International Journal of Sciences:
Basic and Applied Research
(IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

International Journal of
Sciences:
Basic and Applied
Research
ISSN 2307-4531
(Print & Online)
Published by:

A Proposal for ROBPCA Algorithm

Hasan BULUT[a]*, Yüksel ÖNER[b], Çağlar SÖZEN[c]

[a,b] Department of Statistics, University of Ondokuz Mayıs, Samsun, Turkey.
[c] Department of Banking and Finance, University of Giresun, Giresun, Turkey.

[a]*Email: hasan.bulut@omu.edu.tr*
[b]*Email: yoner@omu.edu.tr*
[c]*Email: caglar.sozen@giresun.edu.tr*

Abstract

Principal component Analysis (PCA) is one of the most frequently used multivariate statistical methods. Especially, it is used on the purpose of dimension reduction and obtaining uncorrelated variables. However, classic PCA (CPCA) is sensitive to outlier. Because it is based on classic covariance or correlation matrices influenced by outliers. Therefore, CPCA can give fallacious results in data sets which have outliers. In this study, the robust PCA (RPCA) methods to solve this problem of CPCA are introduced in literature. Moreover, we bring forward a proposal to ROBPCA algorithm which is one of these methods.

Keywords: ROBPCA; Robust Principal Component Analysis; Standardization; High Dimensional Data.

1. Introduction

Principal Component Analysis (PCA) is a statistical method reducing dimension of correlated variables. The new variables are called as components and these are unrelated. The components are linear combinations of original variables. The aim of PCA is dimension reduction and/or obtaining the unrelated variables.

It is known that both the standardized variables and the original variables can be used in PCA. It is benefited from covariance matrix when original data matrix is used in analysis, while the correlation matrix should be

----

* Corresponding author.

E-mail address:hasan.bulut@omu.edu.tr

employed when standardized data matrix is used. These cases might give strongly different results. Measure unit is the most important criterion on the selecting the matrix type. If the measure units and variances of the variables are close enough, covariance matrix is used; otherwise correlation matrix is used. If PCA is practiced to the data set that there are variables with big variance, variables which have big variances have greater weights than other variables on principal components. In that diagonal elements of S matrix are variances. However, the diagonal elements of R are 1 and equal. So, all variables have equal weights on principal components [1, 2, 3, 4, 5, 6].

In classical PCA, components are related to eigenvalues and eigenvectors of empirical covariance/correlation matrix. The first component corresponds to the direction in which the projected observations have the maximum variance. Then, the second component is orthogonal to first component and maximizes the variance of the data points projected on it. Continuing in this way produces all of the principal components. But, both classical variances which is being maximizes and classical covariance/correlation matrix which is being decomposed are very sensitive outliers. So, if there are outliers in data, classical PCA results are unreliable. Consequently, if the data set has outlier, robust principal component analysis methods must be preferred, not classical PCA. In that the aim of robust PCA is to find principal components which are not affected by outliers [7].

Fundamentally, there are two approaches in the robust PCA. Firstly, any robust covariance or correlation matrix is taken instead of classical matrix. Because these robust matrices are not sensitive to outliers, the results are reliable. In this approach, however, the data set should be low dimensional. In that the calculating of robust covariance estimation is restricted with the low dimensional data. The aim of solving this problem, the second approach is submitted. In this approach, firstly k directions are found and these directions are orthogonal each other as CPCA. While these directions are founding, a robust scatter measure called as projection index is maximized, for example MAD (Median Absolute Deviations: median of the absolute deviations from the median). Thus, the data is projected to low dimensional space.

The ROBPCA method combines two approaches. ROBPCA has a complex algorithm. Firstly, the dimension is reduced by being used second approach and low dimensional data is obtained. Then, principal components are computed by basing on first approach. In this method, after eigenvalue-eigenvector of covariance matrix is found without requiring calculating of scatter matrix, the robust covariance matrix is computed by using spectral decomposed [7].

