Birla Institute of Technology and Science Pilani , KK Birla Goa Campus
Machine Learning BITS F464

ROBPCA : A New Approach to Robust Principal Component Analysis

Shikha Bhat (2019A7PS0063G)
Yash Trivedi (2019B4AA0834G)
Viraj Sharma (2020A7PS1011G)

Group 15

Acknowledgements

We would like to thank the instructor of our course Machine Learning (BITS F464) - Dr. Harikrishnan NB for teaching us the concepts of PCA thoroughly and giving us this opportunity to work on a project that gives us deeper insight into the topic. He has always encouraged us to give our all to learning and do better.

We would also like to thank the TAs for this course, Tanmay Devale, Ramanathan Rajaraman and Param Biyani for helping us out whenever needed.

Citation

Mia Hubert, Peter J Rousseeuw & Karlien Vanden Branden (2005) ROBPCA: A New Approach to Robust Principal Component Analysis, Technometrics, 47:1, 64-79, DOI: 10.1198/004017004000000563

Introduction

While solving any ML problem, the data that we use to analyze and feed to the model is of great importance. There are many situations in ML where we come across high dimensional data. In such situations we use **Principal Component Analysis to distill the variables down to their most important features. PCA allows us to represent the dataset as linear combinations of the original variables in a lower dimension.**

PCA often allows for interpretation and better understanding of the different sources of variation.

The Problem?

Outliers!

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations. Outliers increase the variance of the data.

The classic approach to PCA is **sensitive to outliers:** the principal components may then be distorted so as to fit the outlier, which leads to a **bad interpretation of the results**. Thus, the authors of this paper aim to introduce a new, robust method of PCA, which can address this problem. They call it ROBPCA.

Objectives

The paper aims to achieve these 2 objectives:

1. To develop a robust method through which we can accurately apply PCA to high dimensional data having outliers.

2. Get a diagnostic plot that can be used to detect and classify outliers accurately into 3 classes.

Methodology

Datasets

In the paper, the authors illustrate the ROBPCA method and the diagnostic plot on several real datasets.

Cars

Measurements of different cars - length, width, height etc

111 x 11 dimensions
2 principal components retained

Octane

Near-infrared (NIR) absorbance spectra wavelengths of gasoline samples with certain octane numbers.

39 x 226 dimensions
2 principal components retained
6 known outliers

Glass

EPXMA spectra over wavelengths collected on different glass samples

180 x 750 dimensions
3 principal components retained

Methodology

The ROBPCA Method

Previous efforts have been to replace the classical covariance matrix by robust covariance estimators but these cannot resist many outliers or are limited to small to moderate dimensions. The ROBPCA method attempts to combine the advantages of previous approaches - the projection pursuit technique is used for the **initial dimension reduction of the data,** and some ideas based on the **MCD estimator** (minimum covariance determinant) are then applied to this lower-dimensional data space.

1. First, the data are preprocessed such that the transformed data are lying in a subspace whose dimension is at most n − 1 using **Singular Value Decomposition.**
2. Next, a preliminary covariance matrix S0 is constructed and used for selecting the number of components k that will be retained in the sequel, yielding a k-dimensional subspace that fits the data well. **(Projection Pursuit Technique)**
3. Then the data points are projected on this subspace where their location and scatter matrix are robustly estimated **(MCD)**, from which its k nonzero eigenvalues l1,...,lk are computed. The corresponding eigenvectors are the k robust principal components.

Methodology

Equations

In the original space of dimension p, these k components span a k-dimensional subspace. Formally, writing the (column) eigenvectors next to one another yields the p × k matrix $P_{p,k}$ with orthogonal columns. The location estimate is denoted by the p-variate column vector μˆ and called the robust center. The scores are the entries of the n × k matrix T. Moreover, the k robust principal components generate a p × p robust scatter matrix S.

$$\mathbf{T}_{n,k} = (\mathbf{X}_{n,p} - \mathbf{1}_n\hat{\boldsymbol{\mu}}')\mathbf{P}_{p,k},$$

$$\mathbf{S} = \mathbf{P}_{p,k}\mathbf{L}_{k,k}\mathbf{P}'_{p,k},$$

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}},$$

$$OD_i = \|\mathbf{x}_i - \hat{\boldsymbol{\mu}} - \mathbf{P}_{p,k}\mathbf{t}'_i\|,$$

Methodology

Diagnostic Plots

Diagnostic plots help in distinguishing between regular observations and three types of outliers - **orthogonal outliers (5), good leverage points (1 and 4) and bad leverage points (2 and 3)** by plotting the orthogonal distance (OD) of the outliers from the PCA subspace vs the robust score distance (SD) of the observations.

