Robust sparse principal component analysis: situation of full sparseness

B. BARIŞ ALKAN AND I. ÜNALDI

Abstract

Principal Component Analysis (PCA) is the main method of dimension reduction and data processing when the dataset is of high dimension. Therefore, PCA is a widely used method in almost all scientific fields. Because PCA is a linear combination of the original variables, the interpretation process of the analysis results is often encountered with some difficulties. The approaches proposed for solving these problems are called to as Sparse Principal Component Analysis (SPCA). Sparse approaches are not robust in existence of outliers in the data set. In this study, the performance of the approach proposed by Croux et al. (2013), which combines the advantageous properties of SPCA and Robust Principal Component Analysis (RPCA), will be examined through one real and three artificial datasets in the situation of full sparseness. In the light of the findings, it is recommended to use robust sparse PCA based on projection pursuit in analyzing the data. Another important finding obtained from the study is that the BIC and TPO criteria used in determining lambda are not much superior to each other. We suggest choosing one of these two criteria that give an optimal result.

Mathematics Subject Classification 2000: 62hxx, 62-07
Key Words: Principal Component, Outliers, Sparsity, Robustness.

1. INTRODUCTION

Principal component analysis (PCA) is the first referenced dimension reduction method when there is a multidimensional data set. PCA aims to reduce size by finding k $(k \leq p)$ linear combinations of the p original variables in a reduced-dimensional space. Thus, more easily interpretable results are obtained. Principal components (PCs) are vectors that make the maximum variance of k linear combinations (Croux et al., 2013). The first k principal component explains an important part of the total variance of the original data. The PCA uses the sample covariance (or correlation) matrix and the mean vector. Although PCA is a widely used multivariate statistical analysis method in almost all scientific fields, it has too many deficiencies.

In PCA, there is a transformation matrix that identifies the principal components It is called the *loading matrix*. Interpretation of loadings is often a difficult process. Various suggestions have been put forward in the literature to solve the problem of interpretability of PCA results. One of these suggestions is to use the rotation methods

10.2478/jamsi-2022-0001
©2022 Bilal Bariş Alkan, Isil Ünaldi
This is an open access article licensed under the Creative Commons Attribution Licence (http://creativecommons.org/licenses/by/4.0).

to obtain the simple structure of the loading matrix (Jolliffe, 1995). Researchers, rather than the contribution of a variable with moderate loading values, which contribute a weak or uncertain contribution to a principal component, want to be concerned with the contribution of variables that have the big loadings. Otherwise, interpreting principal components can be very difficult. In order to increase the interpretability of PCA results, many approaches, called sparse principal component analysis (SPCA), are presented to estimate the principal components with many zero loadings. When the dataset is multidimensional, the SPCA is useful because only a subset of variables needs to be analyzed. SPCA offers researchers the advantage of easier interpretation of principal components. The easiest way to do this is to set the loadings to zero that have smaller values than a certain threshold. This method is called simple thresholding. Cadima and Jolliffe (1995) found that this method could be misleading. In this context, they emphasized the necessity of examining the standard deviations of variables for determining the contribution of a variable to a given principal component. In the context of PCA, the concept of sparseness was first described by Jolliffe et al. (2003). Later, Zou et al. (2006) suggested SPCA algorithm based on elastic-net regression, which gives better results than SCoTLASS. The adaptations and improvements of the method were made by Wang et al. (2009). D'Aspremont et al. (2007) suggested a direct formulation for SPCA. Sigg et al. (2008) designed EM-PCA based on probabilistic PCA to solve SPCA. Witten et al. (2009) developed a general procedure for the separation of penalty matrix. They showed how this procedure is applied. Journée et al. (2010) derived the GPower algorithm that formulated the SPCA. Guo et al. (2010) have introduced a fusion penalty that captures block structures within the variables. Other recommendations for obtaining SPCA are described in Jenatton et al. (2010) and Bien et al. (2010).

In the presence of outliers in the dataset, many robust alternative solutions have been proposed for PCA. The most important of these were Li and Chen (1985), Hubert et al. (2002), projection pursuit PCA approach given by Croux and Ruiz-Gazen (2005), global PCA approach given by Locantore et al. (1999) as a robust PCA approach. This approach works well as long as there is a robust estimate of multivariate location and scale is possible.

