GÜFBED/*GUSTIJ* (2020) 10 (3): 581-591 **Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi**
DOI: 10.17714/gumusfenbil.641791 Araştırma Makalesi / *Research Article*

Some Regression Methods Based on Principal Components

Temel Bileşenlere Dayalı Bazı Regresyon Yöntemleri

Fatma Sevinç KURNAZ*

Yıldız Teknik Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, 34220, İstanbul

• Geliş tarihi / *Received*: 02.11.2019 • Düzeltilerek geliş tarihi / *Received in revised form*: 14.04.2020 • Kabul tarihi / *Accepted*: 22.04.2020

Abstract

Principal component analysis (PCA) is commonly used technique in data processing and dimensionality reduction. However, PCA is very sensitive to outliers. To deal with this problem, the robust principal component analysis (RPCA) using Projection-Pursuit (PP) is a very appealing method. On the other hand, combining PCA on explanatory variables with least squares regression yields to principal component regression (PCR). Taking into consideration this general structure of PCR, we combine (R)PCA with OLS and MM regression estimators and show the performances of examined methods on extensive simulation studies and real data examples.

Keywords: Ordinary Least Squares, Principal Component Analysis, Principal Component Regression, Robustness

Öz

Temel Bileşen Analizi (TBA) veri işleme ve boyut indirgeme için sıklıkla kullanılan bir tekniktir. Ancak, TBA verideki sapan değerlere karşı oldukça duyarlıdır. Bu problemle başa çıkmak için iz düşüm takibini (projection pursuit) temel alan dayanıklı TBA kullanımı oldukça dikkat çekici bir yöntem olarak ileri sürülmüştür. Diğer taraftan, Temel Bileşen Regresyonu (TBR), TBA ile en küçük kareler regresyonunun birleşimi olarak görülebilmektedir. Bu çalışmada, TBR'nin bu genel yapısı dikkate alınarak, sapan değerlere karşı dirençli olan versiyonları üzerinde çalışılmıştır. İncelenen yöntemlerin performans karşılaştırmaları detaylı bir benzetim çalışması ve çeşitli gerçek veri kümeleri üzerinde gösterilmiştir.

***Anahtar Kelimeler**: En Küçük Kareler, Temel Bileşen Analizi, Temel Bileşen Regresyonu, Dayanıklılık*

*Fatma Sevinç KURNAZ; fskurnaz@yildiz.edu.tr, Tel: (0212) 383 72 67, orcid.org/0000-0002-5958-7366

1. Introduction

The primary work in science and technology is modeling a dependent variable by one or several explanatory variables. For such a modeling, the ordinary least squares (OLS) regression is the most common method performed directly with the values of the explanatory variables. However, in presence of multicollinearity among explanatory variables, the OLS estimator can give wrong information, and if the number of variables exceeds the number of observations ($n \leq p$) it can even not be computed. These problems are quite common in many applications such as chemometrics, medicine, social sciences, etc. For such situations, the principal component regression (PCR) is a very useful alternative that is the combination of the Principal Component Analysis (PCA) and the OLS estimator. The PCR solves the problem of data collinearity reducing the number of variables. But the obtained variables, are no longer the original measured variables but linear combinations thereof. The linear combinations for PCR are the principal component (PC) scores of variables obtained by PCA. Furthermore, the PCR also allows solving the computational problem of high dimensional data. Kendall (1957) proposed the PCR, Hotelling (1957) introduced it in an article in the same year, and a very famous application was implemented by Jeffers (1967), but it still keep to be a cornerstone in the literature and are still a hot research topic of regression methods. For a short overview, see (Varmuza and Filzmoser, 2008).

On the other hand, most real world data sets include outliers that follow different pattern when compared with others in data set. Outliers may cause a negative effect on regression methods besides PCA, especially in high dimensional data, where the influence of observations is more critical because the sample size n is remarkably less than the number of explanatory variables p. Even one massive outlier can heavily distort the results of these methods. In addressing this problem, the robust principal component analysis (RPCA) methods have been proposed with diverse modifications (Maronna, 1976; Campbell, 1980; Li and Chen, 1985; Croux and Haesbroeck, 2000; Hubert et al., 2002; Hubert et al., 2005; Croux and Ruiz-Gazen, 2005; Croux et al., 2007). In this paper we consider the RPCA method proposed by (Li and Chen, 1985), which is constructed using the projection-pursuit (PP) (Huber, 1985). The idea is straightforward: a robust measure of variance is taken instead of taking the variance as a projection index. Theoretical advantages of this idea are examined by Li and Chen (1985) and Cui (2003), and recently this method is studied in detailed by Croux and Ruiz-Gazen (2005). Obtaining RPCs with this way, several algorithms are introduced such as (Li and Chen, 1985; Johnson and Wichern, 1998; Xie et al., 1993), but these algorithms are not made publicly available. Equivalent to Xie et al. (1993) more stable versions are studied by (Hubert et al., 2002) and (Croux and Ruiz-Gazen, 2005). The algorithm proposed by Xie et al. (1993) is called as CR algorithm which is very simple, fast to compute and easy to implement. However, disadvantages of the CR algorithm for high dimensional regression applications are discovered by (Croux et al. 2007). Recently, a new algorithm is proposed to deal with this problem and it is much more precise, while still being computationally efficient (Croux et al., 2007). This algorithm is called as the GRID algorithm. Both of CR and GRID algorithms are freely available R package and it can be downloaded from (http://www.R-project.org) as the library pcaPP. For the other proposals of RPCA, besides the PP based approach, we suggest to see in (Croux and Haesbroeck, 2000; Maronna, 2005).