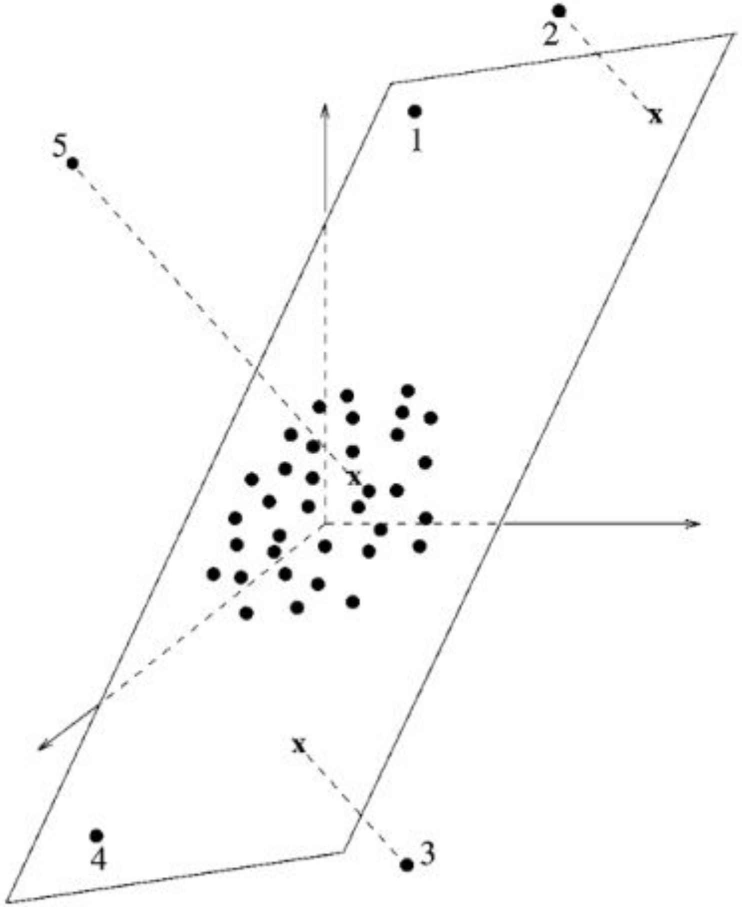
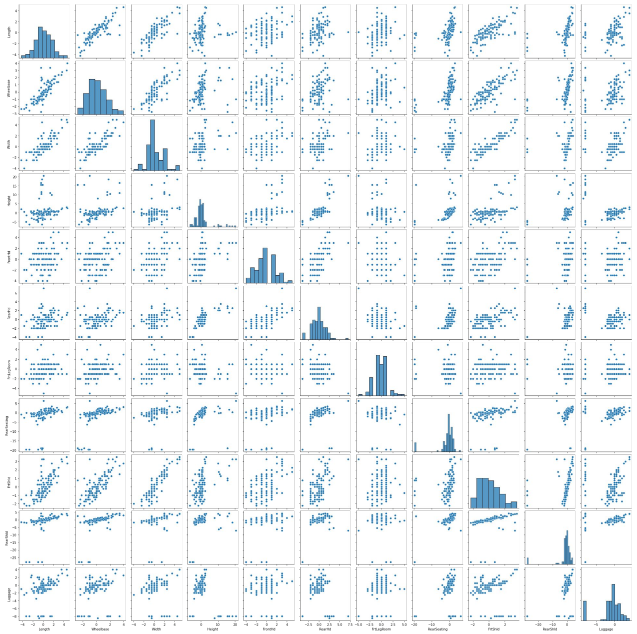