The robust PCA approach proposed by Croux and Ruiz-Gazen (2005) based on the projection pursuit, is suitable when the number of variables is more than the number of observations and the presence of high dimensional data sets. The advantage of the projection pursuit method is that it is not necessary to estimate the covariance matrix for PCA. Robust estimates of eigenvalues and eigenvectors are obtained successively. In the literature, it is possible to find many studies about the robust versions of the PCA and their robustly characteristics. In addition, sparse PCA has also gained importance in the analysis of multidimensional (large data) datasets, which have emerged in recent years as a result of increasingly rapid technological developments, especially in areas such as internet technologies, gene expression and machine learning. He et al. (2014) combining the advantageous properties of Croux et al. (2013) sparse and robust PCA, proposed an algorithm for improving the durability of SPCA. They examined the robustness and effectiveness of the proposed algorithm through artificial and real data sets. A recent study on SPCA was made by Hubert et al. (2016). The main difference from Hubert et al. (2016) 's study of the Croux et al. (2013) is the separation of the sparse step and the steps of identifying outlier observations.

In this study, the performance of the approach proposed by Croux et al. (2013), which combines the important properties of SPCA and Robust Principal Component Analysis (RPCA), will be examined through one real and three artificial datasets in the situation of full sparseness. In the light of the findings, in cases where the number of variables (p) is greater than the number of observations (n), it is recommended to use robust sparse PCA based on projection pursuit in analysing the data. Another important conclusion obtained from the study is that the BIC and TPO criteria used in determining λ are not much superior to each other. We propose to choose the criterion that gives an optimal result. In the second part of the study, the mathematical theory required for SPCA and robust sparse PCA is summarized. In the next section, the applications of the methods discussed in detail in the previous sections on real and artificial datasets are given. In the last section, the results obtained from the study are discussed.

2. SPARSE PCA (SPCA)

The first PCA vector for the $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^p$, n multivariable observations in the rows of the X data matrix is given by,

$$\text{argmax}_{\|\boldsymbol{p}\|=1} V(\boldsymbol{p}'x_1, \dots, \boldsymbol{p}'x_n) \tag{1}$$

Here, V is a measure of variance. In the standard non-robust state V is the experimental variant (var) and the optimal $\boldsymbol{p}_1$ vector corresponds to the first eigenvector of the sample covariance matrix. Equation (1) is the projection pursuit formulation to find the first PC.

Robust PCA vectors can be easily obtained by taking a robust variance measure for V. This measure can be selected as a quadratic median absolute deviation or as a more efficient quadratic $Q_n$estimator.

The Q_n scale estimator is defined as the first quarter of all binary distances, $|y_i - y_j|$, $1 \le i < j \le n$, for a univariate data set $y_1, \dots, y_n$. This first quarter should then be multiplied by the constant 2.219 to obtain a consistent estimator for a normal distribution scale. Therefore, Q_n^2 is a consistent and robust predictor of variance. Croux and Ruiz-Gazen (2005) suggested using the Q_n^2 estimator as the projection tracking index which provides robust and effective estimates for the basic components. In this study, Q_n^2 was taken as a robust variance estimator.

Suppose the first $j-1$ PCA vectors are present ($j>1$). In this case j. vector ($j \le p$) is defined as,

$$\text{argmax}_{\|\boldsymbol{p}\|=1, \boldsymbol{p}\perp\boldsymbol{p}_1, \dots, \boldsymbol{p}\perp\boldsymbol{p}_{j-1}} V(\boldsymbol{p}'x_1, \dots, \boldsymbol{p}'x_n). \tag{2}$$

The sparsity of PCA was first used with SCoTLASS method by Jolliffe et al. (2003). SCoTLASS combines an L_1 constraint with PCA to provide sparse loadings. The resulting objective function is written as

$$max_{\|\boldsymbol{p}\|=1, \boldsymbol{p}\perp\boldsymbol{p}_1, \dots, \boldsymbol{p}\perp\boldsymbol{p}_{j-1}} \boldsymbol{p}'\boldsymbol{S}\boldsymbol{p}, \quad \text{constraint } \|\boldsymbol{p}\|_1 \le \boldsymbol{t}, \tag{3}$$