In this paper, a deficiency of ROBPCA method is introduced and a suggestion is submitted to solve this deficiency. If you view to ROBPCA algorithm, you see that firstly eigenvalues-eigenvectors are estimated robustly and then robust covariance matrix is obtained by using the spectral decomposed. That is, these eigenvectors-eigenvalues pertain to covariance matrix. As stated previously, if the measure units of variables are not close enough, the analysis should be done by using correlation matrix and eigenvalues-eigenvectors of this matrix. Because the ROBPCA method does not take in consideration this problem, it has a deficiency. In this study, an approach is suggested to solve this problem and the idea is supported with simulation study.

2. Methods

2.1. Classical Principal Component Analysis (CPCA)

Generally, the aim of PCA is expressed as follows;

- The dimension reduction
- The obtaining uncorrelated variables
- The data preparing to other statistical methods.

In PCA, it was mentioned how covariance or correlation matrix is used in Section 1. In this section PCA based on correlation matrix is introduced. The multivariate data set is standardized denoted as [5];

$$\boldsymbol{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} = \left(\boldsymbol{\Sigma}^{\frac{1}{2}}\right)^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{1}$$

Then, the covariance matrix of this standard random variable vector is the correlation matrix original variables as follows;

$$Cov(\boldsymbol{Z}) = Corr(\boldsymbol{X}) = \boldsymbol{R}. \tag{2}$$

The principal components are obtained from this matrix $\boldsymbol{R}$. If the eigenvalue-eigenvector pair of the $\boldsymbol{R}$ is (λ_j, e_j), j^{th} principal component is expressed as follows;

$$Y_j = e_j' Z = e_{1j}Z_1 + e_{2j}Z_2 + \cdots + e_{pj}Z_p \ , j = 1,2, \ldots, p \tag{3}$$

and $(Y_j) = \lambda_j$, $Cov(Y_j, Y_k) = 0$. Moreover, the total variance of system is denoted as

$$\sigma_{Tot}^2 = tr(\boldsymbol{R}) = \lambda_1 + \lambda_2 + \cdots \lambda_p = p. \tag{4}$$

In PCA, it is interested in the variance explanations rate and principal component scores. The variance explanations rate of j^{th} variable is calculated as;

$$VER = \frac{Var(Y_j)}{\sigma_{Tot}^2} = \frac{\lambda_j}{p} \tag{5}$$

The principal component scores of i^{th} unit $(i = 1,2, \ldots, n)$ are calculated by replacing standard variable values of it in Equation (3) [6].

2.2. Robust Approaches to Principal Component Analyses

2.2.1. Principal Component Analysis Based on Robust Covariance Matrix

The aim of Robust PCA is to obtain components which are insensitive to outlier. The best popular method for this aim is to replace any robust covariance matrix instead of classical covariance matrix and to obtain components based on eigenvalue-eigenvector pairs of this matrix or correlation matrix produced by using this matrix [8].

Let's be $\widehat{\boldsymbol{\mu}}_r, \widehat{\boldsymbol{\Sigma}}_r$ robust location and scatter estimation. If the measure units of the variables are close enough, principal components are obtained by basing on the eigenvectors of covariance matrix; otherwise, eigenvectors of correlation matrix obtained with Equation (6).

$$\boldsymbol{R}_r = \left(\boldsymbol{D}_r{}^{1/2}\right)^{-1} \; \widehat{\boldsymbol{\Sigma}}_r \left(\boldsymbol{D}_r{}^{1/2}\right)^{-1} \tag{6}$$

where $D_r{}^{1/2}$ is diagonal matrix and diagonal members of it are square root of diagonal members of $\hat{\Sigma}_r$. Principal components found with this method are not impressed by outlier.

The most popular robust multivariate parameter estimator is Minimum Covariance Determinant (MCD). The MCD estimator for location parameter of multivariate data was defined as by Rousseeuw [9];

$$\begin{gathered} T(X) = mean\; of\; the\; h\; point\; of\; X\; for\; which\; the\; determinant \\ of\; the\; covariance\; matrix\; is\; minimal \end{gathered} \tag{7}$$

The covariance matrix of this subset is the MCD estimator of scatter parameter.

