Some robust variants of the principal components analysis

Shibzukhov Z.M.[1,2]

[1]Institute Mathematics and Informatics of Moscow Pedagogical State Univercity,
Moscow, Russia
[2]Institute Applied Mathematics and Automation KBSC RAS, Nalchik, Russia
intellimath@mail.ru

Abstract. One new robust variant of the formulation of the problem of searching for the principal components are considered. It's based on the application of differentiable estimates of the average value, insensitive to outliers. In principle, this approach makes it possible to overcome the influence of outliers in the task of searching for the principal components. The effectiveness of the proposed approach is clearly demonstrated on real data.

Keywords: Principal components analysis; Robust estimation; Data decomposition.

1 Introduction

The *principal component analysis* (PCA) is one of the methods of data decomposition. First classical PCA was originally considered as the problem of the best approximation of a finite set of points by straight lines and planes [1]. The presentation of training data in the basis of the *principal components* (PC) contributes to the reliability of the method of error back propagation in the procedure for finding the global minimum of the quadratic error function of a neural network of linear elements [2]. In machine learning, the PCA is often used as a method of reducing the dimension of input data. However, if part of the data is significantly distorted, then the results of the classical PCA may be inevitably significantly distorted.

An overview of classical and robust PCA can be found in [3]. A detailed overview of modern robust variants of PCA can be found in [4].

In this paper, new robust variant of the standard formulation of the problem of searching for the PC are proposed. It's based on the application of differentiable estimates of the average value, insensitive to outliers. The proposed approach, in principle, allows to overcome the impact of the outliers. The classical and new robust variant of the PCA are described below. The effectiveness of the proposed approach to the construction of robust variants of the principal component method in comparison with their classical versions is clearly demonstrated empirically on several real datasets. The theoretical foundations of the effectiveness of the proposed approach to PCA, shown here empirically, will be presented in future studies. The proposed robust version of the PCA is compared only with the classical version of the PCA. In this work, it was important, first of all, to show experimentally that the proposed robust variant of the classical formulation of the problem, going back to Pearson [1], makes it possible to effectively overcome the influence of outliers.

2 The classical approach

The problem of finding the center vector a_0 and the orthonormal basis $a_1, \dots, a_m$ of an m-dimensional hyperplane in $\mathbb{R}^n$ is solved, such that the sum of the squares of the Euclidean distances to it from the points $x_1, \dots, x_N$ minimal:

$$\sum_{k=1}^{N}\Big(\|x_k - a_0\|^2 - \sum_{j=1}^{m}(x_k - a_0, a_j)^2\Big) \to \min .$$

The solution of this problem is reduced to a chain of problems in which the vectors $a_0, a_1, \dots, a_m$ are sequentially located. This approach allows you to search for the principal components one by one.

The vector a_0 is found as a solution to the problem:

$$a_0 = \arg\min_{a\in\mathbb{R}^n} \sum_{k=1}^{N} \|x_k - a\|^2.$$

Its solution is a sample average: $a_0 = \frac{1}{N}\sum\limits_{k=1}^{N} x_k$.

After finding a_0, centering is performed: $x_k \to x_k - a_0, \quad k = 1, \dots, N$.

Then the vectors a_j $(j = 1, \dots, m)$ are sequentially searched for as a solution to the problem

$$a_j = \arg\min_{\|a\|=1} \frac{1}{N}\sum_{k=1}^{N}\left(\|x_k\|^2 - (a, x_k)^2\right).$$

After finding the next PC a_j, the transformation is performed:

$$x_k \to x_k - a_j(a_j, x_k).$$

The Lagrange multiplier method reduces our problem to the following problem

$$a_j, \lambda_j = \arg\min_{a,\lambda}\Big\{\frac{1}{N}\sum_{k=1}^{N}\left(\|x_k\|^2 - (a, x_k)^2\right) + \lambda\left(\|a\|^2 - 1\right)\Big\}. \tag{1}$$

Since

$$\frac{1}{N}\sum_{k=1}^{N}(a, x_k)^2 = \frac{1}{N}(Xa)^\top(Xa) = \frac{1}{N}a^\top(X^\top X)a = a^\top Sa,$$

where X — a matrix made up of vectors as rows $x_1, \dots, x_N$, $S = \frac{1}{N} X^\top X$ — the covariance matrix. Then, in accordance with the necessary condition of the extremum, a_j and λ_j satisfy the following equations:

$$Sa_j = \lambda_j a_j$$
$$\|a_j\| = 1.$$

That is, the solutions specify orthonormal eigenvectors and eigenvalues of the matrix S.