The aim of this paper is to focus on the estimation methods include robustness, taking the pure idea of the PCR using PCA and RPCA with some regression methods and then compare them with each other. Although the proposed methods can be seen very straightforward, to the best of our knowledge, there is not any performance comparison of them in the literature. In order to fill this gap, during this paper, we aim to provide some practical hints for use of these methods.

The outline of this paper is as follows. In Section 2 we recall the classical OLS estimator and its robust counterpart MM estimator. In Section 3, we remind PCA and RPCA. Then, we introduce the combination of OLS and MM estimators with PCA and RPCA. In Section 4, the comparison results and discussion are given throughout extended simulation studies. Section 5 shows the performances using real data sets, the prostate cancer data, the forest fires data, the glass vessels data and the NCI data. The final Section 6 includes conclusion.

2. Classical Regression Methods

Let us take into consideration the classical linear regression model

$$y = X\xi + \epsilon, \quad (1)$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ is dependent variable with n observations; $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \cdots \mathbf{x_p})$ is the $n \times p$ data matrix which enjoys the information of the p explanatory variables and the n observations; $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_n)$is the true regression coefficient vector; $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$ is the error term with standard normal distribution.

The common method to estimate the true regression coefficients is the ordinary least squares (OLS) estimator, which is defined by

$$\hat{\xi}_{OLS} = (X^T X)^{-1} X^T y. \quad (2)$$

Even though the OLS estimator has many attractive statistical properties as compared to other unbiased linear estimators, it can be affected by outliers in the data sets (Maronna et al., 2006). In order to deal with this problem, several alternative regression estimators are proposed such as S-estimator, M estimator and MM estimator (Maronna et al., 2006). We use the MM estimator that is robust to both types of outliers (Yohai, 1987). The MM regression estimator has a definition based on the M estimator of regression

$$\hat{\xi}_M = \underset{\xi}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\xi)}{\hat{\sigma}(\xi)}\right), \quad (3)$$

where ρ shows a predefined bounded function (Maronna et al., 2006). Here, $\hat{\sigma}$ stands for the scale estimator of the residuals. The M estimator is not robust against outliers in the explanatory variables (Maronna et al., 2006), so-called leverage points, and thus the MM estimator uses as a robust residual scale estimator an M estimator of scale, which is the solution of the equation

$$\sum_{i=1}^{n} \tilde{\rho} \frac{r_i(\xi)}{\sigma(\xi)} = \delta, \quad (4)$$

with $\tilde{\rho}$ taken e.g. as the bisquare function (Maronna at al., 2006), and the tuning constant δ.

In data sets with many explanatory variables and lower observations, the classical low dimensional methods - OLS and MM estimators - may not be applicable because of high multicollinearity problem and the large number of explanatory variables. In such situations, there exist some alternatives providing both regression analysis and interpretative analysis, which is very attractive property for high dimensional data. A first approach can be to use a principal component analysis, followed by a (potentially OLS) regression, so called PCR. Another one is the partial least squares regression (PLSR) (Varmuza and Filzmoser, 2008) providing a projection onto a few latent components, and it yields a vector of regression coefficients based on those latent components. The other alternative is to use the Lasso regression that is quite attractive in last decades which only uses a subset of explanatory variables (Zou and Hastie, 2005). In this paper we focus on the methods based on PCA because it is a mainstay tool and still a hot topic for applied sciences.

3. Principal Component Methods

PCA searches a linear combination of explanatory variables ensure that the maximum variance is extracted from the explanatory variables. Then it removes the obtained variance and searches the linear combination which explains the maximum proportion of the remaining variance, continue to computation of the next component until the number of components explains at least 80%, eventually 90% of the total variance.

Although PCA is very useful tool for high dimensional data structure, outliers may have severe influence on the results. Similar to the classical PCA, PP approach is proposed which yields the RPCA finding projections of the data which have maximal dispersion. Instead of using the variance as a measure of dispersion, a robust scale estimator S is used for the maximization problem. This approach was introduced by (Li and Chen, 1985), who proposed estimators based on maximizing (or minimizing) a robust scale. In this way, for n observations $\mathbf{x_1}, \mathbf{x_2}, \cdots \mathbf{x_p}$ - all of the column vectors of dimension p - the first robust principal component (RPC) is defined as

$$a_1 = \underset{\|a\|=1}{\operatorname{argmax}} S^2(a^T x_1, \cdots, a^T x_n), \quad (5)$$

where S^2 shows the variance. A particular feature is that the PP-based approach for PCA allows sequential estimation of the PCs. The subsequent PCs are obtained by imposing orthogonality conditions. Clearly, let us have already computed the (k-1)th PC at hand. Then the direction of the kth component for $1 \leq k \leq p$ is obtained by the unit vector maximizing the index S^2 of the data projected on it. Each new obtained PC has to be orthogonal to all previously obtained components. This situation can be written mathematically as follows:

$$a_k = \underset{\|a\|=1, a \perp a_1, \cdots, a \perp a_{k-1},}{\operatorname{argmax}} S^2(a^T x_1, \cdots, a^T x_n), \quad (6)$$

