Faculty of Applied Sciences

High-dimensional data analysis

MATH2021-1

Project 2:
Mixture detection

Authors :
Delcour Florian - s181063
Makedonsky Aliocha - s171197

Instructors :
Haesbroeck Gentiane

November 2021

First we plot all the variables depending one on the other, and we observe clearly the shifting on the three last variables, we have something linear for all the others and no difference in the two classes, but for the three last one there clearly are two separate groups, which is normal since for the group of initial outliers we create it by playing on the three last variables.

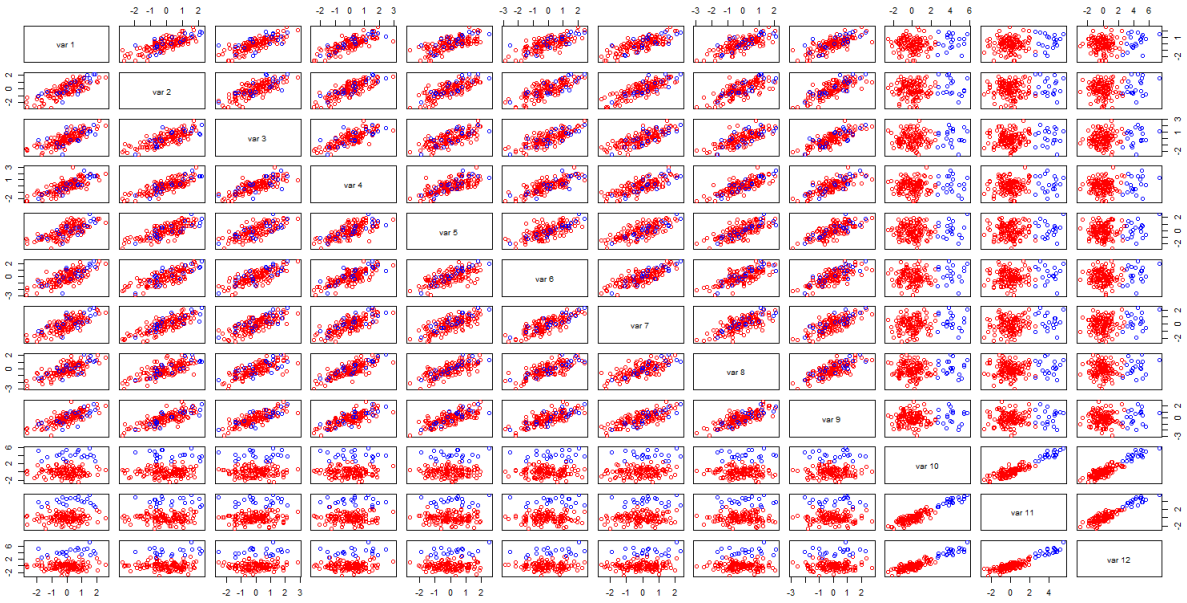

Figure 1: Distribution F

1 Outlier detection

To detect outlier, we will use the χ^2 cutoff which relies on the normality assumption of the data. This assumption is verified because distribution F is the sum of two normal distributions. For this project, we will use a confidence interval of 95%.

1. As in the first project, we will use the MCD estimator to compute the robust distances, where we keep 75% of the observations to compute the mean of the data set.
2. Something that is interesting to notice in the Table 1 is that PTP is way smaller for $\epsilon = 0.2$ than for $\epsilon = 0.1$. Based on mahalanobis distances formula,

$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

because there are more initial outliers in the second case, the mean is bigger, and thus the mahalanobis distances are smaller, so there are less outliers. PFP is always smaller than PTP, which is expected since the observations for PFP are the "good" observations and not the initial outliers.
The percentages are constantly higher for robust distances than for classical Mahalanobis distances, since the robust method excludes the worst values (75% coverage

Mean	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0.2$
Classical PTP	-	19.34%	8.33%
Robust PTP	-	68%	25.9%
Classical PFP	4.09%	3.15%	3.16%
Robust PFP	8.54%	6.34%	6.53%

Table 1: Average PTP and PFP for 500 simulations, in original space

here) before computing the mean, so it's lower, thus the distances are bigger, and consequently there are more outliers.

3. On Figure 2, we represented 3 boxplots corresponding to each value of ϵ.
We observe that for each value of ϵ, the PTP is much higher for the robust distances than for the classical distances, which is normal since there are more outliers with the robust method. It's a good thing since the initial outliers should be tagged as outliers, so higher the PTP, better the outlier detection technique.
But on the other hand, PFP is slightly higher for the robust distances than for the classical ones, which is explained by the same reason than for PTP. It means that more of the "good" observations are tagged as outliers, which is not a good thing.
But overall, the performances in detecting the true outliers justify the loss of accuracy in tagging false positives, so the robust distances method should be prefered.