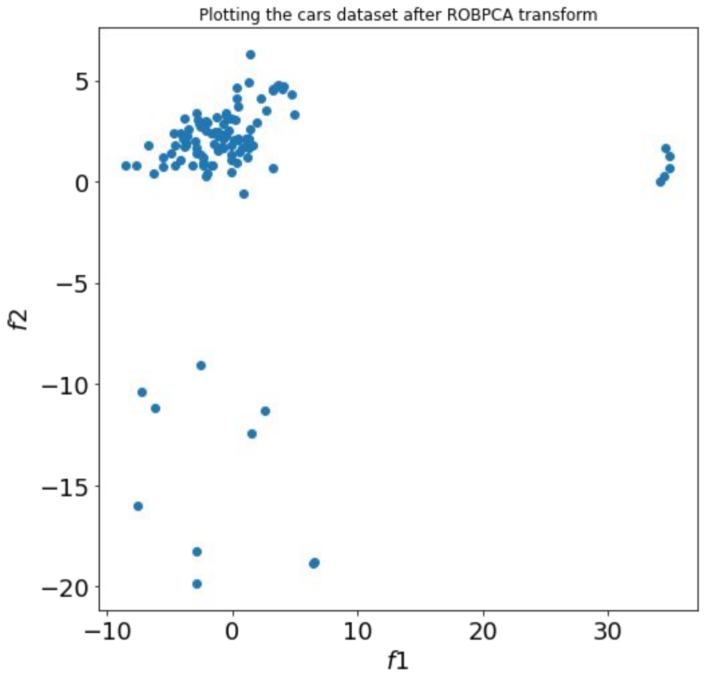
Figure 1. Different Types of Outliers When a Three-Dimensional Dataset Is Projected on a Robust Two-Dimensional PCA Subspace.

Experiments and Results

Cars Dataset (111 samples, 11 features)



Dataset after ROBPCA



Preliminary Analysis indicated that there are high correlations among the variables, hence PCA seems to be an appropriate method for finding the most important sources of variation in this dataset.

No clear outliers were known beforehand. We were able to successfully perform the ROBPCA method and **reduce the dimensionality of the data to 2.**

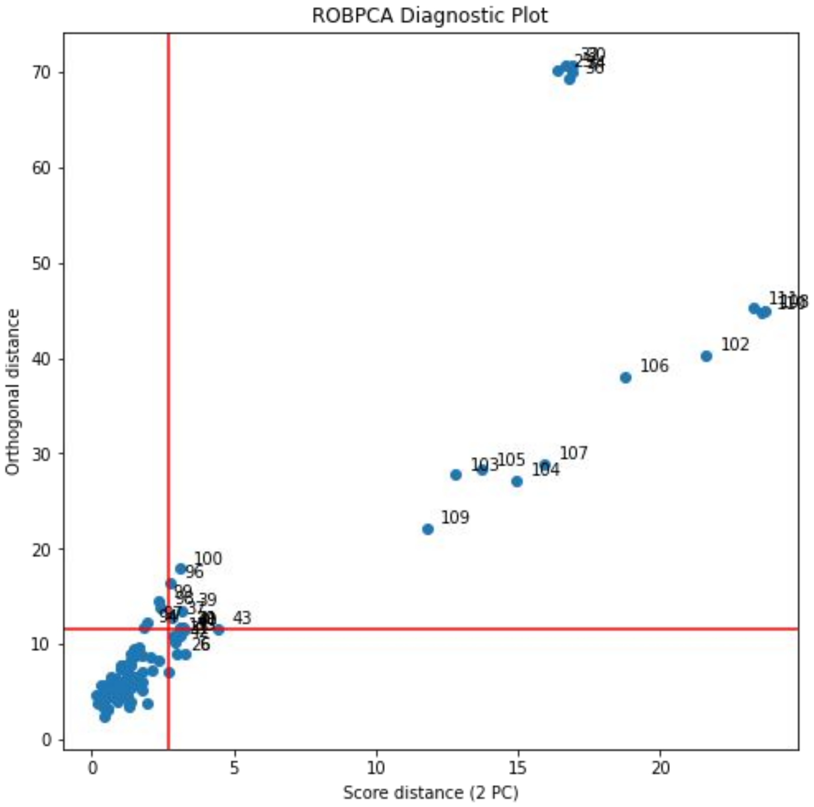
Cars Dataset (111 samples, 11 features)