This objective function tries to find the vertical loadings $\boldsymbol{p}_j$that maximize the variance described. Here, $\|\boldsymbol{p}_j\|_1$ is the L_1 norm of $\boldsymbol{p}_j$, $\|\boldsymbol{p}\|_1 = \sum_{j=1}^{p} |\boldsymbol{p}_j|$. The dual formulation of the above form is given by

$$max_{\|\boldsymbol{p}\|=1,\boldsymbol{p}\perp\boldsymbol{p}_1,\ldots,\boldsymbol{p}\perp\boldsymbol{p}_{j-1}}(\boldsymbol{p}'\boldsymbol{S}\boldsymbol{p}-\lambda_1\|\boldsymbol{p}\|_1) \quad (4)$$

Where $\boldsymbol{p_j}$ is the $\boldsymbol{j}$th PCA vector. $\boldsymbol{\lambda_1}$ is the sparsity parameter for SCoTLASS and is used instead of $\boldsymbol{t}$. The larger values of $\boldsymbol{\lambda_j}$ mean greater sparsity and the zero value means no sparsity.

2.1. Robust Sparse PCA

Croux et al. (2013) proposed the robust SPCA method, which combines the projection pursuit (PP) approach and SPCA. Their approach consists of adding L_1 penalty to PP equations. This method searches for vectors that maximize the scale of the data projected onto these vectors under the constraint that vector loadings are not too large. The principal components are obtained directly using the PP approach without estimating a covariance matrix.

To find the first sparse PCA vector, the L_1 constraint is added and

$$\tilde{\boldsymbol{p}}_1 = \text{argmax}_{\|\boldsymbol{p}\|=1}(V(\boldsymbol{p}'x_1,\ldots,\boldsymbol{p}'x_n)-\lambda_1\|\boldsymbol{p}\|_1) \quad (5)$$

equation is obtained. Here, the vector $\tilde{\boldsymbol{p}}_1$ is the first SPCA vector, its sparsity is controlled by parameter λ_1. If $\lambda_1 = 0$, the constraint is not added to the first PCA vector $\boldsymbol{p}_1$. Similarly, jth SPCA vector ($1 < j \leq p$),

$$\tilde{\boldsymbol{p}}_j = \text{argmax}_{\|\boldsymbol{p}\|=1,\boldsymbol{p}\perp\tilde{\boldsymbol{p}}_1,\ldots,\boldsymbol{p}\perp\tilde{\boldsymbol{p}}_{j-1}}(V(\boldsymbol{p}'x_1,\ldots,\boldsymbol{p}'x_n)-\lambda_j\|\boldsymbol{p}\|_1) \quad (6)$$

as defined. Here, V is the variance estimator, a measure of the scale. This is the empirical (experimental) variant for the classical basic components. For the robust sparse PCA, this variance estimator is a robust variance estimator such as squared Q_n. If $V = var$ exist, then equality (4) and equality (5) are the same. Croux et al. (2013), Rousseeuw and Croux (1993) proposed the robust Q_n estimator. This Q_n estimator is the first quartile of the binary distances between the elements of a vector.

It is not easy to solve the optimization problems given in equations (5) and (6). Croux et al. (2013) is an algorithm based on iterative grid searches in spaces stretched by binary pairs of variables to find sparse vectors. Croux et al. (2007), the grid algorithm

used to obtain PCA vectors is a well-working algorithm. This algorithm was developed by Croux et al. (2013) expanded to sparse PCA.

Grid algorithm is a kind of coordinate origin methods. The extended grid algorithm always provides robust and sparse solutions with an appropriate calculation time (Croux et al., 2013).

2.2. Selection of Sparsity Parameter λ

The selection of the sparsity parameter λ is made by optimizing the objective function calculated for a set of n. λ can take different values, ranging from zero to the maximum value. If the maximum value that can be taken by λ cannot be determined, then the minimum λ value that provides full sparseness is preferred. The absolute sparseness state refers to the minimum absolute sum of the loadings, which usually consist only of zero and one in the loadings matrix. Researchers can complete this process using one of two optimization approaches for λ selection. These approaches are given as TPO (Tradeoff Product Optimization) and BIC (Bayes Information Criteria) respectively. The BIC approach selects the same λ value for all PCs and expresses a specific choice of the number of PCs considered. However, the TPO approach is optimized for each PC separately, so that different λ values are derived in a model that is not bound to a decision over k. In the TPO approach, tradeoff in the context of SPCA refers to the sparsity gained against the explained variance loss. The TPO maximizes the product of the variance explained by the number of zero loadings of the sparse principal component jth. Here, λ_j values for k PC according to TPO approach are obtained by equation (7).