In this method, α is trimming ratio and $h = (1 - \alpha) * n$. Moreover, the breakdown point of MCD estimators is equal to trimming ratio (α). In this case, when $h = 0{,}5 * n$, the estimator has maximum breakdown point, %50. However, to balance between robustness and efficiency, generally, h is defined as $h = 0{,}75 * n$ and breakdown point is % 25 [10].

On the other hand, if one is certain that the fraction of outliers is at most α, (where 0< α <0,5), one can work with the estimators MCD obtained by replacing h by, $k = \lfloor n(1 - \alpha) \rfloor + 1$ in Equation (7) [9].

Rousseeuw and Van Driessen were suggested a fast algorithm called as FAST-MCD to obtain MCD estimator [10].

Principal components obtained in this way are not influenced by outliers. However, the data should be low dimensional $(n > p)$ for this aproach, as CPCA. Otherwise, the determinant of covariance matrix is zero. Therefore, this approach is restricted with low dimensional data.

2.2.2. Projection Pursuit (PP) Approach

The Projection Pursuit (PP) algorithm is another approach which ensures robustness in PCA and does not need robust covariance estimation. PP algorithm is based upon projecting lower dimension multivariate data. The algorithm finds ways by maximizing a gradient called as Projection index [11]. Huber showed that CPCA is a special case of PP algorithm and it uses variance as projection index [12]. Li and Cheng used this method to make robust PCA by taking a robust scale estimator [13]. Croux and Ruiz-Grazen investigated robustness properties of this method and suggested a fast calculation algorithm [14].

The advantages of PP algorithm are below that:

- As previously mentioned, robust estimations calculating is easier and faster in low dimension. However, projection ways again finding is time consuming after estimations are obtained.
- Robust covariance estimation is restricted with low dimensional data. But PP algorithm can be used high dimensional data sets.
- The projection ways searching is a sequential process. Thus user can determine exactly way number and the eigen analysis of covariance matrix is not difficult. Especially, the calculating time reduces signally in high dimensional data sets [15].

2.2.3. ROBPCA Approach

The ROBPCA method uses together both robust covariance matrix estimation and projection pursuit approaches and it was suggested in 2005. Firstly, PP method is used in dimension reduction. Then, the principal components are obtained by basing on MCD estimations from low dimensional data which is obtained first step. The combined approach gives results more quickly than PP algorithm [7].

Let's be $\boldsymbol{X}: nxp$ original data matrix where n is the number of observation and p is the number of original variable. The ROBPCA is practiced in three steps. Firstly, the data is reduced to a subspace which has maximum (n-1) dimension. Then, $\boldsymbol{\Sigma}_0$ the inception covariance matrix is obtained and k the number of components is determined. While it is determined, it is careful that the subspace with k dimension adapts well to data. Finally, the data points are projected on this subspace where their location vector and k-nonzero eigenvectors $(\lambda_1, \lambda_2, \dots, \lambda_k)$ are calculated and the scatter matrix are robustly estimated. The eigenvectors related to these eigenvalues give k robust principal components [16].

In the original space with p dimension, these k components indicate k dimensional space. $\boldsymbol{P}_{pxk}$ eigenvector matrix is obtained by collecting k eigenvectors. $\widehat{\boldsymbol{\mu}}_{px1}$ vector is called as robust location. The score matrix is given below as:

$$\boldsymbol{T}_{n,k} = \left(\boldsymbol{X}_{n,p} - \boldsymbol{1}_n\, \widehat{\boldsymbol{\mu}}'\right)\boldsymbol{P}_{p,k} \tag{8}$$

where 1_n is vector consisted of ones. The robust scatter matrix is calculated as:

$$\boldsymbol{\Sigma}_{p,p} = \boldsymbol{P}_{p,k}\boldsymbol{L}_{k,k}\boldsymbol{P}'_{k,p} \tag{9}$$

where $\boldsymbol{L}_{k,k}$ is diagonal matrix which the diagonal elements are $\lambda_1, \lambda_2, \dots, \lambda_k$ [7].

The ROBPCA has orthogonal equivariance property as CPCA. That is, when an orthogonal transformation is practiced to data, principal component scores do not change.