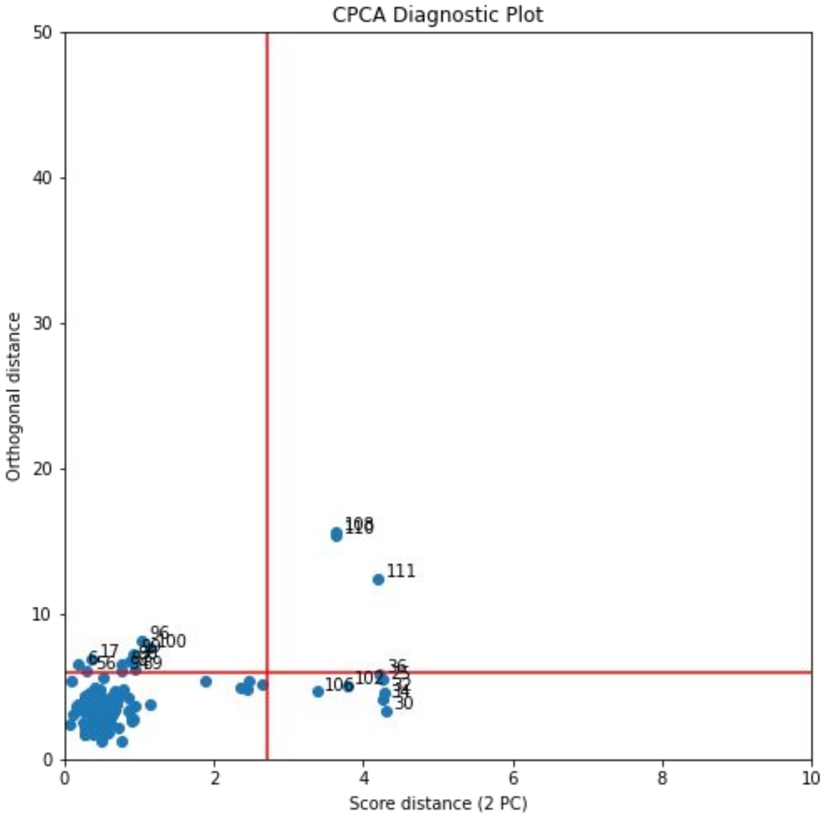
The resulting diagnostic plot is shown below, compared to the diagnostic plot we obtained for classical PCA approach. They are a little different from the paper's but that is expected because of the changes we made in the implementations (explained in challenges).

ROBPCA identified more outliers than CPCA.

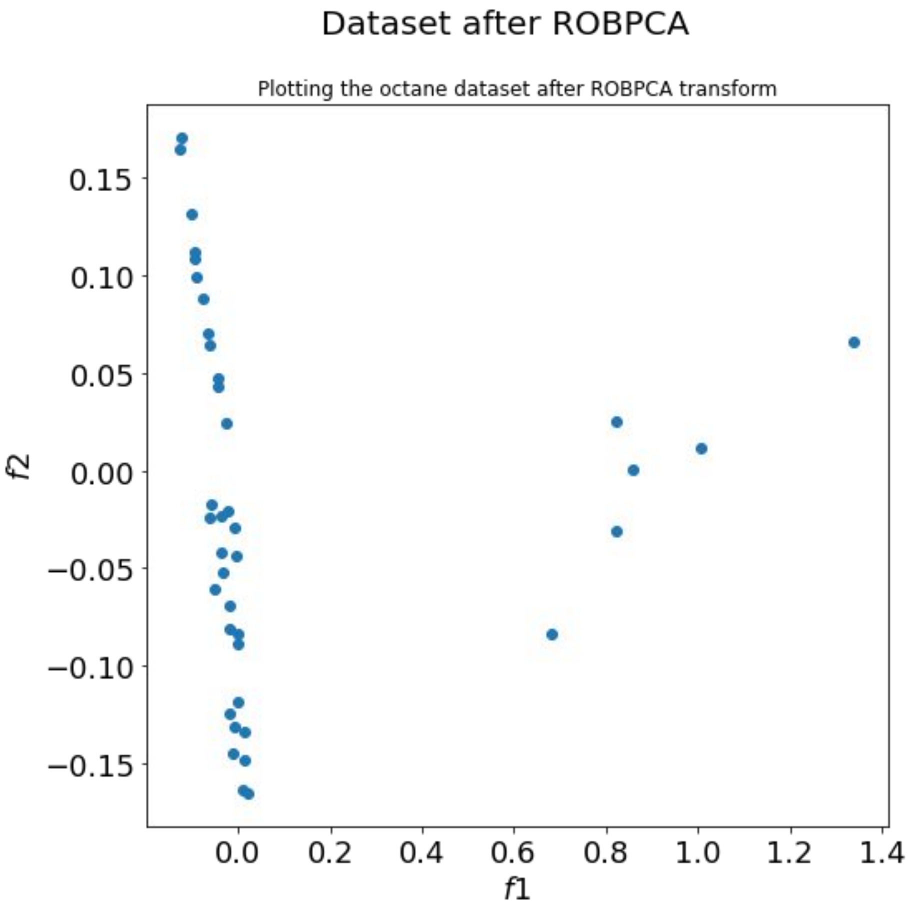
The most striking difference is that the group of bad leverage points from ROBPCA is converted into good leverage points in CPCA.