Note that if the variance is defined as a projection index, then the resulting method is the standard PCA. In this case, $\mathbf{a}_1$ is the eigenvector of the sample covariance matrix of the data corresponding to the largest eigenvalue (Johnson and Wichern, 1998). Eq. (1) cannot be solved easily for the other choices of S. Some approximative algorithms can be used for such a case. In this paper, we use the GRID algorithm which is quite effective to computation of RPCA (Croux et al., 2007) and has the R package (Filzmoser et al., 2018).

On the other hand, from the perspective of regression, PCR is very common method that solves the problem of multicollinearity and reduces the number of explanatory variables by means of generating linear combinations of explanatory variables. The main idea of PCR is that instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as explanatory variables, which means PCR is a combination of PCA and OLS estimator. This idea makes the PCR some kind of a regularized procedure and it is gained multitasking skills to the regression methods such as dimension reduction, easy interpretation and the most importantly such multitasking skills are available when data sets are high dimensional. Theoretically, PCA decomposes a data matrix **X** into scores **T** and loadings **P**. For a predetermined number a of PCs, **X** is reconstructed by

$$\mathrm{X} = \mathrm{TP}^T + \mathrm{E}, \tag{7}$$

where **E** is the error term, which would be zero if all possible PCs would be used in the model. Note that **y** is not considered in this step. In the linear regression model given Eq. (1), data matrix **X** is replaced by scores matrix **T** and the resulting equation is

$$\mathrm{y} = (\mathrm{TP}^T)\beta + e_T = \mathrm{T}\theta + e_T, \tag{8}$$

where $\boldsymbol{\theta} = \mathbf{P}^T\boldsymbol{\beta}$ and $\boldsymbol{e_T}$ is the error term. This indeed solves multicollinearity problem because the information of the highly correlated explanatory variables is compressed in few score vectors that are uncorrelated. OLS estimator can now be computed by

$$\hat{\theta} = (\mathrm{T}^T\mathrm{T})^{-1}\mathrm{T}^T\mathrm{y}, \tag{9}$$

where note that the new explanatory variables, so called PCs, are no longer the original measured x-variables and the estimator $\boldsymbol{\theta} = \mathbf{P}^T\boldsymbol{\beta}$ corresponds to space of PCA scores. One could obtain the regression coefficients in the original space as follows:

$$\theta = \mathrm{P}^T\beta. \tag{10}$$

Because of PCR still keeps to be cornerstone in high dimensional data structure, taking the general idea of PCA, RPCA can be used in regression as a solution to the problem of outliers in data besides the problems of multicollinearity and high dimensionality. As in PCR, considering RPCs as in inputs for the linear regression, we use OLS estimator and called the resulting method as rpcaOLS. However, in this part, an obvious question emerges, whether the OLS should be replaced by any choice of robust estimator in order to improve the prediction ability with the components selection. Therefore, another approach is to construct the MM regression estimators regressing on RPCs spaces. The obtained estimators with this idea are called as pcaMM and rpcaMM estimators. From this perspective, one could generally think these methods as a kind of regularized procedure by means of using only a subset of all RPCs for regression. More clearly, the proposed estimators have two main stages analogously to PCR. First stage is to apply PCA or RPCA on the explanatory variables and second stage is to perform a regression of response variable on the obtained PCs in the first stage.

3.1. Choosing the Optimal Number of Components

The classical approach is to determine the number of PCs is leave one-out cross-validation (CV) based on root mean squared prediction error (RMSPE). We take into consideration the similar idea to estimate the number of PCs. But rather than the classical approach CV, we use generalized cross-validation (GCV) which can be seen as a rotation-invariant form of the CV.

On the other hand, we would like to draw attention that, in the context of PCR, it is popular to center the data matrix **X** at the PCA stage since PCR required the use of PCA on data matrix **X** and PCA is sensitive to centering of the data. From this perspective, as in PCA, RPCA are performed on the centered data matrix **X**.

4. Simulation Studies

We design $16(=4\times2\times2)$ different cases for each of "low dimensional data" and "high

dimensional dat" sets using arguments the sizes (4 cases), the magnitude of correlations among variables (2 cases) and the different level of contamination in data - with %0 and %10 (2 cases). Therefore, in total, we examine the performance of the compared methods based on 32 different scenarios. In each case, the number of replications is 100.