Other advantage of ROBPCA is to give diagnostic plot. The diagnostic plot gives information about observations in data sets. In Figure 1 (a), the PCA is performed to three dimensional data and the principal components space is obtained with two components. The black points in figure are called as regular observations. Because 1 and 2 observations are akin to PCA space but their projections to PCA space are far regular observations, they are called as good leverage points. As 3 and 4 observations are far to PCA space but their projections to PCA space are inside of regular observations, they are called as bad leverage points. Finally, because both differences of 5 and 6 observations to PCA space are high and their projections to PCA space are far to regular points, they are called as outliers. The diagnostic plot categorizes observations and the diagnostic plot of this data is given Figure1 (b) [17].

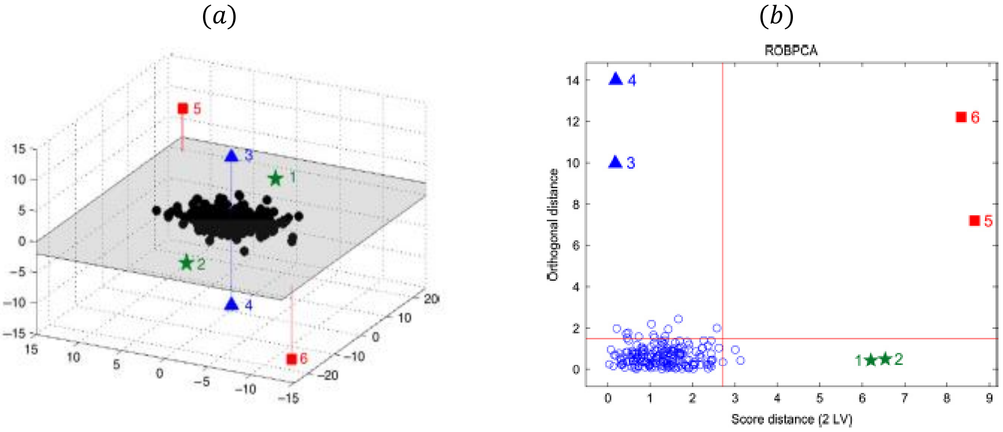

Figure 1: (a) Projecting and Dimension Reduction; (b) The Diagnostic Plot [17]

3. The Simulation Study

When the ROBPCA algorithm is investigated, it is seen that firstly the eigenvalues-eigenvectors are estimated robustly and then the robust covariance matrix is obtained by using spectral decomposition. That is, eigenvalues-eigenvectors are connected with the covariance matrix. So, the PCA is practiced on original data. But it is known that principal components should be obtained from standardized data if the units of measures and variances of variables are very different. It is obvious that the eigenvalues-eigenvectors of correlation matrix should be used in this case. Because the ROBPCA uses original data, it is unpractical when the units of measures and variances of variables are very different. In that there are rather different among units of measure in multivariate real data sets, generally. To solve this problem, two approaches can be suggested.

The Suggestion 1: The robust correlation matrix is calculated by using finally obtained robust covariance matrix and principal components are again got by basing eigenvalues-eigenvectors of this correlation matrix.

The disadvantaged of Suggestion 1: While the advantages of PP algorithm are introduced, it is mentioned that there is not time consuming because firstly eigenvalues-eigenvectors are estimated robustly. This suggestion to the ROBPCA algorithm which uses PP algorithm in the first stage causes losing mentioned advantage and time consuming.

Suggestion 2: As it is mentioned in the classical approach, the covariance matrix of standardized data set is equal to the correlation matrix of original data set. Therefore, if ROBPCA algorithm is started standardized data, finally obtained covariance matrix is correlation matrix and discrepancies among units of measure of variables do not negative effect on results. But a question is needed for this suggestion: “Which approach should we use in standardizing: classic or robust?”