The subspace found by CPCA is attracted toward the bad leverage points.





Octane Dataset (39 samples, 226 features)

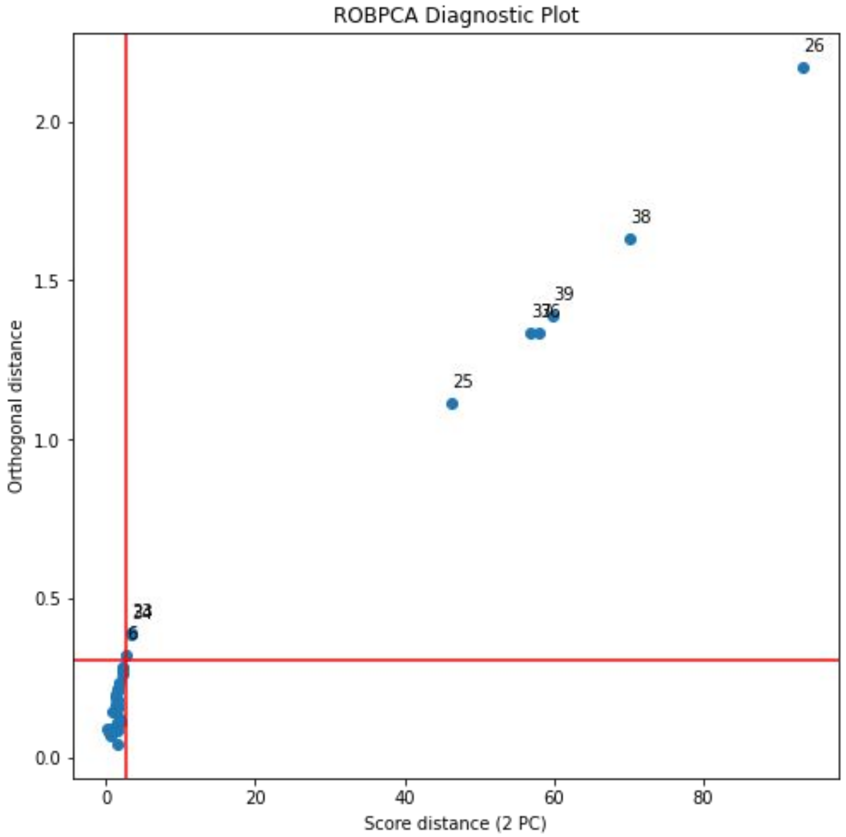


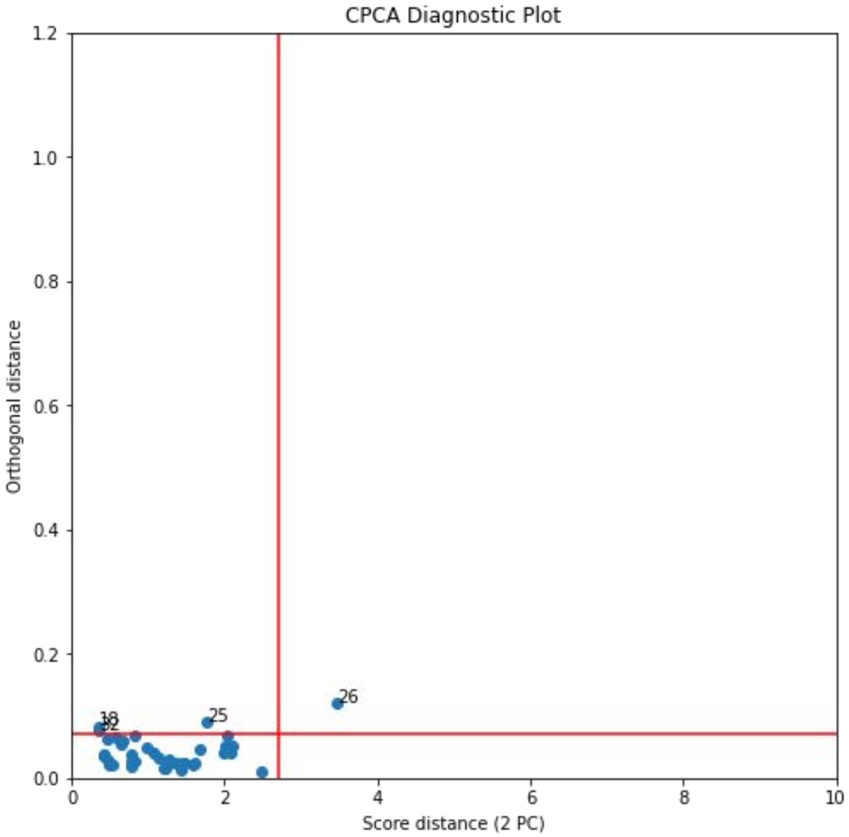
Our second experiment is with the octane dataset described by Esbensen, Schönkopf, and Midtgaard (1994). This dataset contains near-infrared (NIR) absorbance spectra over p = 226 wavelengths of n = 39 gasoline samples with certain octane numbers.

Outliers: In this dataset, **it is known that six of the samples (25, 26, and 36–39) contain added alcohol.**