To search for a_j, one can apply an iterative procedure:

$$a^{t+1} = \frac{1}{\lambda^t}\left(Sa^t\right), \qquad \lambda^t = \frac{(a^t)^\top S a^t}{(a^t, a^t)}.$$

The outlier problem occurs when the distribution of values $\{\|x_k - a\|^2 \colon k = 1, \dots, N\}$, $\{\|x_k\|^2 - (a, x_k)^2 \colon k = 1, \dots, N\}$ contains outliers. The calculation of the empirical mean in this case gives a significantly biased value.

3 About robust variant of target functionals

In order to overcome the problem of outliers, it is proposed to use estimates of the average value, which are insensitive to outliers. And in order to be able to apply gradient minimization procedures, it is also proposed to use differentiable estimates of the average.

The proposed robust formulation of the problem involves replacing the minimization of the functional

$$\mathcal{Q}(a) = \frac{1}{N}\sum_{k=1}^{N} \varPhi_k(a)$$

with the minimization of the following functional

$$\mathcal{Q}_{\mathsf{M}}(a) = \mathsf{M}\{\varPhi_1(a), \dots, \varPhi_N(a)\}, \tag{2}$$

where $\mathsf{M}\{z_1, \dots, z_N\}$ is a differentiable estimation of the average value (M-estimator [5]), insensitive to outliers [6,7]. This formulation of the problem can significantly reduce the impact of outliers.

The problem of minimizing $\mathcal{Q}_{\mathsf{M}}$ is reduced to solving the equation

$$\sum_{k=1}^{N} \frac{\partial \mathsf{M}}{\partial z_k} \nabla \varPhi_k(a) = 0.$$

To solve it, one can use the method of iterative reweighting:

$$a^{t+1} = \arg\min \sum_{k=1}^{N} v_k^t \varPhi_k(a), \tag{3}$$

where

$$v_k^t = \frac{\partial \mathsf{M}\{\mathit{\Phi}_1(a^t), \ldots, \mathit{\Phi}_N(a^t)\}}{\partial z_k}. \tag{4}$$

In this paper, the following robust estimate of the mean is used — the censored arithmetic mean:

$$\mathsf{CP}_\alpha\{z_1, \ldots, z_N\} = \frac{1}{N}\sum_{k=1}^{N} \min(z_k, \bar{z}_\alpha),$$

where $0 < \alpha < 1$. It uses an estimate of the smoothed variant of the α-quantile. It's a example of M-mean:

$$\mathsf{M}_\rho\{z_1, \ldots, z_N\} = \arg\min_u \sum_{k=1}^{N} \rho(z_k - u),$$

where $\rho(r)$ is a positive strictly convex function with minimum $r(0) = 0$. If ρ is twice differentiable then

$$\frac{\partial \bar{z}_\rho}{\partial z_k} = \frac{\rho''(z_k - \bar{z}_\rho)}{\rho''(z_1 - \bar{z}_\alpha) + \cdots + \rho''(z_N - \bar{z}_\rho)}.$$

For smoothed α-quantile $\rho(r) = \rho_\alpha(r)$:

$$\rho_\alpha(r) = \begin{cases} (1-\alpha)\rho_\varepsilon(r), & \text{if } r < 0 \\ 0, & \text{if } r = 0 \\ \alpha\rho_\varepsilon(r), & \text{if } r > 0, \end{cases}$$

$\rho_\varepsilon(r)$ is such that 1) $\lim\limits_{\varepsilon\to 0} \rho_\varepsilon(r) = |r|$; 2) $\lim\limits_{\varepsilon\to 0} \rho'_\varepsilon(r) = \operatorname{sign} r$; 3) $\lim\limits_{\varepsilon\to 0} \rho''_\varepsilon(r) = \delta(r)$[1]. For example, $\rho_\varepsilon(r) = \sqrt{\varepsilon^2 + r^2}$ (it was used for calculations in the illustrative examples below).