4.1. Design of Simulation Schemes for Low Dimensional Data

We take into consideration scenarios by means of generating low dimensional data sets with $(\mathrm{n}, \mathrm{p}) = \{(100,50), (50,20), (500,20), (500,100)\}$. The true response variable is computed by

$$y_i = 1 + \mathrm{x}_i^T \beta + \varepsilon_i, \qquad (11)$$

where the error term ε_i is distributed according to a standard normal distribution $\mathbf{N}(0,1)$ for $i = 1, \cdots \mathrm{n}$. Here, the data matrix $\mathbf{X} = (\mathbf{x_j})$ is generated using two different ways. Firstly, we think the design matrix $\mathbf{X}$ from $\mathbf{N}_{\mathrm{p}}(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \boldsymbol{\rho}^{|j-k|}$, $1 \le j, k \le \mathrm{p}$, where high level of correlation is determined by $\rho = 0.9$. Secondly, we take the explanatory variables forming a block structure (BS) generating the observations of the blocks of design matrix $\mathbf{X}$ from $\mathbf{X}_{\mathrm{a}} \sim \mathbf{N}_{\mathrm{p_a}}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{a}})$ with $\boldsymbol{\Sigma}_{\mathrm{a}} = \boldsymbol{\rho}^{|j-k|}$, $1 \le j, k \le \mathrm{p_a}$, and from $\mathbf{X}_{\mathrm{b}} \sim \mathbf{N}_{\mathrm{p_b}}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{b}})$ with $\boldsymbol{\Sigma}_{\mathrm{b}} = \boldsymbol{\rho}^{|j-k|}$, $1 \le j, k \le \mathrm{p_b}$. Here, $\mathbf{X} = (\mathbf{X}_{\mathrm{a}}, \mathbf{X}_{\mathrm{b}})$ with $\mathrm{p} = \mathrm{p_a} + \mathrm{p_b}$. We take $\rho_{\mathrm{a}} = 0.9$ to allow for a high correlation among the half of explanatory variables, and $\rho_{\mathrm{b}} = 0.5$ for a middle correlation among the half of explanatory variables.

For the simulation scenarios with outliers, we think explanatory variables with BS and the first 10% of the observations is replaced by values of independent normal distributions $\mathbf{N}(50,1)$ for each variables, respectively. Therefore, adding leverage points to the data is finished. In order to add vertical outliers, the error terms for these outliers are replaced by values from $\mathbf{N}(20\hat{\sigma}_y, 1)$ instead of $\mathbf{N}(0,1)$, where $\hat{\sigma}_y$ corresponds to the estimated standard deviation of the clean response vector. With this way, in total, we consider 16 scenarios with 8 clean data and 8 contaminated data.

4.2. Design of Simulation Studies for High Dimensional Data

We take into consideration scenarios by means of generating high dimensional data sets with $(\mathrm{n}, \mathrm{p}) = \{(60,100), (50,1000), (40,200), (80,5000)\}$. The true response variable is computed as in Eq. (11), where the error term ε_i is distributed according to a standard normal distribution $\mathbf{N}(0,1)$, for $i = 1, \cdots \mathrm{n}$. Here, the data matrix $\mathbf{X} = (\mathbf{x_j})$ is obtained forming a block structure (BS) generating the observations of the blocks of design matrix $\mathbf{X}$ from $\mathbf{X}_{\mathrm{a}} \sim \mathbf{N}_{\mathrm{p_a}}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{a}})$ with $\boldsymbol{\Sigma}_{\mathrm{a}} = \boldsymbol{\rho}^{|j-k|}$, $1 \le j, k \le \mathrm{p_a}$, and from $\mathbf{X}_{\mathrm{b}} \sim \mathbf{N}_{\mathrm{p_b}}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathrm{b}})$ with $\boldsymbol{\Sigma}_{\mathrm{b}} = \boldsymbol{\rho}^{|j-k|}$, $1 \le j, k \le \mathrm{p_b}$. Here, $\mathbf{X} = (\mathbf{X}_{\mathrm{a}}, \mathbf{X}_{\mathrm{b}})$ with $\mathrm{p} = \mathrm{p_a} + \mathrm{p_b}$. We take $\rho_{\mathrm{b}} = 0.9$ to allow for a high correlation among the each of blocks. Additionally, we take $\rho_{\mathrm{a}} = 0.9$ and $\rho_{\mathrm{b}} = 0.5$ to have both high correlation and middle correlation among the different blocks.

For the simulation scenarios with outliers, we take the same idea as in low dimensional data. That is, in order to add leverage points, we replace the first 10% of the observations with the values of independent normal distributions $\mathbf{N}(50,1)$ for each variables, respectively. Then, the error terms for these outliers are replaced by values from $\mathbf{N}(20\hat{\sigma}_y, 1)$ instead of $\mathbf{N}(0,1)$, where $\hat{\sigma}_y$ represents the estimated standard deviation of the clean response vector. Therefore, in total, we consider 16 scenerios with 8 clean data and 8 contaminated data.

4.3. Performance Measures

The prediction ability of the discussed methods are done using training and test data sets generated according to the sampling schemes explained in Section 4.1 and 4.2. In order to fit the models, we used the training data and to evaluate the models, we used the test data. Each test data is generated outlier free.

As a performance measure, we consider the root mean squared prediction error (RMSPE) given by

$$\mathrm{RMSPE} = \frac{1}{\mathrm{m}} \sum_{j=1}^{m} \left(\sqrt{\frac{1}{n} \sum_{i=1}^{\mathrm{n}} (y_i - \hat{y}_i)^2} \right), \qquad (12)$$

where $\hat{y}_i = \mathrm{x}_i^T \hat{\xi}$ and m is the replication number.