No preliminary analysis was conducted since pairwise calculations with 226 features would take very long. We successfully executed the ROBPCA method on this dataset with 2 retained principal components, shown alongside.

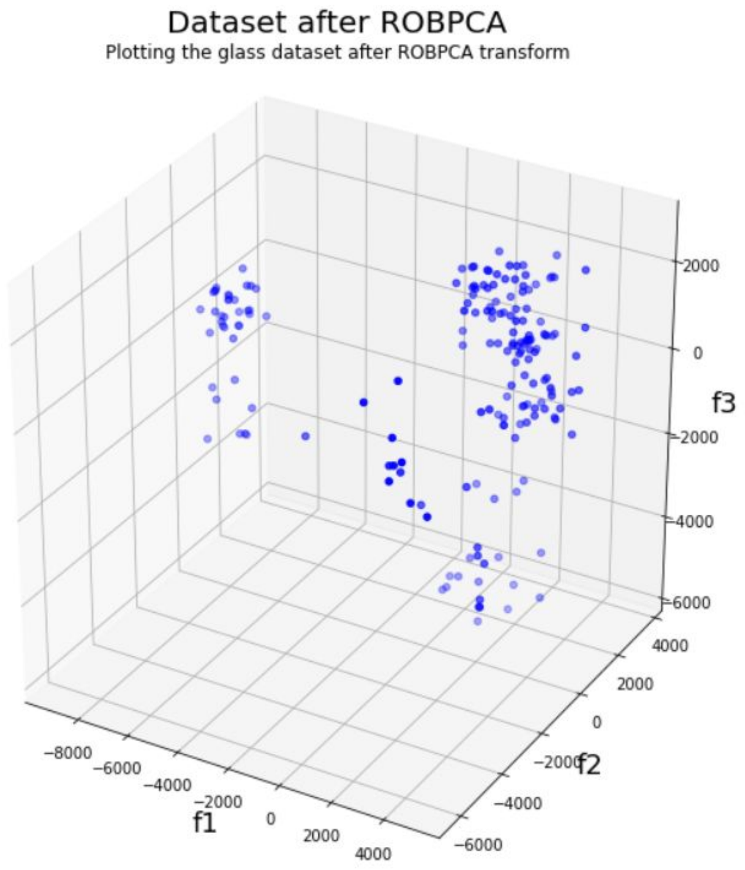
Octane Dataset (39 samples, 226 features)





The CPCA diagnostic plot shows that the classical analysis detects only the outlying spectrum 26 and barely 25, which does not stick out much above the border line. In contrast, we immediately clearly spot the six samples with added alcohol on the ROBPCA diagnostic plot.