$$\lambda_j^{TPO} = V\big((STB)_j\big)\alpha_j, \;\; j = 1,2,\ldots,k \tag{7}$$

where α_j is the number of zero loadings for the jth SPC. Croux et al. (2013) used a BIC type criterion to select the λ. Let P_k^s and P_k^c be the matrices of loadings that contain sparse and non-sparse basic components that contain the first k PC, respectively. From this, residual matrices $R^s = X - XP_k^s(P_k^s)'$ and $R^c = X -$

$XP_k^c(P_k^c)'$are defined. The jth columns of residual matrices are defined as $r_j^s = (r_{1j}^s, r_{2j}^s, \dots, r_{nj}^s)$ and $r_j^c = (r_{1j}^c, r_{2j}^c, \dots, r_{nj}^c)$ respectively.

The BIC criterion is defined by the equation

$$\lambda^{BIC} = \frac{\sum_{j=1}^{k} V(r_j^s)}{\sum_{j=1}^{k} V(r_j^c)} + \beta \frac{\log n(n)}{n} \tag{8}$$

Here, β is the number of non-zero loadings.

In practice, the selection of λ is made by minimizing λ in a grid $[0, \lambda_{max}]$ range. Here, λ_{max} is the result of the full sparseness of k-component sparse PCA results. In addition to λ, it is also necessary to select the number of components of k. The appropriate choice of k is an old and common problem in PCA. Many suggestions have been made to solve this problem. In this study, the number of k is determined by the robust variance explanation ratio (VER),

$$VER_k = \frac{V(T_k)}{V(X)} \tag{9}$$

Here, T_k contains the principal component scores. For $V = var$, VER_k, k is the ratio of the sum of the biggest eigenvalues to the sum of all eigenvalues of the sample covariance matrix. For this value of k, a selected λ must have a more sparse matrix and a smaller robust variance.

3. ARTIFICIAL AND REAL DATA EXAMPLES

In this section, the performance of RPCA based on PP (PP-RPCA) nd robust sparse PCA based on PP (PP-RSPCA) methods based on projection-pursuit (PP) will be compared over one real and three artificial data sets. The codes written with the help of pcaPP, pls, lars, lasso2, spls, lattice, elasticnet, stats, stats4, robust, robustbase, rrcov, rrcovHD libraries were used in the R program to apply PP-RPCA and PP-RSPCA to real and artificial data sets (R Core Team,2019).

Artificial data sets were used in the design given in the study. As the actual data set was made for the monthly stock exchange traded on the Istanbul Stock Exchange between January 2005 and March 2013.

Artificial Data Designs

1) Artificial Data Set 1

10% outlier situation: A data set of 100 observations (n = 100) and 5 variables (p = 5) was generated from a multivariate t distribution (T_5):10% and $N_5(0, I)$:90%.

20% outlier situation: A data set of 100 observations (n = 100) and 5 variables (p = 5) was generated from a multivariate t distribution (T_5): 20% and $N_5(0, I)$:80%.

30% outlier situation: A data set of 100 observations (n = 100) and 5 variables (p = 5) was generated from a multivariate t distribution (T_5): 30% and $N_5(0, I)$:70%

2) Artificial Data Set 2

It was planned to produce a 215 × 6 data set consisting of 215 observations (n = 215) and 6 variables (p = 6). Of these 215 observations, N_6 ($\mu 1, \Sigma 1$): 200 and N_6 ($\mu 2, \Sigma 2$): 15.

200 observations were generated from the multivariate normal distribution and 15 from the multivariate normal distribution.

$\mu 1 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0],\ \Sigma 1 = diag\ [5 \;\; 1 \;\; 1 \;\; 1 \;\; 1 \;\; 1]$

$\mu 2 = [0 \;\; 20 \;\; 20 \;\; 20 \;\; 20 \;\; 20], \Sigma 2 = diag\ [1 \;\; 1 \;\; 1 \;\; 1 \;\; 1 \;\; 1]$

3) Artificial Data Set 3 ($p > n$ status)

An 18 × 20 data set consisting of 18 observations and 20 variables was produced. Of the 18 observations, N_{20} ($\mu 3, \Sigma 3$): 15 and N_{20} ($\mu 4, \Sigma 4$): 3.