Another performance measure is the accuracy of the coefficient estimate as a quality criterion, which is called ``bias" as follows

$$\mathrm{bias}(\hat{\boldsymbol{\xi}}) = \frac{1}{\mathrm{m}} \sum_{j=1}^{m} \left(\sqrt{\frac{1}{n} \sum_{i=1}^{\mathrm{n}} \left(\xi_i - \hat{\xi}_i \right)^2} \right), \qquad (13)$$

where m is the replication number.

These evaluation measures are calculated for the generated data in each of simulation replications separately, and then averaged in Tables. The replication number m is 100. The smaller the value for these criteria, the better the performance of the method.

4.4. Simulation Results

4.4.1. Simulation Results for Low Dimensional Data

Table 1 lists the results of RMSPE and bias for clean low dimensional data with different data structures explained in Subsect. 4.1. These results are sorted in ascending order to see changes according to the rate of p/n. For the lowest rate 2 with both multicollinearity levels, the pcaOLS has the best RMSPE and bias values. The good performance of the pcaOLS continues for the rate 2.5 with BS. The rpcaOLS outperforms with the same rate but multicollinearity level is highest. With the higher rate 5 and 25, OLS is better than the other competitors. Most of the results of the bias for compared methods are quite close to each other.

Table 1. The Root Mean Squared Prediction Error (RMSPE) and bias results for clean low dimensional data, averaged over $m = 100$ runs.

		The results of RMSPE and bias					
		OLS	MM	pcaOLS	pcaMM	rpcaOLS	rpcaMM
(100,50)	(BS)						
	RMSPE	2.08	2.71	**1.83**	1.90	1.89	1.99
	bias	0.12	0.15	**0.11**	0.12	**0.11**	**0.12**
	($\rho = 0.9$)						
	RMSPE	1.99	2.61	**1.49**	1.52	1.51	1.54
	bias	**0.11**	0.15	**0.11**	**0.11**	**0.11**	**0.11**
(50,20)	(BS)						
	RMSPE	1.70	2.02	**1.52**	1.58	1.54	1.60
	bias	**0.14**	0.16	**0.14**	0.15	0.15	0.16
	($\rho = 0.9$)						
	RMSPE	1.75	2.01	1.25	1.26	**1.24**	1.25
	bias	0.15	0.18	**0.13**	**0.13**	**0.13**	**0.13**
(500,100)	(BS)						
	RMSPE	**1.25**	1.28	1.42	1.44	1.44	1.45
	bias	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**
	($\rho = 0.9$)						
	RMSPE	**1.25**	1.29	1.53	1.54	1.54	1.55
	bias	**0.04**	0.05	0.05	0.05	0.05	0.05
(500,20)	(BS)						
	RMSPE	**1.05**	**1.05**	10.8	1.08	1.08	1.08
	bias	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**
	($\rho = 0.9$)						
	RMSPE	**1.03**	1.04	1.09	1.09	1.09	1.09
	bias	**0.04**	**0.04**	**0.04**	0.05	**0.04**	0.05

Table 2 shows the results of RMSPE and bias for contaminated low dimensional data with different structures explained in Subsect. 4.1 We sort the results in ascending order of the rate p/n. First three level of p/n., which are $2, 2.5$ and 5, gives the best performance with pcaMM for both BS and highest level of multicollinearity. The last scenario with the highest rate shows different results. In this case, classical MM estimator performs better than other competitors.

Table 2. The Root Mean Squared Prediction Error (RMSPE) and bias results for contaminated low dimensional data, averaged over m = 100 runs.

		The results of RMSPE and bias					
		OLS	MM	pcaOLS	pcaMM	rpcaOLS	rpcaMM
(100,50)	(BS)						
	RMSPE	14.08	14.14	9.60	**1.83**	9.71	1.87
	bias	0.35	0.35	0.29	**0.10**	0.32	0.11
	($\rho = 0.9$)						
	RMSPE	61.92	66.02	53.83	**1.61**	54.19	1.61
	bias	0.57	0.63	0.60	**0.10**	0.62	0.11
(50,20)	(BS)						
	RMSPE	7.09	1.93	5.18	**1.54**	5.26	1.56
	bias	0.31	0.18	0.29	**0.16**	0.30	0.17
	($\rho = 0.9$)						
	RMSPE	16.40	2.01	13.25	**1.33**	13.35	1.35
	bias	0.53	0.18	0.50	**0.15**	0.50	**0.15**
(500,100)	(BS)						
	RMSPE	7.31	1.96	5.36	**1.62**	5.43	1.60
	bias	0.29	0.17	0.27	**0.16**	0.28	**0.16**
	($\rho = 0.9$)						
	RMSPE	74.51	67.71	123.78	**4.19**	123.77	4.24
	bias	0.33	0.31	0.42	**0.04**	0.43	0.05
(500,20)	(BS)						
	RMSPE	4.16	**10.5**	4.16	1.08	4.16	1.08
	bias	8.76	**0.04**	0.08	**0.04**	0.08	**0.04**
	($\rho = 0.9$)						
	RMSPE	8.76	**1.05**	10.02	1.09	10.03	1.09
	bias	0.10	**0.04**	0.11	**0.04**	0.11	**0.04**

4.4.2. Simulation Results for High Dimensional Data

Table 3 summaries the results of RMSPE and bias belong to clean high dimensional data with different structures explained in Subsect. 4.2. These results are sorted in descending order for the rate of p/n. Because of the computational problems, we exclude the classical methods OLS and MM estimators for high dimensional data in this part. Table 3 shows that the pcaOLS outperforms in all cases.