In classical approach, standardizing is denoted as Equation (10) [1];

$$\frac{X-\mu}{\sigma} \tag{10}$$

and for robust approach it is denoted as Equation (11) ;

$$\frac{X-Med(X)}{MAD(X)} \tag{11}$$

where Median (Med) and median absolute deviation (MAD) are robust alternatives of sample mean and standard deviation, respectively. And these are calculated as Equation (12) [18] and Equation (13) [19], respectively;

$$Med(x)=\begin{cases} x_{\left(\frac{n+1}{n}\right)} \;, n\ uneven \\ \dfrac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}, n\ even \end{cases} \tag{12}$$

$$MAD(x) = Med\{|x - Med(x)|\}. \quad (13)$$

The simulation study is done to determine which standardizing give better results. In this study, the data sets are generated from multivariate normal and contaminated multivariate normal distributions. The general equation used to generate data sets is given Equation (14).

$$(1-\varepsilon)\, N_p(\boldsymbol{\mu},\boldsymbol{\Sigma}) + \varepsilon N_p\big(\,\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}}\big). \quad (14)$$

The parameters are below.

When n>p (For Low Dimensional Data)

$$\mu = \mathbf{0}_{5x1}\,, \boldsymbol{\Sigma} = Diag[16,8,4,2,1] \quad ; \quad \widetilde{\boldsymbol{\mu}} = [0,0,0,0,15]_{5x1}, \widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}x\left(\frac{1}{15}\right)$$

When n<p (For High Dimensional Data)

$$\boldsymbol{\mu} = \mathbf{0}_{75x1}\,, \boldsymbol{\Sigma} = Diag[100,80,75,\dots,0.15,0.10,0.05]\,; \,\widetilde{\boldsymbol{\mu}} = [0,\dots,60,0,\dots,60]_{75x1}, \widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}x\left(\frac{1}{15}\right)$$

In Table 1, for $(p = 5, n = 20, \varepsilon = 0)$, while average Explained Variance Ratio (EVR) of ROBPCA is %83.61 in classical standardized (Cstd) data, average Explained Variance Ratio (EVR) of ROBPCA is %84.57 in robustly standardized (Rstd) data. On the other hand, for $(p = 5, n = 20, \varepsilon = 0.10)$, while average Explained Variance Ratio (EVR) of ROBPCA is %90.61 in classical standardized (Cstd) data, average Explained Variance Ratio (EVR) of ROBPCA is %84.74 in robustly standardized (Rstd) data. In that case, it can be considered that classical standardizing is more preferable than robustly standardizing. However we use a different criterion called as the Outlier Detection Ratio (ODR) and these values are %12.5 and %89, respectively. In other words, if the data with outliers is standardized by using classic approach, the outlier detection power of ROBPCA method reduces. Thus, results of ROBPCA are unreliable in Cstd data. When the data is standardized robustly, EVR value is almost %84.6 while there are outliers or not in data. It is noticed that the results are stable and reliable.

According to Table 1, similarly, $for\ (p = 5, n = 50, \varepsilon = 0)$, while average Explained Variance Ratio (EVR) of ROBPCA is %76.03 in classical standardized (Cstd) data, average Explained Variance Ratio (EVR) of ROBPCA is %77.03 in robustly standardized (Rstd) data. On the other hand, for $(p = 5, n = 50, \varepsilon = 0.10)$, while average Explained Variance Ratio (EVR) of ROBPCA is %84.14 in classical standardized (Cstd) data, average Explained Variance Ratio (EVR) of ROBPCA is %77.66 in robustly standardized (Rstd) data. In that case, it can be again considered that classical standardizing is more preferable than robustly standardizing.

Table 1: When p<n, the Results of ROBPCA for standardized data sets with classical and robust estimators

		Cstd	Rstd	Explanation

p=5
n=20
	$\varepsilon = 0$	%83.61	%84.57	EVR

	$\varepsilon = 0,10$	%90.31	%84.74	EVR
		%12.5 (25/200)	%89 (178/200)	ODR

p=5
n=50
	$\varepsilon = 0$	%76.03	%77.03	EVR

	$\varepsilon = 0,10$	%84.14	%77.66	EVR
		%2.8 (14/500)	%81.6 (408/500)	ODR

However the ODR values are %2.8 and %81.6, respectively. In other words, if the data with outliers is standardized by using classic approach, the outlier detection power of ROBPCA method reduces. Thus, results of ROBPCA obtained with Cstd data are unreliable. When the data is standardized robustly, EVR value is almost %77.3 while there are outliers or not in data set. So, its results are stable and reliable.