At the same time,

$$\frac{\partial \mathsf{CM}_\alpha}{\partial z_k} = \begin{cases} \left(\dfrac{1}{M} + \dfrac{m}{M}\right)\dfrac{\partial \bar{z}_\alpha}{\partial z_k}, & \text{if } z_k < \bar{z}_\alpha \\ \dfrac{m}{M}\dfrac{\partial \bar{z}_\alpha}{\partial z_k}, & \text{if } z_k \geqslant \bar{z}_\alpha, \end{cases}$$

The following iterative procedure is used to find $\bar{z}_\alpha$:

$$u^{t+1} = \frac{\sum\limits_{k=1}^{N} \varphi(z_k - u^t) z_k}{\sum\limits_{k=1}^{N} \varphi(z_k - u^t)},$$

where $\varphi(r) = \rho'_\alpha(r)/r$.

[1] $\delta(r)$ is delta function of Dirac.

For comparison, here is another common use of M-estimators in the construction of $\mathcal{Q}$. The target functionality is defined as follows:

$$\mathcal{Q}(a) = \frac{1}{N}\sum_{k=1}^{N}\varrho(\varPhi_k(a)). \tag{5}$$

For example, in regression problems

$$\mathcal{Q}(a) = \frac{1}{N}\sum_{k=1}^{N}\varrho(f(x_k;a) - y_k).$$

Just such a method is used also in the robust version of PCA in [8].

Note that minimization of (5) is equivalent to minimization of Kolmogorov mean of $\varPhi_1(a), \dots, \varPhi_N(a)$:

$$\mathcal{Q}_\varrho(a) = \varrho^{-1}\Big(\frac{1}{N}\sum_{k=1}^{N}\varrho(\varPhi_k(a))\Big).$$

It is also an example of more general M-mean with $\rho(r) = r^2$:

$$\mathsf{M}_{\rho,\varrho}\{z_1, \dots, z_N\} = \varrho^{-1}\Big(\mathsf{M}_\rho\{\varrho(z_1), \dots, \varrho(z_N)\}\Big).$$

To find the optimal a^* by minimizing (5), the iterative reweighting method (3) is also used, which differs from our method proposed above in its method of recalculation of weights:

$$v_k^t = \frac{\varphi(\varPhi_k(a^t))}{\varphi(\varPhi_1(a^t)) + \cdots + \varphi(\varPhi_N(a^t))}. \tag{6}$$

The weights in both variants of the procedure decrease with the growth of $\varPhi_k(a^t)$. However, in (4), the weights depend on the magnitude of the deviation $\varPhi_k(a^t)$ from the robust estimate of the mean value (2) as opposed to (6). Here is a brief explanation: both $\varphi(r)$ and $\rho''(r)$ are positive and decrease toward to 0 as $r \to +\infty$. If the value of the average value is significantly separated from zero, then, as a rule,

$$\frac{\varphi(z_k)}{\varphi(z_1) + \cdots + \varphi(z_N)} > \frac{\rho''(z_k - \overline{z}_\rho)}{\rho''(z_1 - \overline{z}_\alpha) + \cdots + \rho''(z_N - \overline{z}_\rho)}$$

and therefore the weights of the outliers will have a smaller value in case of (4).

4 The robust PCA

The robust version of the statement of the search problem a_0 takes the form:

$$a_0 = \arg\min_{a\in\mathbb{R}^n} \mathsf{M}\big\{\|x_1 - a\|^2, \dots, \|x_N - a\|^2\big\}.$$

This problem is reduced to solving the equation

$$a = \sum_{k=1}^{N} \frac{\partial \mathsf{M}\{\|x_1 - a_0\|^2, \dots, \|x_N - a_0\|^2\}}{\partial z_k} x_k,$$

that can be solved using the following iterative procedure:

$$a^{t+1} = \sum_{k=1}^{N} v_k^t x_k,$$

where $$v_k^t = \frac{\partial \mathsf{M}\{\|x_1 - a^t\|^2, \dots, \|x_N - a^t\|^2\}}{\partial z_k}.$$

After finding a_0, centering is also performed: $x_k \to x_k - a_0, \quad k = 1, \dots, N$.

The robust version of the search problem a_j takes the following form:

$$a_j = \arg\min_{\|a\|=1} \mathsf{M}\left\{\|x_1\|^2 - (a, x_1)^2, \dots, \|x_N\|^2 - (a, x_N)^2\right\}.$$

Using the iterative reweighing method, its solution can also be reduced to a chain of tasks:

$$a_j^{t+1} = \arg\min_{\|a\|=1} \sum_{k=1}^{N} v_k^t \left(\|x_k\|^2 - (a, x_k)^2\right)$$

with the following point weights

$$v_k^t = \frac{\partial \mathsf{M}\left\{\|x_1\|^2 - (a^t, x_1)^2, \dots, \|x_N\|^2 - (a^t, x_N)^2\right\}}{\partial z_k}.$$

In all cases $v_1^t + \dots + v_N^t = 1$ by definition $\sum_{k=1}^{N} \frac{\partial \mathsf{M}\{z_1, \dots, z_N\}}{\partial z_k} = 1$.