$\mu 3 = [0 \;\; 0 \;\; 0 \;\; \ldots \;\; 0], \Sigma 3 = diag[5 \;\; 1 \;\; 1 \;\; \ldots \;\; 1]$

$\mu 4 = [0 \;\; 20 \;\; 20 \;\; \ldots \;\; 20], \Sigma 4 = diag[1 \;\; 1 \;\; 1 \;\; \ldots \;\; 1]$

Real Data (Stock Exchange Data)

The data set containing the quarterly return rates of eight stocks traded between January 2005 and March 2013 on the Istanbul Stock Exchange was used. The analyzed dataset has a size of 33 × 8.

3.1. Comparison of PP-RPCA and PP-RSPCA over Artificial Data Set 1

The Artificial Data Set 1 has been analyzed with the written R code and the results are presented in Table 3.1, Table 3.2, Table 3.3 according to 10%, 20%, 30% outlier rates. For the 10% outlier rate, λ's were determined by using TPO and BIC criteria, and the most optimal λ value was selected as $\lambda = 7.8$ according to TPO criteria. When Table 3.1 is examined, it is seen that PP-RPCA explains 88% of the total variance with the first four principal components. When PP-RSPCA results are analyzed, it is seen that 84% of the total variance is explained by the first four main components. In this case, the cost of using the sparsity feature is equivalent to the loss of 4% variance explanation rate. However, in Table 3.1, it is seen that the number of non-zero loadings in the PP-RPCA is 20 while the number of non-zero loadings in the PP-RSPCA is reduced to 4. It is seen that the sparsity feature added to the loadings matrix facilitates the interpretation of the main components with extremely good performance and a low error rate.

Table 3.1. In the presence of 10% outlier of the variables, the first four non sparse robust ($\lambda = 0$) and sparse robust ($\lambda = 7.8$) main component loadings

Variables	PP-RPCA				PP-RSPCA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.59	0.30	-0.59	-0.07	1.00	0.00	0.00	0.00
X2	0.18	-0.09	0.65	0.24	0.00	0.00	0.00	1.00
X3	0.74	0.03	0.30	0.22	0.00	0.00	0.00	0.00
X4	-0.14	0.92	0.31	-0.21	0.00	0.00	1.00	0.00
X5	-0.21	0.25	-0.22	0.92	0.00	1.00	0.00	0.00
VER* %	0.25	0.24	0.23	0.16	0.25	0.21	0.19	0.18
Cumulative VER %	0.25	0.48	0.72	0.88	0.25	0.46	0.66	0.84

For the 20% outlier rate, λ's were determined by using TPO and BIC criteria and $\lambda = 12.14$ was selected as the most optimal λ value according to TPO criteria.

When Table 3.2 is examined, it is seen that PP-RPCA explains 88% of the total variance with the first four main components. When PP-RSPCA results are analyzed, it is seen that 86% of the total variance is explained by the first four main components.

In this case, the cost of using the sparsity feature is equivalent to the loss of 2% variance explanation rate. However, in Table 3.2, it is seen that the number of non-zero loads in PP-RPCA is 19 while the number of non-zero loads in PP-RSPCA is reduced to 4. It is seen that the sparsity feature added to the loadings matrix facilitates the interpretation of the principal components with extremely good performance and a low error rate.

Table 3.2. In the presence of 20% outlier of the variables, the first four non sparse robust ($\lambda = 0$) and sparse robust ($\lambda = 12.14$) main component loadings

Variables	PP-RPCA				PP-RSPCA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.04	0.73	-0.67	0.01	0.00	1.00	0.00	0.00
X2	-0.03	0.57	0.71	0.00	0.00	0.00	1.00	0.00
X3	0.68	-0.27	-0.12	-0.29	0.00	0.00	0.00	1.00
X4	-0.28	-0.24	-0.14	0.74	1.00	0.00	0.00	0.00
X5	0.67	0.15	0.12	0.61	0.00	0.00	0.00	0.00
VER* %	0.31	0.23	0.19	0.15	0.23	0.23	0.21	0.19
Cumulative VER* %	0.31	0.54	0.73	0.88	0.23	0.46	0.67	0.86

For the case of 30% outlier, λ's were determined by using TPO and BIC criteria and $\lambda = 3.87$ according to TPO criterion.

