approach RPCA-FCD performs better than RPCA. In this case, if the ratio of outliers is approximately 10%, RPCA method can be useful. But, if ratio of outliers is about 20% and 30%, RPCA-FCD should be preferred. In the case of n = 100, for 10%, 20% and 30% outlier ratios, RPCA-FCD performed well. As a result, when the outliers ratio is more than 20%, the use of RPCA-FCD is recommended.

Artificial dataset 2- low dimensional (p=5, n=30, 100) situation, scenario 2

In artificial dataset 2 we examine the actions of the methods in a low dimensional (p=5, n=30,100) situation. We generated data points from the 5-variate multivariate standard normal distribution $N_5(\mathbf{O},\mathbf{I})$, and outliers in the different proportions (10%, 20%, 30%) were generated from multivariate t distribution with 5 degrees of freedom (T_5).

Table 2. Artificial dataset 2- low dimensional (p=5, n=30, 100) situation, scenario 2

		%10 Outliers			**%20 Outliers**			**%30 Outliers**		
		PC1	**PC2**	**CVEP***	**PC1**	**PC2**	**CVEP***	**PC1**	**PC2**	**CVEP***
n=30	**CPCA**	0.3368	0.2837	**0.6206**	0.2735	0.2453	**0.5188**	0.3359	0.2233	**0.5593**
	RPCA	0.4415	0.3318	**0.7733**	0.3226	0.2409	**0.5635**	0.4143	0.2311	**0.6455**
	RPCA-FCD	0.4133	0.2793	**0.6927**	0.4327	0.2495	**0.6822**	0.3537	0.2990	**0.6527**
n=100	**CPCA**	0.2548	0.2358	**0.4907**	0.2600	0.2165	**0.4766**	0.2828	0.2292	**0.5120**
	RPCA	0.2891	0.2224	**0.5115**	0.3121	0.2412	**0.5533**	0.3268	0.2458	**0.5726**
	RPCA-FCD	0.3459	0.2564	**0.6022**	0.3492	0.2905	**0.6397**	0.3611	0.2404	**0.6014**

* Cumulative Variance Explained Proportion

When the artificial dataset 2 is analyzed by CPCA, RPCA and RPCA-FCD methods, the findings given in Table 2 are obtained. It is seen that the results obtained from Table 2 strongly support the results presented in Table 1 obtained from artificial dataset 1. According to Table 2, in the presence of 10% outliers, in the case of n = 30, according to the cumulative total variance explained proportion of the first two dimensions, RPCA-FCD shows no improvement when compared to RPCA. However, if the outlier rates increase to 20% and 30%, it is shown that the proposed approach RPCA-FCD performs better than RPCA. In this case, if the ratio of outlier observations is approximately 10%, RPCA method can be useful. But, if ratio of outliers is about 20% and 30%, RPCA-FCD should be preferred. In the case of n = 100, for 10%, 20% and 30% outlier ratios, RPCA-FCD performed well.

Artificial dataset 3 -High dimensional (n=1000, p=30) situation, scenario 3

In artificial dataset 3 we examine the actions of the methods in a high dimensional (p=30, n=1000) situation. We generated data points from the p-variate multivariate standard normal distribution $N_{30}(\mathbf{O},\mathbf{I})$, and outliers in the proportion of 30% were generated from multivariate t distribution with 5 degrees of freedom (T_5).

When the artificial dataset 3 is analyzed by CPCA, RPCA and RPCA-FCD methods, the findings given in Table 3 were obtained. When the findings presented in Table 3 are examined, it is seen that the proposed RPCA-FCD approach works well in the analysis of high-dimensional dataset.

Table 3. Artificial dataset 3 -High dimensional (n=1000, p=30) situation, scenario 3

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**CVEP** *
CPCA	0.0509	0.0442	0.04400	0.04262	0.04153	0.04082	0.03943	0.03881	0.03740	0.0366	**0.4164**
RPCA	0.0471	0.04543	0.04426	0.04262	0.04188	0.04033	0.03966	0.03807	0.03775	0.0366	**0.4138**
RPCA -FCD	0.0504	0.04778	0.04497	0.04483	0.04225	0.04021	0.03991	0.03933	0.03791	0.0367	**0.4243**