This problem is a weighted version of the original search problem a_j within the framework of the classical formulation of the problem.

The covariance matrix takes the form:

$$S^t = X^\top \begin{pmatrix} v_1^t & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_N^t \end{pmatrix} X.$$

The vector a_j^{t+1} is the solution of the system:

$$\begin{gathered} S^t a = \lambda a \\ \|a\|^2 = 1, \end{gathered}$$

that is, it is an orthonormal eigenvector of the matrix S^t, and λ_j^{t+1} is its eigenvalue.

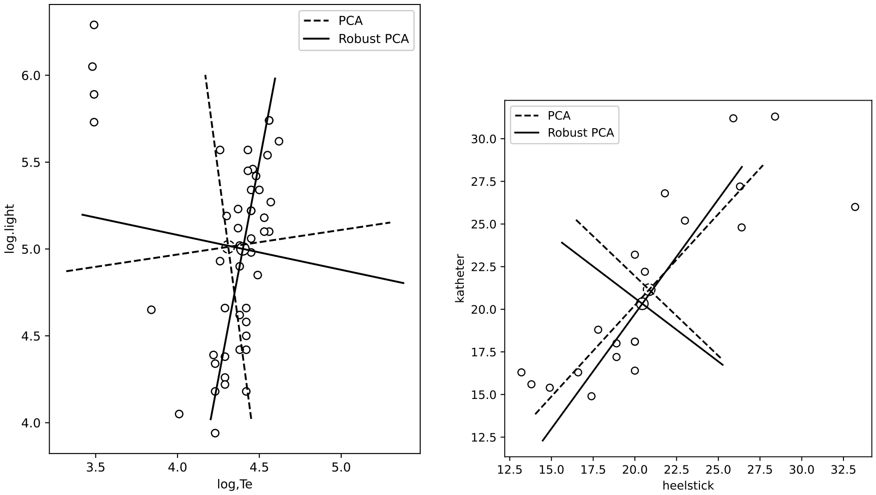
An iterative procedure is used to search for a_j:

$$a^{t+1} = \frac{1}{\lambda^t} \left(S^t a^t\right), \qquad \lambda^t = \frac{(a^t)^\top S^t a^t}{(a^t, a^t)}.$$

5 Experiments

For experimental confirmation of the effectiveness of the approach proposed here, examples of the application of classical and robust PCA for several data sets are considered. The robust approach proposed here is considered as a natural robust extension of the classical approach to the construction of PCA. Therefore, only the proposed robust approach and the classical approach are experimentally compared here in order to clearly show the ability of the proposed robust extension to overcome the outliers available in the data.

All calculations were performed using the open source library MLGRAD (`https://bitbucket.org/intellimath/mlgrad.git`) both for robust and classical variants of PCA. In the `/example` folder in the repository there are jupiter notebooks that contain calculations for the examples presented below. In some cases datasets was preprocessed using `scale` or `robust_cale` routines of the module `preprocessing` from SCIKIT-LEARN library (`https://scikit-learn.org`).



(a) PCs for STARSCYG dataset.

(b) PCs for KELLY1984 dataset.

Fig. 1: Experiments with 2 plain datasets.

Dataset starsCYG

Consider a data set for constructing a Hertzsprung-Russell diagram of the CYG_OB1 star cluster [5] (table 3), which describes the relationship between the logarithm of luminosity (log.light) and the logarithm of temperature (log.Te) of stars. In figure 1a there are 7 points that can be attributed to outliers. The application of the classical PCA gives the vectors of the PC rotated counterclockwise in the direction of outliers. The application of the robust PCA methods

($\alpha = 0.87$, $\varepsilon = 0.001$) makes it possible to find the unbiased position of the center and the PC that do not deviate under the influence of outliers.

Dataset Kelly1984

Consider a dataset Kelly1984 [9], which describes simultaneous pairs of measurements of serum kanamycin levels in blood samples drawn from 20 babies. In figure 1b there are some points that can be attributed to outliers. The application of the classical component method gives the vectors of the PC rotated counterclockwise in the direction of outliers. The application of the robust PCA ($\alpha = 0.8$, $\varepsilon = 0.001$) makes it possible to find the unbiased position of the center and the PC that do not deviate under the influence of outliers.