When Table 3.3 is examined, it is seen that PP-SPCA explains 91% of the total variance with the first four main components. When PP-RSPCA results are analyzed, it is seen that 81% of the total variance is explained by the first four main components. In this case, the cost of using the sparsity feature is equivalent to the loss of variance explanation rate of 10%. However, in Table 3.3, it is seen that the number of non-zero loads in PP-RPCA is 19 while the number of non-zero loadings in PP-RSPCA is reduced to 7. It is seen that the sparsity feature added to the loadings matrix facilitates the interpretation of the principal components with extremely good performance and a low error rate.

Table 3.3. In the presence of 30% outlier of the variables, the first four non sparse robust ($\lambda = 0$) and sparse robust ($\lambda = 3.87$) main component loadings

Variables	PP-RPCA				PP-RSPCA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.04	0.73	-0.67	0.01	0.00	1.00	0.00	0.00
X2	-0.03	0.57	0.71	0.00	0.00	0.00	1.00	0.00
X3	0.68	-0.27	-0.12	-0.29	0.00	0.00	0.00	1.00
X4	-0.28	-0.24	-0.14	0.74	1.00	0.00	0.00	0.00
X5	0.67	0.15	0.12	0.61	0.00	0.00	0.00	0.00
VER %	0.31	0.23	0.19	0.15	0.23	0.23	0.21	0.19
Cumulative VER %	0.31	0.54	0.73	0.88	0.23	0.46	0.67	0.86

3.2. Comparison of PP-RPCA and PP-RSPCA over Artificial Data Set 2

The Artificial Data Set 2 was analyzed by R code and the results are presented in Table 3.4.

Table 3.4. The first four non sparse robust ($\lambda = 0$) and sparse robust ($\lambda = 11.03$) principal component loadings of variables in Artificial Dataset 2

Variables	PP-RPCA				PP-RSPCA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.99	-0.06	0.03	0.11	1.00	0.00	0.00	0.00
X2	0.04	0.26	0.86	-0.18	0.00	0.00	0.00	1.00
X3	0.08	0.79	-0.23	-0.47	0.00	0.00	0.00	0.00
X4	0.07	0.00	0.08	-0.18	0.00	0.00	1.00	0.00
X5	0.07	0.40	-0.36	0.39	0.00	0.00	0.00	0.00
X6	-0.11	0.37	0.27	0.74	0.00	1.00	0.00	0.00
VER %	0.23	0.18	0.17	0.16	0.18	0.18	0.18	0.16
Cumulative VER %	0.23	0.41	0.59	0.74	0.18	0.36	0.54	0.71

For Artificial Data Set 2, λ's were determined by using TPO and BIC criteria, and the most optimal λ value was chosen as $\lambda = 11.03$ according to BIC criteria.

When Table 3.4 is examined, it is seen that PP-RPCA explained 74% of the total variance with the first four main components. When PP-RSPCA results are analyzed, it is seen that 71% of the total variance is explained by the first four main components.

In this case, the cost of using sparsity is equivalent to loss of 3% variance explanation rate. However, in Table 3.4, the number of non-zero loadings in PP-RPCA is 23 while it is seen that the number of non-zero loadings in PP-RSPCA is reduced to 4.

It is seen that the sparsity feature added to the loadings matrix facilitates the interpretation of the principal components with extremely good performance and a low error rate.

3.3. Comparison of PP-RPCA and PP-RSPCA over Artificial Data Set 3 ($p > n$)

The Artificial Data Set 3 has been analyzed with R code written and the results are presented in Table 3.5. For Artificial Data Set 3, λ's were determined by using TPO and BIC criteria, and the most optimal λ value was selected as λ = 50.69 according to BIC criteria.

When Table 3.5 is examined, it is seen that PP-RPCA explains 87% of the total variance with the first four main components. When PP-RSPCA results are analyzed, it is seen that 84% of the total variance is explained by the first four main components. In this case, the cost of using sparsity is equivalent to loss of 3% variance explanation rate. However, it is seen that the number of non-zero loadings in PP-RPCA in Table 3.5 is 71 while the number of non-zero loadings in PP-RSPCA is reduced to 4. It is seen that the sparsity feature added to the loadings matrix facilitates the interpretation of the principal components with extremely good performance and a low error rate.