When Table 2 is investigated, similar results are noticed and similar interpretations are done. Some diagnostic plots obtained from simulation are given below. These plots are only relevant to data with outliers.

The data set used in Figure 2 (a) and (b) is same and the ROBPCA did not determine outliers (19 and 20 observations) for Cstd data in (a). But it achieved easily for Rstd data (b).

Table 2: When p>n, the Results of ROBPCA for standardized data sets with classical and robust estimators

		Cstd	Rstd	Explanation

p=75
n=20
	$\varepsilon = 0$	%79.82	%81.06	EVR

	$\varepsilon = 0,10$	%81.19	%81.21	EVR
		%0 (0/200)	%89 (178/200)	ODR

p=75
n=50
	$\varepsilon = 0$	%52.54	%53.61	EVR

	$\varepsilon = 0,10$	%55.30	%53.98	EVR
		%0 (0/500)	%100 (500/500)	ODR

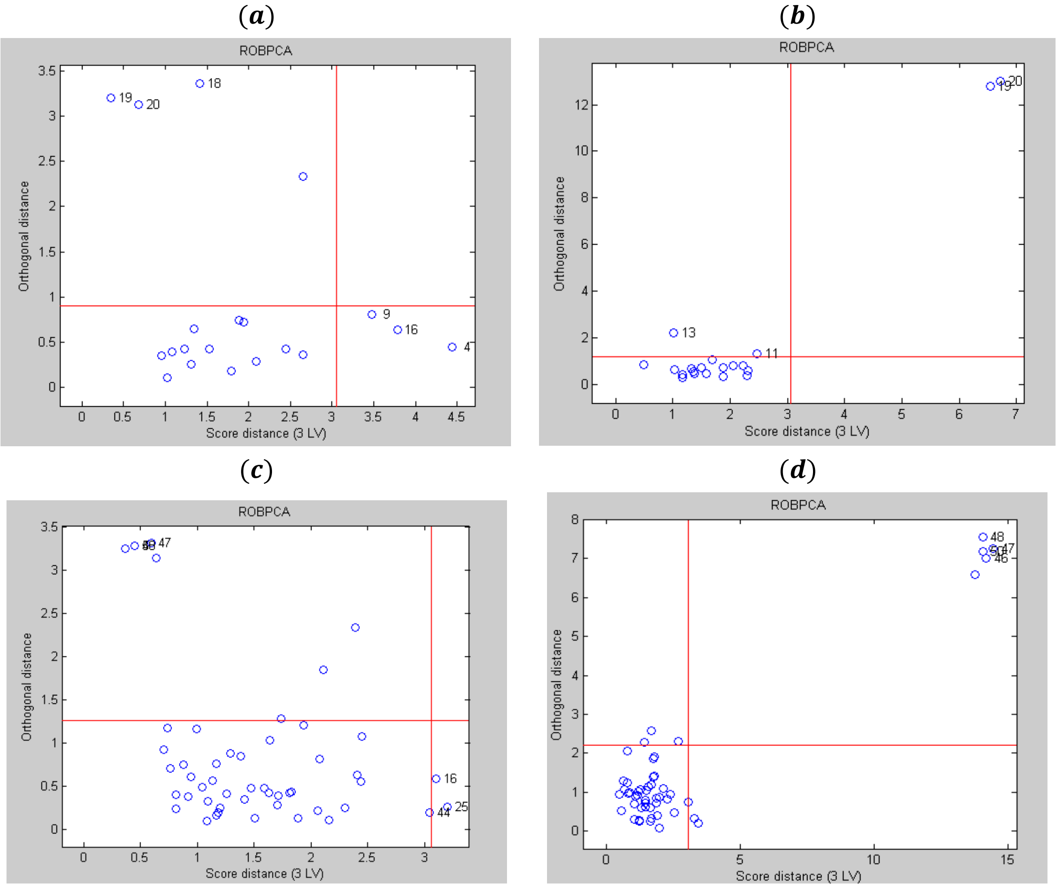
(*a*) (*b*) (*c*) (*d*)