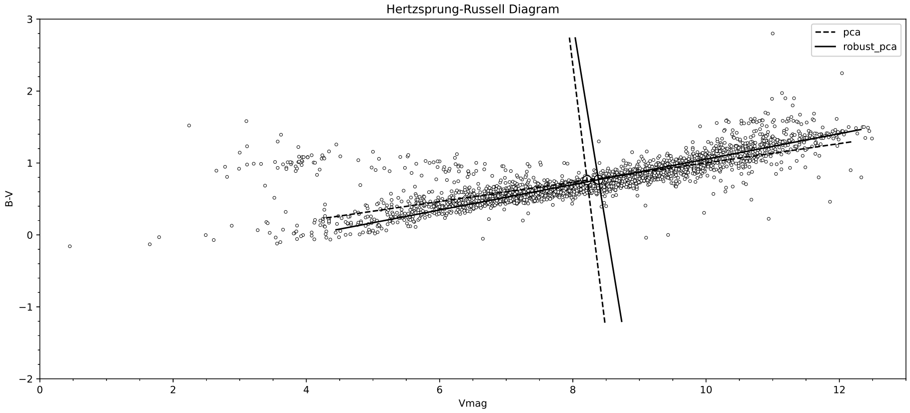


Fig. 2: The positions of the centers and PCs for HIP_STARS dataset.

Dataset HIP_stars

Consider a data set for plotting a chart for stars from a dataset [10]. In figure 2 shows a projection on a pair of Vmag and B-V parameters. The classical PCA gives offset PC for a given projection. Both proposed robust PCA ($\alpha = 0.95$, $\varepsilon = 0.001$) makes it possible to overcome the influence of outliers.

Dataset CigarettesSW

Consider panel data on cigarette consumption in 48 continental US states for 1985 – 1995 years [11]. The PCA is used for tabular data that covers 7 features (cpi, population, packs, income, tax, price, taxs; state, year are excluded). First data was preprocessed using `scale` routine of the module `preprocessing` from SCIKIT-LEARN library. Fig. 3 clearly shows that the data in projections on PC1×PC2×PC3, which are obtained on the basis of the robust PCA ($\alpha = 0.55$, $\varepsilon = 0.001$), have a more contrasting appearance: the data lines up along two straight lines. Only the first robust approach was applied here because the second approach does not show efficiency in this dataset.

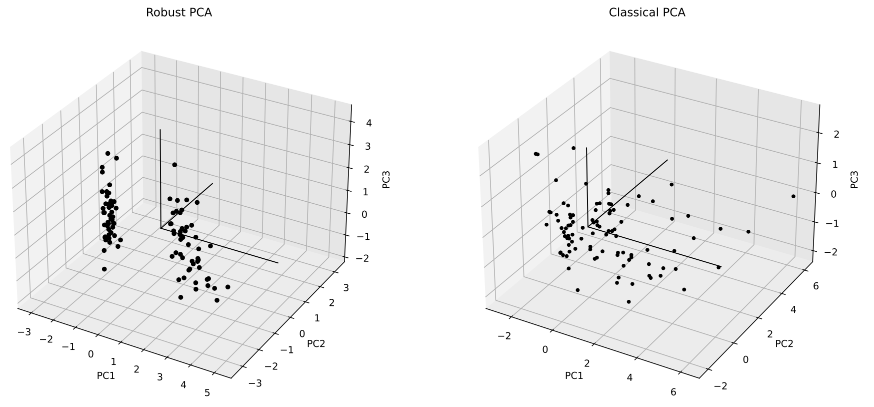


Fig. 3: Projections of the data from the CigarettesSW dataset on PC1×PC2×PC3.

It's easy to see that robust variant allow us to use the projection of data on the plain PC1×PC2×PC3 so that clustering linear regression method can be applied to distinguish the linearly shaped clusters.