Table 3.5. The first four non sparse robust (λ = 0) and sparse robust (λ = 50.69) principal component loadings of the variables in Artificial Dataset 3

Variables	PP-RPCA				PP-RSPCA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.62	-0.18	0.36	-0.25	0.00	0.00	0.00	1.00
X2	0.35	0.03	-0.10	0.13	0.00	0.00	0.00	0.00
X3	-0.22	0.25	-0.22	-0.10	0.00	0.00	0.00	0.00
X4	-0.04	-0.15	0.00	-0.02	0.00	0.00	0.00	0.00
X5	-0.24	-0.14	-0.13	0.33	1.00	0.00	0.00	0.00
X6	0.00	0.00	0.04	0.04	0.00	0.00	0.00	0.00

X7	0.00	-0.14	-0.07	0.00	0.00	0.00	0.00	0.00
X8	0.13	0.03	-0.22	-0.07	0.00	0.00	0.00	0.00
X9	0.02	0.50	0.23	0.13	0.00	0.00	0.00	0.00
X10	0.00	-0.13	-0.41	-0.50	0.00	1.00	0.00	0.00
X11	-0.20	0.16	-0.11	-0.08	0.00	0.00	0.00	0.00
X12	-0.08	0.13	0.05	-0.03	0.00	0.00	0.00	0.00
X13	0.11	-0.24	-0.12	0.05	0.00	0.00	0.00	0.00
X14	0.00	0.50	0.07	0.00	0.00	0.00	0.00	0.00
X15	-0.03	-0.21	-0.28	0.55	0.00	0.00	1.00	0.00
X16	0.01	-0.08	-0.07	0.01	0.00	0.00	0.00	0.00
X17	0.00	-0.22	0.14	0.03	0.00	0.00	0.00	0.00
X18	-0.08	0.06	-0.34	-0.41	0.00	0.00	0.00	0.00
X19	0.17	0.25	-0.10	0.07	0.00	0.00	0.00	0.00
X20	-0.53	-0.23	0.51	-0.22	0.00	0.00	0.00	0.00
VER %	0.28	0.24	0.21	0.14	0.27	0.23	0.19	0.16
Cumulative VER %	0.28	0.53	0.73	0.87	0.27	0.50	0.68	0.84

3.4. Comparison of PP-RPCA and PP-RSPCA on Real Data 1- Stock Market Data

Real Data 1-Stock Market Data is analyzed with R code written and the results are presented in Table 3.6.

Table 3.6. The first four non sparse robust ($\lambda = 0$) and sparse robust ($\lambda = 3.31$) principal component loadings of the variables in Real Data 1-Stock Market Data

Variables	PP-RPCA				PP-RSPCA			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
X1	0.36	-0.12	-0.48	0.19	0.00	0.00	0.00	1.00
X2	-0.02	0.56	-0.22	-0.26	0.79	-0.01	0.00	0.00
X3	0.29	0.39	0.19	-0.45	0.00	0.00	0.17	0.01
X4	0.22	0.00	-0.34	0.18	0.00	0.00	0.00	0.00
X5	0.15	0.54	0.11	0.02	0.00	0.00	0.99	0.00
X6	0.59	0.00	-0.35	0.01	0.01	0.10	0.00	0.00
X7	0.46	0.09	0.61	0.55	0.00	1.00	0.00	0.00
X8	-0.39	0.47	-0.26	0.61	0.62	0.01	0.00	0.00

VER %	0.34	0.24	0.15	0.15	0.28	0.19	0.19	0.10
Cumulative VER %	0.34	0.58	0.73	0.87	0.28	0.47	0.66	0.76

For the Stock Market Data, $\boldsymbol{\lambda}$'s were determined by using TPO and BIC criteria, and the most optimal $\boldsymbol{\lambda}$ value was selected as $\boldsymbol{\lambda} = 3.31$ according to TPO criteria.

Table 3.6 shows that PP-RPCA explains 87% of the total variance with the first four main components. When PP-RSPCA results are analyzed, it is seen that 76% of the total variance is explained by the first four main components. In this case, the cost of using the sparsity feature is equivalent to the loss of 11% variance explanation rate. However, in Table 3.6, it is seen that the number of non-zero loadings in PP-RPCA is 30 while the number of non-zero loadings in PP-RSPCA is reduced to 11. It is seen that the sparsity feature added to the loadings matrix facilitates the interpretation of the principal components with extremely good performance and a low error rate.

