#45

BICA VPS 2023

One robust variant of the principal components analysis.

Z.M. Shibzukhov[1,2]. [1]Moscow Pedagogical State University, Moscow, Russia. [2]Moscow Institute of Physics Technologies, Dolgoprudnyi, Moscow Region, Russia.

SUMMARY

- New robust approach for principal components analysis.
- New robust variant of classical algorithm.
- Effectiveness of the proposed approach on some real datasets.

INTRODUCTION

Principal components coincide with straight lines passing through a given set of points to which the sum of the squared distances is minimal. If there are outliers in the data, this approach does not work.

The robust variant assumes a robust search for the center and a search for lines to which the robust estimate of the mean from the squares of the distances is minimal.

The robust search option is reduced to a chain of classical problems of searching for the main components, provided that weights are assigned to the starting points.

As a result of the search, the outliers will have fairly small values of weights compared to normal points.

APPROACH

Robust variant of finding centers takes the form:

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

It is reduced to solving the equation

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

where M is differential robust mean estimate.
It can be solved using the iterative procedure:

$$a^{t+1} = \sum_{k=1}^{N} v_k^t x_k, \quad v_k^t = \frac{\partial M\{\| x_1 - a^t \|^2, \ldots, \| x_N - a^t \|^2\}}{\partial z_k}.$$

After finding the center, centering is also performed:

$$x_k \to x_k - a_0, k = 1, \ldots, N.$$

The robust variant of finding PC takes the form:

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

It also boils down to solving of equations:

$$S_a a = \lambda a, \quad \| a \|^2 = 1, \qquad S_a = \sum_{k=1}^{N} v_k (x_k)^\top x_k$$

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

The search for the solution is based on an iterative procedure:

$$S^p a = \lambda a, \quad \| a \|^2 = 1, \qquad S^p = \sum_{k=1}^{N} v_k^p (x_k)^\top x_k$$

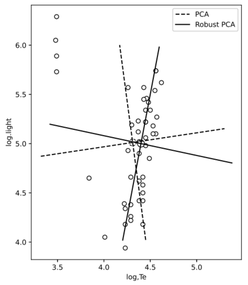
$$v_k^p = \frac{\partial M\left(\| x_1 \|^2 - (a^p, x_1)^2, \ldots, \| x_N \|^2 - (a^p, x_N)^2\right)}{\partial z_k}.$$

Solving matrix equations are based on an iterative procedure:

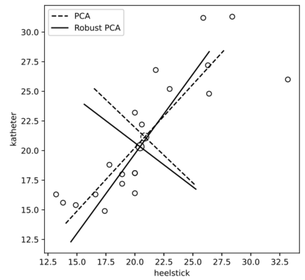
$$a^{t+1} = \frac{1}{\lambda^t}(S a^t), \lambda^t = \frac{(a^t)^\top S a^t}{(a^t, a^t)}.$$

RESULTS

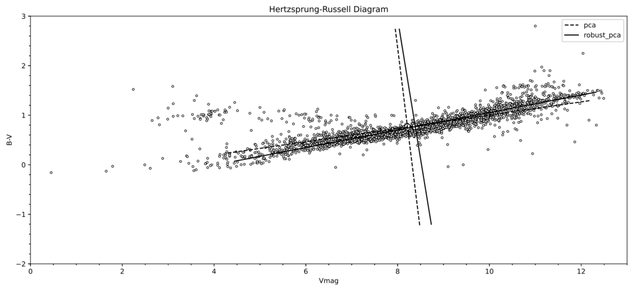
Dataset starsCYG



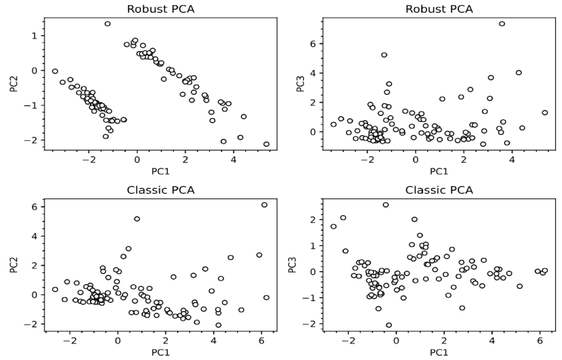
Dataset Kelly1984



Dataset HIP_stars



Dataset CigarettesSW



NOTES ON ROBUST AVERAGE

M is censored average:

$$CM_\alpha\{z_1, \ldots, z_N\} = \frac{1}{N} \sum_{k=1}^{N} \min(z_k, \acute{z}_\alpha), \quad 0 < \alpha < 1$$

$$\acute{z}_\alpha = \arg\min_u \sum_{k=1}^{N} \rho_\alpha(z_k - u),$$

$$\rho_\alpha(r) = \begin{cases} (1-\alpha)\rho(r), & \text{if } r < 0 \\ 0, & \text{if } r = 0 \\ \alpha\,\rho(r), & \text{if } r > 0, \end{cases} \quad \rho(r) = \sqrt{\varepsilon^2 + r^2}.$$

$$\frac{\partial CM_\alpha}{\partial z_k} = \begin{cases} \left(\frac{1}{M} + \frac{m}{M}\right)\frac{\partial \acute{z}_\alpha}{\partial z_k}, & \text{if } z_k < \acute{z}_\alpha \\ \frac{m}{M}\frac{\partial \overline{z}_\alpha}{\partial z_k}, & \text{if } z_k \geq \acute{z}_\alpha \end{cases}$$

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

$$u^{t+1} = \frac{\sum_{k=1}^{N} \varphi(z_k - u^t) z_k}{\sum_{k=1}^{N} \varphi(z_k - u^t)}, \quad \varphi(r) = \rho_\alpha'(r)/r$$

CONCLUSIONS

The robust version of the formulation of the PCA problem have possibility to find unbiased vectors of the PC.

This method also makes it possible to identify outliers by analyzing the empirical distribution of distances to straight lines passing through the center along the vectors of the PC.