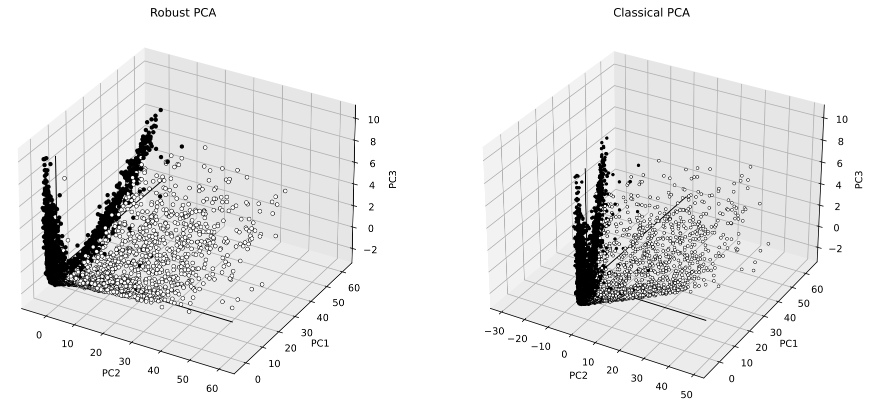


Fig. 4: Projections of the data from the HTRU2 dataset on PC1×PC2×PC3.

Dataset HTRU2

HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South). It contains 17898 rows, 8 columns and divided into 2 classes. The PCA used for data that covers 8 features. First data was preprocessed using `robust_scale` routine of the module `preprocessing` from SCIKIT-LEARN library. It is easy to see in the figure 4 that the projection of data on PC1×PC2×PC3 for the case of the classical PCA turns out to be rotated relative to PCs, while in the case of the robust PCA variant ($\alpha = 0.9$, $\varepsilon = 0.001$), the data is almost located along PCs. For example, the projection on the PC2 allows you to distinguish pulsars from others.

6 Conclusion

Robust variants of the formulation of the PCA problem, based on minimizing differentiable estimates of the mean, which can be significantly more resistant to outliers, allows us to find unbiased vectors of the PC. The α indicator in the smoothed quantile estimate roughly corresponds to the proportion of data that are not outliers (a more accurate value is found experimentally in its neighborhood). The first robust method makes it possible to identify outliers by analyzing the empirical distribution of distances to straight lines passing through the center a_0 along the vectors of the PC. The second robust method also makes it possible to identify outliers by analyzing the empirical distribution of Mahalanobis distances from the center to all points. It is also interesting because after finding a robust variant of the covariance matrix S, standard algorithms of PCA can be applied. Expreiments also was demonstrated that the first and the second robust approaches may have different levels of efficiency in same dataset with ouliers.

References

1. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. **2**, 559–572. (1901). doi: 10.1080/14786440109462720
2. Baldi P., Hornik R. Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima. Neural Networks. **2**, 53–58. (1989). doi: 10.1016/0893-6080(89)90014-2
3. Jolliffe I.T. Principal Component Analysis. Springer International Publishing. (2002). doi: 10.1007/b98835
4. Bouwmans T., Sobral A., Javed S., Jung S.K., Zahzah E.-H. Decomposition into Low-rank plus Additive Matrices for Background/Foreground Separation: A Review for a Comparative Evaluation with a Large-Scale Dataset. arXiv:1511.01245 (2015). doi: 10.1016/j.cosrev.2016.11.001
5. Rousseeuw P.J., Leroy A.M. Robust Regression and Outlier Detection. New York: John Wiley and Sons. 1987.
6. Shibzukhov Z.M. Machine Learning Based on the Principle of Minimizing Robust Mean Estimates. Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020. Springer International Publishing. 472–477. (2020). doi: 10.1007/978-3-030-65596-9_56
7. Shibzukhov Z.M. Minimizing Robust Estimates of Sums of Parameterized Functions Journal of Mathematical Sciences. Springer Science and Business Media LLC. **260**, 249–264. (2022). doi: 10.1007/s10958-022-05689-z

8. Polyak B. T., Khlebnikov M. V. Principle component analysis: robust versions. // Automation and Remote Control. 2017. Volume 78, Issue 3, PP. 490–506. DOI: https://doi.org/10.1134/S0005117917030092
9. Kelly B. The Influence Function in the Errors in Variables Problem. Annals of Statistics. **12(1)**, 87–100. (1984). doi: 10.1214/aos/1176346394
10. Dataset Hipparcos. `https://www.astrostatistics.psu.edu/datasets/HIP_star.html`
11. Dataset CigarettesSW: Cigarette Consumption Panel Data. `https://rdrr.io/rforge/gmm4/man/CigarettesSW.html`
12. Lyon R. HTRU2. UCI Machine Learning Repository. (2017). doi: 10.24432/C5DK6R