Figure 2: The some diagnostic plots of ROBPCA (a) When (n=20,p=5, ε=0.10) in Cstd Data, (b) When (n=20,p=5, ε=0.10) in Rstd Data, (c) When (n=50,p=5, ε=0.10) in Cstd Data, (d) When (n=50,p=5, ε=0.10) in Rstd Data

Similarly, the data set used in Figure 2 (c) and (d) is same. While the ROBPCA did not determine outliers (46, 47, 48, 49 and 50 observations) for Cstd data in (c), it achieved easily for Rstd data in (d).

4. Results

The ROBPCA algorithm is alternative method to CPCA when the data set has outlier(s). But the algorithm obtains principal components by basing on eigenvalues-eigenvectors of covariance matrix. Therefore, it may give fallacious results when the units of measure of variables are different. To solve this problem, the algorithm should be performed on covariance matrix of standardized data set as the classical approach. With simulation, however, it is shown that the ODR value of ROBPCA reduces for Cstd data.

If data set consists of variables with different measure units, firstly data set should be standardized robustly and then principal components should be obtained by using ROBPCA.

References

[1] Alpar R. *Applied Multivariate Statistical Methods*. Ankara: Detay Publishing, 2011.

[2] Bulut H., Öner Y. "The Evaluation of the Development Agency Regions in Turkey In Terms Of Some Socioeconomic Indicator with Factor Analyses". *Alphanumeric Journal*, 3.1:81-88, 2015.

[3] Koch I. *Analysis of Multivariate and High-Dimensional Data*. New York: Cambridge University Press, 2014.

[4] Özdamar K. *Statistical Data Analysis with Package Softwares-2*. Eskişehir: Nisan Publishing, 2013.

[5] Rencher A. C. *Methods of Multivariate Analysis*. New York: John Wiley & Sons; 2003.

[6] Tatlıdil H. *Applied Multivarite Statistical Analysis*. Ankara: Akademi Publishing, 1996.

[7]Hubert M., Rousseeuw P. J., Branden K. V. "ROBPCA: A New Approach to Robust Principal Component Analysis". *Technometrics,* 47: 64-79, 2005.

[8] Filzmoser P. ,Todorov V. "Review of robust multivariate statistical methods in high dimension". *Analytica chimica acta*, 705.1: 2-14, 2011.

[9]Rousseeuw P. J. "Multivariate Estimation with High Breakdown Point". *Mathematical Statistics and Applications,* 283-297,1985.

[10] Rousseeuw P. J., van Driessen K. "A Fast Algorithm for the Minimum Covariance Determinant Estimator". *Technometrics*, 41: 212-223, 1999.

[11] Friedman, J. H., Tukey, J. W. "A projection pursuit algorithm for exploratory data analysis". 881-889, 1979.

[12] Huber P. J."Projection Pursuit". *The Annals of Statistics,* 13: 435-475, 1985.

[13] Li G.Y., Cheng P. "Some Recent Developments in projection pursuit in China". *Statistica Sinica,* 35-51, 1993.

[14] Croux C., Grazen R. "High Breakdown Estimators for Principal Components: The Projection-Pursuit Approach Revisited". *Journal of Multivariate Analysis*, 95.1:206-226, 2005.

[15] Filzmoser P., Serneels S., Croux C. et al. "Robust multivariate methods: The projection pursuit approach".*Berlin: Springer Berlin Heidelberg,* 270-277, 2006.

[16] Moller S. F., Frese J. V., Bro R., "Robust methods for multivariate data analysis". *Journal of Cheomemetrics,*19: 549-563,2006.

[17] Hubert, M., Rousseeuw, P., Verdonck, T. "Robust PCA for skewed data and its outlier map". *Computational Statistics & Data Analysis*, 53(6): 2264-2274, 2009.

[18] Erbaş S. O. *Probability and Statistics*. Ankara: Gazi Publishing, 2008.

[19] Maronna R. A., Martin R. D., Yohai, V. J. *Robust statistics*. Chichester: John Wiley & Sons, 2006.