Table 3. The Root Mean Squared Prediction Error (RMSPE) and bias results for clean high dimensional data, averaged over m = 100 runs.

		The results of RMSPE and bias			
		pcaOLS	pcaMM	rpcaOLS	rpcaMM
(60,100)	(BS)				
	RMSPE	**5.05**	5.15	13.75	18.93
	bias	**0.54**	0.55	1.12	2.08
	(BS $\rho = 0.9$)				
	RMSPE	**1.84**	1.91	9.68	13.69
	bias	**0.17**	0.18	0.88	1.58
(40,200)	(BS)				
	RMSPE	**19.09**	20.12	32.39	43.95
	bias	**2.55**	2.63	3.91	8.42
	(BS $\rho = 0.9$)				
	RMSPE	**9.14**	9.39	25.39	42.89
	bias	**1.35**	1.44	3.90	6.89
(50,1000)	(BS)				
	RMSPE	**96.66**	96.87	100.80	105.82
	bias	**10.79**	10.82	11.44	15.82
	(BS $\rho = 0.9$)				
	RMSPE	**128.12**	128.26	131.30	136.71
	bias	**16.25**	16.47	16.74	20.40
(80,5000)	(BS)				
	RMSPE	**234.76**	234.84	235.87	241.27
	bias	20.29	20.75	**20.20**	28.09
	(BS $\rho = 0.9$)				
	RMSPE	**305.29**	305.72	306.12	312.25
	bias	**28.34**	30.15	28.63	32.29

Table 4 shows the results of RMSPE and bias for contaminated high dimensional data with different structures explained in Subsect. 4.2. With the rate of $p/n = 0.6, 0.2$ ve 0.05, the pcaMM performs better than other alternatives. In the last situation, which corresponds to quite high number of variables and low sample size with the rate 0.016, the pcaMM outperforms under the RMSPE.

Table 4. The Root Mean Squared Prediction Error (RMSPE) and bias results for contaminated high dimensional data, averaged over $m = 100$ runs.

		The results of RMSPE and bias			
		pcaOLS	pcaMM	rpcaOLS	rpcaMM
(60,100)	(BS)				
	RMSPE	3.71	**2.21**	11.89	13.69
	bias	0.36	**0.23**	2.94	2.78
	(BS $\rho = 0.9$)				
	RMSPE	7.44	**1.69**	11.31	9.49
	bias	0.78	**0.18**	2.24	1.92
(40,200)	(BS)				
	RMSPE	4.78	**3.81**	60.69	60.71
	bias	0.74	**0.56**	12.01	12.94
	(BS $\rho = 0.9$)				
	RMSPE	8.72	**2.78**	22.07	39.14
	bias	1.25	**0.36**	7.43	8.80
(50,1000)	(BS)				
	RMSPE	9.74	**9.23**	355.16	126.16
	bias	1.21	**1.11**	66.71	34.51
	(BS $\rho = 0.9$)				
	RMSPE	11.66	**8.94**	162.81	185.32
	bias	1.42	**1.01**	50.45	55.89
(80,5000)	(BS)				
	RMSPE	**305.29**	305.72	306.12	312.25
	bias	**28.34**	30.15	28.63	32.29
	(BS $\rho = 0.9$)				
	RMSPE	21.50	21.00	3075.10	326.60
	bias	**2.22**	2.37	572.25	46.79

Note that the evaluation measures of compared methods are inflating with increasing rate of p/n. Moreover, the results of rpcaOLS and rpcaMM are inflating more than the other competitors pcaOLS and pcaMM. Although RPCA based methods mostly yield such poor results, PCA based results do as well. For example, the RMSPE results with $p/n = 0.016$ (BS), the pcaOLS outperforms with 305.29. The reason can be the poor results of the PCA and RPCA spaces for these specific data structures.

5. Real Data Applications

Although PCA based regression methods are of great use for high dimensionality problems, they can already be useful when applied to low dimensional problems. Therefore, the first and second examples are dedicated to show the performances of the discussed methods on low dimensional data sets. The third and fourth real data examples focus on the high dimensional data sets. In the matter of prediction performance for each real data example, the trimmed root mean squared prediction error (trimmed RMSPE) is computed using leave-one-out cross-validation (CV). At the beginning of the analysis for each data set, the number of component is determined by GCV as described in Subsection 3.1.

5.1. Prostate Cancer Data

Firstly, we consider the Prostate Cancer data set in order to show the performances of the discussed methods on a real data example. This data set is studied by (Stamey et al., 1989), analyzed by (Friedman et al., 2001) for prediction and available in R (R Foundation for Statistical Computing, 2010) It includes $n = 97$ observations and 9 variables. We determine the *lbph*, which corresponds to 4th variable, as a response variable and it represents the log of the amount of benign prostatic hyperplasia. Therefore, the number of explanatory variables is 8.

Table 5. Prostate Cancer data: number of components is determined as $k = 2$ by GCV, and trimmed root mean squared prediction error from leave-one-out cross validation of the optimal models.