4. DISCUSSION AND CONCLUSION

The sparse principal component analysis is very useful in finding vectors in the direction of maximizing the variance described and improving the interpretability of the principal components. The robust sparse principal component analysis also allows the identification of principal component analysis vectors, which also facilitate interpretability and are unaffected by outliers.

The projection-pursuit approach discussed maximizes a robust variance to find the principal component analysis vectors. One of the advantages of the projection-pursuit approach is that the principal components are calculated sequentially and the algorithm stops when the desired number of principal components is reached. This feature is particularly useful in cases where the number of variables is greater than the number of observations $(p > n)$ in high-dimensional data sets.

In the study, the performance of the RSPCA method suggested by Croux et al. (2013) was examined on one real and three artificial data sets with different outliers ratios. Artificial dataset 3 includes the number of variables (p) > number of observations (n). Therefore, the findings to be obtained in the solution of such data

set with RSPCA are very important. The projection-pursuit method is also a widely used robust method in case of $p > n$.

The optimal value of the sparse parameter λ used to obtain the optimal result in terms of both interpretability and variance explanation rate can be found by BIC and TPO approaches. The BIC and TPO criteria used in the determination of lambda did not show much superiority compared to each other. However, in each application, the optimum of the λ determination criteria (BIC and TPO), which gives the λ value, has been selected.

In the future study, RSPCA based on projection-pursuit discussed in this study can be compared with RSPCA results based on different robust variance estimators. For small sample and large sample cases, the performance of the methods for different outlier ratios can be evaluated. The effects of the use of different approaches such as MCD instead of projection-pursuit approach on RSPCA can also be examined.

REFERENCES

Bien, J., Xu, Y., Mahoney, M. W. 2010. CUR from a sparse optimization viewpoint. *In Advances in Neural Information Processing Systems* (pp. 217-225).

Cadima, J., Jolliffe, I. T. 1995. Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics*, 22(2): 203-214.

Croux, C., Ruiz-Gazen, A. 2005. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1): 206-226.

Croux, C., Filzmoser, P., Oliveira, M. R. 2007. Algorithms for Projection–Pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2): 218-225.

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

D'Aspremont, A., El Ghaoui, L., Jordan, M. I., Lanckriet, G. R. 2007. A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49(3): 434-448.

Guo, J., James, G., Levina, E., Michailidis, G., Zhu, J. 2010. Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical Statistics*, 19(4): 930-946.

He, Y., Zhou, W., Qian, G., Chen, B. 2014. Methane storage in metal–organic frameworks. *Chemical Society Reviews*, 43(16): 5657-5678.

Hubert, M., Rousseeuw, P. J., Van Aelst, S. 2008. High-breakdown robust multivariate methods. *Statistical Science*, 92-119.

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

Jenatton, R., Obozinski, G., Bach, F. 2010, March. Structured sparse principal component analysis. *In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 366-373).

Jolliffe, I. T. 1995. Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics*, 22(1): 29-35.

Jolliffe, I. T., Trendafilov, N. T., Uddin, M. 2003. A modified principal component technique based on the LASSO. *Journal of computational and Graphical Statistics*, 12(3): 531-547.

Journée M., Nesterov Y., Richtárik P., Sepulchre R. 2010 Generalized power method for sparse principal component analysis. *J Mach Learn Res* 11:517–553.

Li, G., Chen, Z. 1985. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association*, 80(391): 759-766.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., ..., Fan, J. 1999. Robust principal component analysis for functional data. *Test*, 8(1): 1-73.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rousseeuw, P. J., Croux, C. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424): 1273-1283.

Rousseeuw, P. J., Driessen, K. V. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3): 212-223.

Sigg, C. D., Buhmann, J. M. 2008, July. Expectation-maximization for sparse and non-negative PCA. *In Proceedings of the 25th international conference on Machine learning*, ACM, pp. 960-967.

Wang, H., Li, B., Leng, C. 2009. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 671-683.

Zou, H., Hastie, T., Tibshirani, R. 2006. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2): 265-286.

Bilal Barış ALKAN
Department of Educational Measurement and Evaluation,
University of Akdeniz, Antalya 07000, Turkey.
Email: bbarisalkan@gmail.com (corresponding author)

Işıl ÜNALDI
Department of Department of Biostatistics and Medical Informatics,
University of Ondokuz Mayıs, Samsun 55000, Turkey.
Email: isil.unaldi@omu.edu.tr