* Cumulative Variance Explained Proportion

Real Dataset- Daudin's milk composition dataset (p=8, n=86), 20 % Outliers

In the study, for demonstrating the functionality of the RPCA-FCD, Daudin et al. (1988)’s milk composition dataset are selected for real data application. A number of investigators such as Todorov et al. [39], Atkinson [40], Rock and Woodruff [41] have used Daudin's milk composition dataset as a sample dataset for comparison with classical methods to validate proposed approaches in literature for detecting outliers and examining robust statistical inferences. For this reason, this dataset is selected for application in our study. Adjusted quantile method is used to determine the multivariate outliers in the dataset. The adjusted quantile compares the difference between the distribution function of the chi-square distribution and the empirical distribution of the quadratic robust distance [42]. With this method, 18 observations (20%) are found as outliers. Daudin's milk composition dataset is analyzed by CPCA, RPCA based on MCD and proposed RPCA-FCD respectively and obtained results are presented in Table 4. It is seen that CPCA explained 77.12% of the total variance with the first two major components. However, it would not make sense to use CPCA as a criterion, since there are 18 outliers (20%) in the dataset, and in the presence of outliers, the CPCA may swell in variance explanatory ratios and may change the direction of the first major component. RPCA method explains 87.13% of the total variance with the first two major components. This method is a robust method if there are observations that are outliers in dataset. When compared to CPCA, it gives us a better approach. Another result given in Table 4 is that the total variance explanation ratio obtained by proposed RPCA-FCD has the maximum total variance explanation ratio with 92.45%. In this case, if the results in Table 4 are examined. it can be seen that the use of the proposed RPCA-FCD approach instead of CPCA and RPCA methods in the presence of outliers in the dataset is more appropriate. It is seen that this result also supports the results of different artificial dataset analysis, which are covered by three different scenarios.

Table 4. Real Data- (p=8, n=86) situation, *20 % Outliers*

	PC1	**PC2**	**CVEP***
CPCA	0.6306	0.1406	**0.7712**
RPCA	0.5929	0.2785	**0.8713**
RPCA-FCD	0.8265	0.0980	**0.9245**

*Cumulative Variance Explained Proportion

7. CONCLUSIONS

In this study, we have proposed RPCA-FCD approach, which improves the results of robust PCA used in the presence of outliers in multivariate dataset. In comparison with CPCA and RPCA based on MCD, the proposed RPCA-FCD approach is more robust in the presence of outliers, for the analysis of both low dimensional and high dimensional dataset. The efficiency of the proposed approach was examined on one real dataset and three different scenarios having artificial dataset. Examining the positive or negative effects of the use of different membership functions on the proposed approach in the fuzzy coding process can be considered as a future study.

REFERENCES

[1] Croux C, Haesbroeck G. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, Biometrika, 2000, 87, 603–618.

[2] Croux C, Ruiz-Gazen A. High breakdown estimators for principal components: the projection-pursuit approach revisited, Journal of Multivariate Analysis 2005, 95, 206–226.

[3] Maronna R. Principal components and orthogonal regression based on robust scales,Technometrics, 2005, 47(3), 264-273.

[4] Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K. Robust principal components for functional data, 1999, Test 8, 1–28.

[5] Alkan BB, Atakan C, Alkan N. A comparison of different procedures for principal component analysis in the presence of outliers, Journal of Applied Statistics, 2015, 42(8), 1716-1722.

[6] Alkan BB. Robust Principal Component Analysis Based on Modified Minimum Covariance Determinant In The Presence of Outliers (in Turkish). Alphanumeric Journal, 2016, 4(2).

[7] Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation of dispersion matrices and principal components, Journal of the American Statistical Association, 1981, 76(374), 354-362.

[8] Campbell NA. Robust procedures in multivariate analysis I: Robust covariance estimation, Applied statistics, 1980, 231-237.

[9] Rousseeuw PJ. Least median of squares regression, Journal of the American statistical association, 1984. 79(388), 871-880.

[10] Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis, J. Statist Softw, 2009, 32(3) 1–47.

[11] Lauro CN, Palumbo F. Principal component analysis of interval data: a symbolic data analysis approach. Comput Stat, 2000, 15(1):73–87.

[12] Zimmermann HJ. Fuzzy set theory-and its applications. Springer, 2001.

[13] Taheri SM. Trends in fuzzy statistics. Austrian J Stat, 2003, 32(3):239– 257.