Method	Trimmed RMSPE
OLS	0.8538
MM	**0.7399**
pcaOLS	1.3189
pcaMM	1.3081
rpcaOLS	1.2978
rpcaMM	1.2855

Note that the prostate cancer data set has quite low number of explanatory variables, and the determined number of PCs by GCV is 2. Table 5 displays trimmed RMSPE values for each method. To the best of our knowledge, there is no evidence that the prostate cancer data has any outliers. Table 5 shows that the classical MM estimator performs better than other competitors, and the classical OLS estimator follows it.

5.2. Forest Fires Data

For an illustration on a real data example with low dimensionality we consider the Forest Fires data with the number of observations $n = 517$ and the number of variables $p = 13$. This data set is recently gathered from the northeast region of Portugal and includes spatial, temporal, components from the Canadian Fire Weather Index (FWI) and four weather conditions. This problem was modeled as a regression problem aiming to estimate the burned area. Therefore, 13th variable, *area*, is determined as a response variable. This data set is available in webpage http://archive.ics.uci.edu/ml/datasets/\\Forest+Fires.

Table 6. Forest Fires data: number of components is determined as $k = 2$ by GCV, and trimmed root mean squared prediction error from leave-one-out cross validation of the optimal models.

Method	Trimmed RMSPE
OLS	120.9887
MM	**3.2184**
pcaOLS	103.5546
pcaMM	3.3139
rpcaOLS	109.9360
rpcaMM	3.3150

Table 6 represents the results of trimmed RMSPE of compared methods based on leave-one-out CV. The number of PCs is computed by GCV as 2. The OLS, pcaOLS and rpcaOLS perform quite poor. One can say because of the affect of the outliers in data set, they yield such violated results. On the other hand, the classical MM estimator outperform, followed by pcaMM and rpcaMM.

5.3. Glass Spectra Data

First example for high dimensional data is the archaeological glass vessels data, which is analyzed by (Janssens et al., 1998), from the 16^{th} and 17^{th} century. The number of glass vessels is$n = 180$ and each of these glass vessels was analyzed by an electron-probe X-ray microanalysis (EXPMA) leading to $p = 1920$ spectra for each vessel. The idea of taking the spectra in the range 15 to 500, which have highest frequencies, instead of taking all variables of size $p = 1920$ is used as in (Maronna, 2011). The resulting data set has $n = 180$ observations and $p = 486$ variables which have high multicollinearity. As response variable, we choose the oxide lead (PbO).

Table 7. Glass Spectra data: number of components is determined as $k = 78$ by GCV, and trimmed root mean squared prediction error from leave-one-out cross validation of the optimal models.

Method	Trimmed RMSPE
pcaOLS	0.0091
pcaMM	**0.0077**
rpcaOLS	0.0116
rpcaMM	0.0117

Table 7 displays the performances of the discussed methods on the Glass Vessels data set. The number of PCs computed by GCV is 78. The pcaMM estimator yields a best performance, followed by the pcaOLS. The rpcaMM shows poorest performance.

5.4. NCI Data

The prediction ability of the discussed methods is done using training and test data sets generated according to the sampling schemes explained in Section 4.1 and 4.2. In order to fit the models, we used the training data and to evaluate the models, we used the test data. Each test data is generated

outlier free. The other example for high dimensional data is the cancer data, which measures 60 human cancer cell lines, from the National Cancer Institute (NCI). This data set is available in webpage http://discover.nci.nih.gov/cellminer/. It is known that the 40th observation includes all missing values. Therefore, we exclude it from calculation and the number of observations is $n = 59$. The gene expression data is obtained with an Affymetrix HG-U133A chip and normalized using the GCRMA method. The resulting data includes $p = 22.283$ predictors. As a response variable we choose one of the expression variables, which is ADPRT-6, which corresponds to 4th of them. In this data, 13th observation determined as an outlier (Alfons et al., 2013).

In Table 8, we show the performances of the compared methods on a quite high dimensional data set. The number of PCs is determined as 9 at the beginning of the computation. Results show that the pcaMM outperforms for this data set and followed by pcaOLS.

Table 8. NCI data: number of components is determined as $k = 13$ by GCV, and trimmed root mean squared prediction error from leave-one-out cross validation of the optimal models.

Method	**Trimmed RMSPE**
pcaOLS	1.8590
pcaMM	**1.2094**
rpcaOLS	5.6125
rpcaMM	3.3342

6. Conclusions

PCA based methods are very attractive in many respects, especially for high dimensional data. By combining different scale measures as Projection Index, different types of PCA are introduced in literature. One important version of such different types is the RPCA. In this paper, we take into consideration RPCA using PP. Using the idea of PCR, we combine the PCA with MM besides the RPCA with OLS and MM, and thus we have obtained the pcaMM, the rpcaOLS and the rpcaMM estimators. These methods are compared throughout extensive simulation studies and real data examples. We consider the low/high dimensional and clean/contaminated data structures for different sizes in simulation studies. Four real data examples - two for low dimensional and two for high dimensional - are performed to illustrate the effectiveness of the methods.