Glass Spectra Dataset (180 samples, 750 features)

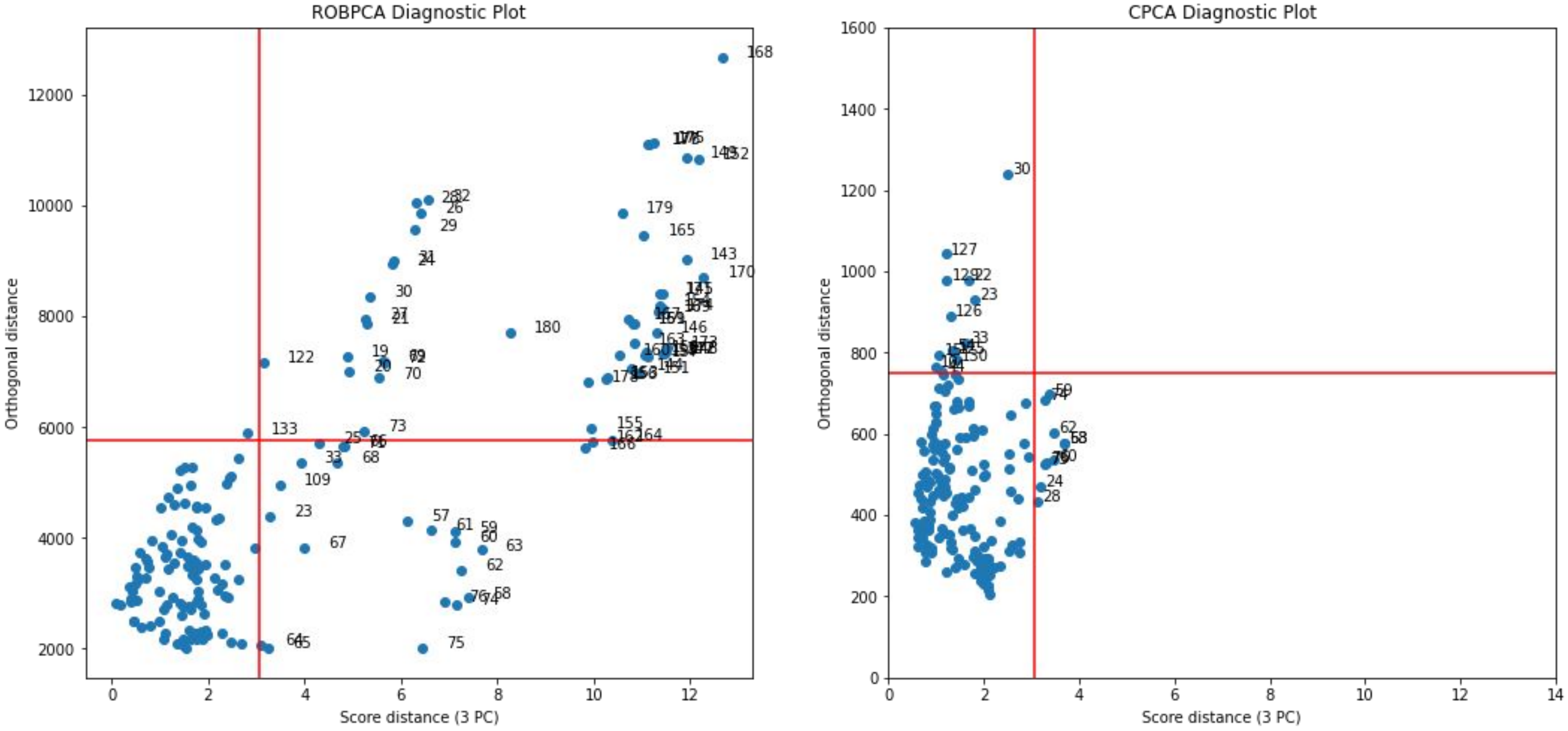


Our third experiment is with Glass Spectra dataset which consists of EPXMA spectra over 750 wavelengths collected on 180 glass samples.(Lemberge, De Raedt, Janssens, Wei, and Van Espen 2000).

It turned out that the window of the detector system had been cleaned before the last 38 spectra were measured. As a result, less radiation (X-rays) was absorbed and more could be detected, resulting in higher X-ray intensities. (57–63) and (74–76), are samples with high concentrations of calcic. (22, 23, and 30) indicate a larger concentration of phosphor.