[14] Douzal-Chouakria A, Billard L, Diday E. Principal component analysis for interval-valued observations. Stat Anal Data Min, 2011, 4(2):229–246

[15] Viertl R. 2011 Statistical methods for fuzzy data. Wiley

[16] Calcagnì A, Lombardi L & Pascali E. A dimension reduction technique for two-mode non-convex fuzzy data. Soft Computing, 2016, 20(2), 749-762.

[17] Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984, 10(2-3), 191-203.

[18] Dumitrescu D, Sârbu C, Pop H. A fuzzy divisive hierarchical clustering algorithm for the optimal choice of sets of solvent systems. Analytical letters, 1994, 27(5), 1031-1054.

[19] Pop HF, Sârbu C, Horowitz O, Dumitrescu D. A Fuzzy Classification of the Chemical Elements. Journal of chemical information and computer sciences, 1996, 36(3), 465-482.

[20] Sârbu C, Pop HF. Fuzzy clustering analysis of the first 10 MEIC chemicals. Chemosphere, 2000, 40(5), 513-520.

[21] Yang TN, Wang SD. Robust algorithms for principal component analysis. Pattern Recognition Letters, 1999, 20(9), 927-933.

[22] Sârbu C, Pop HF. Fuzzy robust estimation of central location. Talanta, 2001; 54(1), 125-130.

[23] Sarbu C, Pop HF. Fuzzy soft-computing methods and their applications in chemistry. Reviews in Computational Chemistry, 2004, 20, 249.

[24] Sarbu C, Pop HF. Principal component analysis versus fuzzy principal component analysis: a case study: the quality of Danube water (1985–1996). Talanta, 2005, 65(5), 1215-1220.

[25] Guitonneau GG, Roux M. Sur la taxinomie du genre Erodium. Les cahiers de l'analyse des données, 1977, 2(1), 97-113.

[26] Asan Z, Greenacre M. Biplots of fuzzy coded data. Fuzzy sets and Systems, 2011, 183(1), 57-71.

[27] Asan Z, Senturk S. An Application of Fuzzy Coding in Multiple Correspondence Analysis for Transforming Data from Continuous to Categorical. Journal of Multiple-Valued Logic & Soft Computing, 2011, 17.

[28] Jang JSR, Sun CT, Mizutani E, 1997. Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence.

[29] Smithson M, Verkuilen J. Fuzzy set theory: Applications in the social sciences, 2006, (No. 147). Sage.

[30] Croux C, Filzmoser P, Fritz H. Robust sparse principal component analysis. Technometrics, 2013, 55(2), 202-214.

[31] Hubert M, Reynkens T, Schmitt E, Verdonck T. Sparse PCA for high-dimensional data with outliers. Technometrics, 2016, 58(4), 424-434.

[32] Farcomeni A, Greco L. Robust methods for data reduction. CRC press, 2015.

[33] Rousseeuw PJ. Multivariate estimation with high breakdown point, Mathematical statistics and applications, 1985, 8, 283-297.

[34] Hubert M, Engelen S. Robust PCA and classification in biosciences, Bioinformatics, 2004, 20(11), 1728-1736.

[35] Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator, Technometrics, 1999, 41(3), 212-223.

[36] R Development Core Team, 2011. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna.

[37] Rousseeuw PJ, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Maechler M., 2009. Robustbase: basic robust statistics, R package version 0.4–5. Available at http://CRAN. R-project. org/package = robustbase.

[38] Todorov V, 2009. rrcov: Scalable Robust Estimators with High Breakdown Point, R package version 0.5–03, Availableat http://CRAN. R-project. org/package = rrcov.

[39] Todorov V, Neyko N, Neytchev P. Stability of High Breakdown Point Robust PCA, in Short Communications, COMPSTAT'94; Physica Verlag, Heidelberg, 1994.

[40] Atkinson AC. Fast very robust methods for the detection of multiple outliers, J Amer Statist Assoc, 1994; 89, 1329–1339.

[41] Rocke DM, Woodruff DL. Identification of Outliers in Multivariate Data, J Amer Statist Assoc 1996; 91 (435), 1047–1061.

[42] Filzmoser P, Reimann C, Garrett RG. Multivariate outlier detection in exploration geochemistry, Technical ReportTS, 2003; 03–5, Department of Statistics, Vienna University of Technology, Austria.