For low dimensional data examples (both simulations and real data) we also consider the classical OLS and MM estimators for comparisons. While the real data examples are resulting with best performances of the pcaMM estimators, simulation studies are resulting mostly with best performances of the pcaOLS and pcaMM. When the number of observations increased, the classical OLS (clean cases) and MM (contaminated cases) estimators have become better in low dimensional data examples. High dimensional real data examples yield to best results with the pcaMM.

The performance of different estimation methods usually depends on the data set and therefore on the nature of data structure. Unluckily, there is no general rule or guideline for the choice of the method for both low and high dimensional data sets that is ideally convenient for the data at hand in the literature. But our extensive simulation studies and real data examples show using RPCA does not improve the results, but combining PCA with MM estimator provides very promising results. We deduce that the classical estimators, OLS and MM, can be preferred for low dimensional data. But for high dimensional data, the pcaMM works quite well.

References

Alfons, A., Croux, C. and Gelper, S., 2013. Sparse Least Trimmed Squares Regression for Analyzing High-Dimensional Large Data Sets. Annals of Applied Statistics, 7, 226-248.

Campbell, N.A., 1980. Robust Procedures in Multivariate Analysis: Robust Covariance Estimation. Applied Statistics, 29, 231-237.

Croux, C. and Haesbroeck, G., 2000. Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. Biometrika, 87, 603-618.

Croux, C. and Ruiz-Gazen, A., 2005. High Breakdown Estimators for Principal Components: The Projection-Pursuit Approach Revisited. Journal of Multivariate Analysis, 95, 206-226.

Croux, C., Filzmoser, P. and Oliveira, M.R., 2007. Algorithms for Projection-Pursuit Robust Principal Component Analysis. Chemometrics and Laboratory Systems, 87, 218-225.

Cui, H., He, X. and Ng, K.W., 2003. Aysptotic Distribution of Principal Components Based on Robust Dispersions. Biometrika, 90, 953-966.

Filzmoser, P., Fritz, H. and Kalcher, K., 2018. pcaPP: Robust PCA by Projection Pursuit. R Foundation for Statistical Computing, Vienna, Austria, http:// CRAN.R-project.org/package=pcaPP R Package Version 1.9-73.

Friedman, J., Hastie, T. and Tibshirani, R., 2001. The Elements of Statistical Learning. vol. 1: Berlin Springer Series in Statistics Springer.

Hotelling, H., 1957. The Relations of the Newer Multivariate Statistical Methods to Factor Analysis. British Journal of Mathematical and Statistical Psychology, 10, 69-79.

Huber, P.J., 1985. Projection Pursuits. The Annals of Statistics, 13, 435-525.

Hubert, M., Rousseeuw, P.J. and Verboven, S., 2002. A Fast Method for Robust Principal Components with Applications to Chemometrics. Chemometrics and Intelligent Laboratory Systems, 60, 101-111.

Hubert, M., Rousseeuw, P.J. and Vanden Branden, K., 2005. ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics, 47(1), 64-79.

Janssens, K., Deraedt, I., Freddy, A. and Veekman, J., 1998. Composition of 15-17th Century Archeological Glass Vessels Excavated in Antwerp, Belgium. Mikrochimica Acta, 15, 253-267.

Jeffers, J.N.R., 1967. Two Case Studies in the Application of Principal Component Analysis. Applied Statistics, 16, 225-236.

Johnson, R.A. and Wichern, D.W., 1998. Applied Multivariate Statistical Analysis, 4th edition, Prentice Hall, New York.

Li, G. and Chen, Z., 1985. Projection-Pursuit Approach to Robust Dispersion and Principal Components: Primary Theory and Monte Carlo. Journal of the American Statistical Association, 80, 759-766.

Kendall, M.G., 1957. A Course in Multivariate Anaysis, London: Griffin.

Maronna, R.A., 1976. Robust M-Estimators of Multivariate Location and Scatter. The Annals of Statistics, 4, 51-67.

Maronna, R.A., 2005. Principal Components and Orthogonal Regression Based on Robust Scales. Technometrics, 47, 264-273.

Maronna, R.A., Martin, R.D. and Yohai, V.J., 2006. Robust Statistics Theory and Methods, John Wiley and Sons Ltd.: England.

Maronna, R.A., 2011. Robust Ridge Regression for High-Dimensional Data. Technometrics, 53, 44-53.

R Development Core Team, 2013. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL http://www.R-pro ject.org

Stamey Thomas, A., Kabalin John, N., McNeal John, E. et al., 1989. Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate. II. Radical Prostatectomy Treated Patients. The Journal of Urology, 141, 1076-1083.

Varmuza, K. and Filzmoser, P. 2008. Multivariate Statistical Anaysis in Chemometrics, CRC Press: Taylor and Francis Group.

Xie, Y., Wang, Y., Liang, Y., Sun, L., Song, X. and Yu, R.Q., 1993. Robust Principal Component Analysis by Projection Pursuit. Chemometrics, 7, 527-541.

Yohai, V.J., 1987. High Breakdown Point and High Efficiency Robust Estimates for Regression. The Annals of Statistics, 15, 642-656.

Zou, H. and Hastie, T. 2005. Regularization and Variable Selection via the Elastic Net. Journal of Royal Statistical Society Series B, 67, 301-320.