No preliminary analysis was conducted since pairwise calculations with 750 features would take very long. We successfully executed the ROBPCA method on this dataset with **3 retained principal components,** shown alongside.

Glass Spectra Dataset (180 samples, 750 features)



From the classical diagnostic plot, we see that CPCA does not find important outliers. In contrast, the ROBPCA plot clearly distinguishes 143-179, 57-63 and 74-76 as outliers.

Discussions

The results we got show that **ROBPCA is a promising method that gives robust estimates even when data contains outliers.** The associated outlier maps are very useful to **visualize and classify the different outliers.** A side by side comparison of these plots for different PCA methods as shown in the paper show the superiority of ROBPCA, as it is able to identify the outliers which are not identified by CPCA . Our implementation was slightly different than the paper's but we were still able to get good results using ROBPCA. Overall, it was a successful endeavour.

Conclusion

In this study, the authors construct a fast and robust algorithm ROBPCA which can apply PCA on high dimensional data. They apply PP techniques and these results are used to project the observations on smaller dimension subspace. Within this subspace, they apply ideas of robust covariance estimation. The results were promising and definitely a step up from the CPCA approach. **The ROBPCA method thus opens a door to practical robust multivariate calibration and to the analysis of regression data with both outliers and multicollinearity.**

Challenges

We followed the methods as described in the Appendix of the paper. We made a few changes when we encountered problems -

1. To find the h “least outlying” data points, we could not understand exactly how they calculated the outlyingness score for each observation. They have used a variant of the Stahel–Donoho affine-invariant outlyingness, but it was very complex and we were not clear on what **B** was exactly. So we used the sklearn package **LocalOutlierFactor** to calculate the outlyingness score for each observation.
2. Since we changed that, the criteria for selecting k also had to be changed to make it equivalent to what the authors had obtained (so that we could compare our results). We **changed the threshold to 80%** for the cars dataset to get k = 2, and 95% for the glass dataset to get k=3, instead of 90%.

$$\mathrm{outl}_O(\mathbf{x}_i) = \max_{\mathbf{v}\in B} \frac{|\mathbf{x}_i'\mathbf{v} - t_{MCD}(\mathbf{x}_j'\mathbf{v})|}{s_{MCD}(\mathbf{x}_j'\mathbf{v})}.$$

$$\sum_{j=1}^{k} \tilde{l}_j \Big/ \sum_{j=1}^{r} \tilde{l}_j \approx 90\%,$$

Challenges

3. The authors slightly adapt the FAST–MCD algorithm of Rousseeuw and Van Driessen, but the variation was not easily implementable, so we resorted to directly using the **MCD estimator package provided by sklearn, which implements the standard Fast-MCD algorithm.**
4. The scoring matrix T for CPCA is not defined in the paper, and we implemented what we thought was right.
5. We implemented the ROBPCA and CPCA algorithm from scratch, but were not able to reproduce the exact diagnostic plots in terms of measurements made by the authors (the scale on x and y axis are different).

These changes affected the results as well, which is why our diagnostic plots are correct, but different from the paper.

Code

Our code can be found in the github repository
https://github.com/MLGroup15/ROBPCA

Drive folder link:
https://drive.google.com/drive/folders/1Tw-O5lpdLiXRSnPVlCSGm3JpTtiv-LKm?usp=share_link

References and Sources

1. Mia Hubert, Peter J Rousseeuw & Karlien Vanden Branden (2005) ROBPCA: A New Approach to Robust Principal Component Analysis, Technometrics, 47:1, 64-79, DOI: 10.1198/004017004000000563
2. Shlens, Jonathon. "A tutorial on principal component analysis." arXiv preprint arXiv:1404.1100 (2014).
3. Smith, Lindsay I.. “A tutorial on Principal Components Analysis.” (2002).
4. LIBRA: the Matlab Library for Robust Analysis (Verboven and Hubert 2004) https://github.com/duncombe/matlab/tree/master/LIBRA
5. Octane Data: https://www.impopen.com/software/octane-data-set
6. Glass Data: https://search.r-project.org/CRAN/refmans/cellWise/html/data_glass.html
7. Cars Data: https://cran.r-project.org/web/packages/msos/msos.pdf