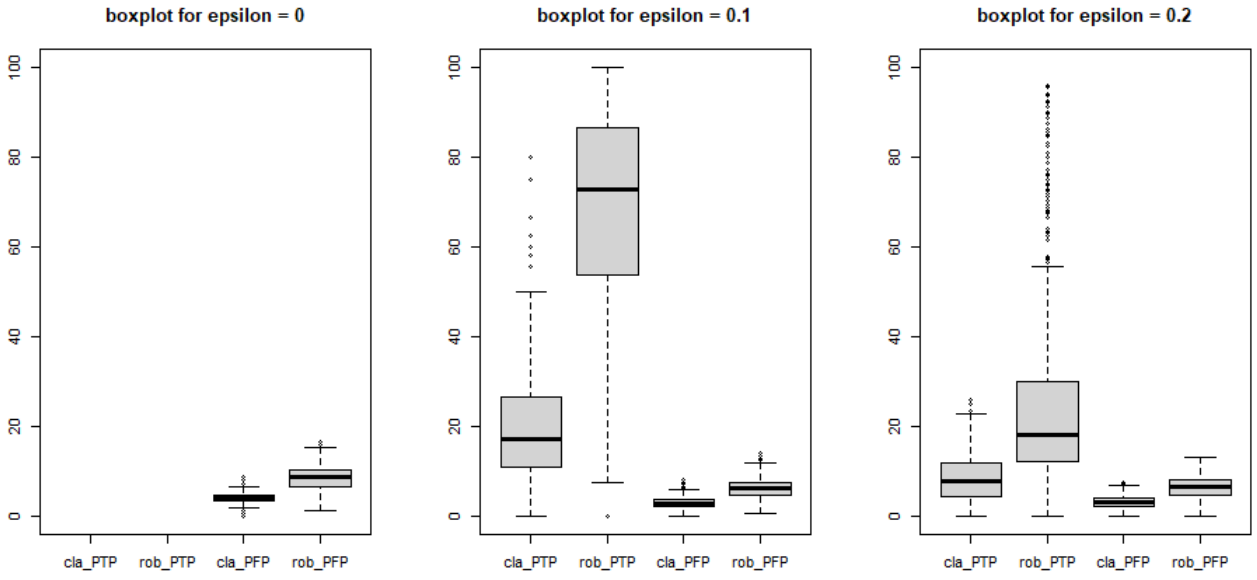

Figure 2: PTP and PFP boxplots with classical and robust detection, in original space

2 Outlier detection after dimension reduction

1. (a) First, we decided to make a 2D projection of our data because there are only two different classes which are coming from two distributions shifted from each other. Hence, a 2D projection is sufficient to visualize the clusters.
Some options are available for the `Rtsne`, among which the option `pca` which, if set to `TRUE`, will apply a PCA on the covariance matrix in order to keep the 50 (default) principal components (the components that explain the most the variance). All the dimensions are kept and this is simply a rotation of the initial data. Euclidean distances are preserved and so the results do not change a lot.
An other option is the `normalize` option which, if set to `TRUE`, centers and scales the data. It doesn't change the closeness or outlyingness of the neighbors so again the results do not change.
Having tried every combinations of those 2 options, we observed that the results were quite the same, keeping the same repartition. For the following, we will set both options to `FALSE`.
An option that influences the map is the `perplexity` as shown on Figure 3. As we increase this value, trends appear clearer, and we observe that the red points (which are the "good" samples) are separated from the blue points (which are the initial outliers), so there is a clear separation between the two groups of observations, even if some red points are close to the blue cloud, those are the outliers from the "good" observations. However, there is a limit to this value, which is set to

$$3 * \text{perplexity} < \text{nb\_samples} - 1$$

But taking 10 is clearly enough here. On Figure 4, the final 2D map is provided for all values of ϵ, with parameters `pca` and `standardize` set to `FALSE`.

(b) We performed the PCA on the covariance matrix because we wanted to minimize the loss of information between the two groups. In this part, the goal is to detect outliers which are observations coming from the second distribution. This second distribution is shifted with the first one and this shift is an important information to separate the groups. So we do not standardize the data to keep the difference in magnitude in order to better separate the two groups. Comparing the covariance matrix and the correlation matrix of distribution F is useful to see the loss of magnitude.

Applying PCA for $\epsilon = 0.2$, we obtain that the most important component explains 54.5% of the variance, and the second one 34.29%, for a total of 88.8% with just those two. The third one only explains 1.55% of the variance, which is nothing compared to the others and that's why we keep only two principal components. Results for $\epsilon = 0.1$ were quite similar and for $\epsilon = 0$, there is only one group so not very interesting.

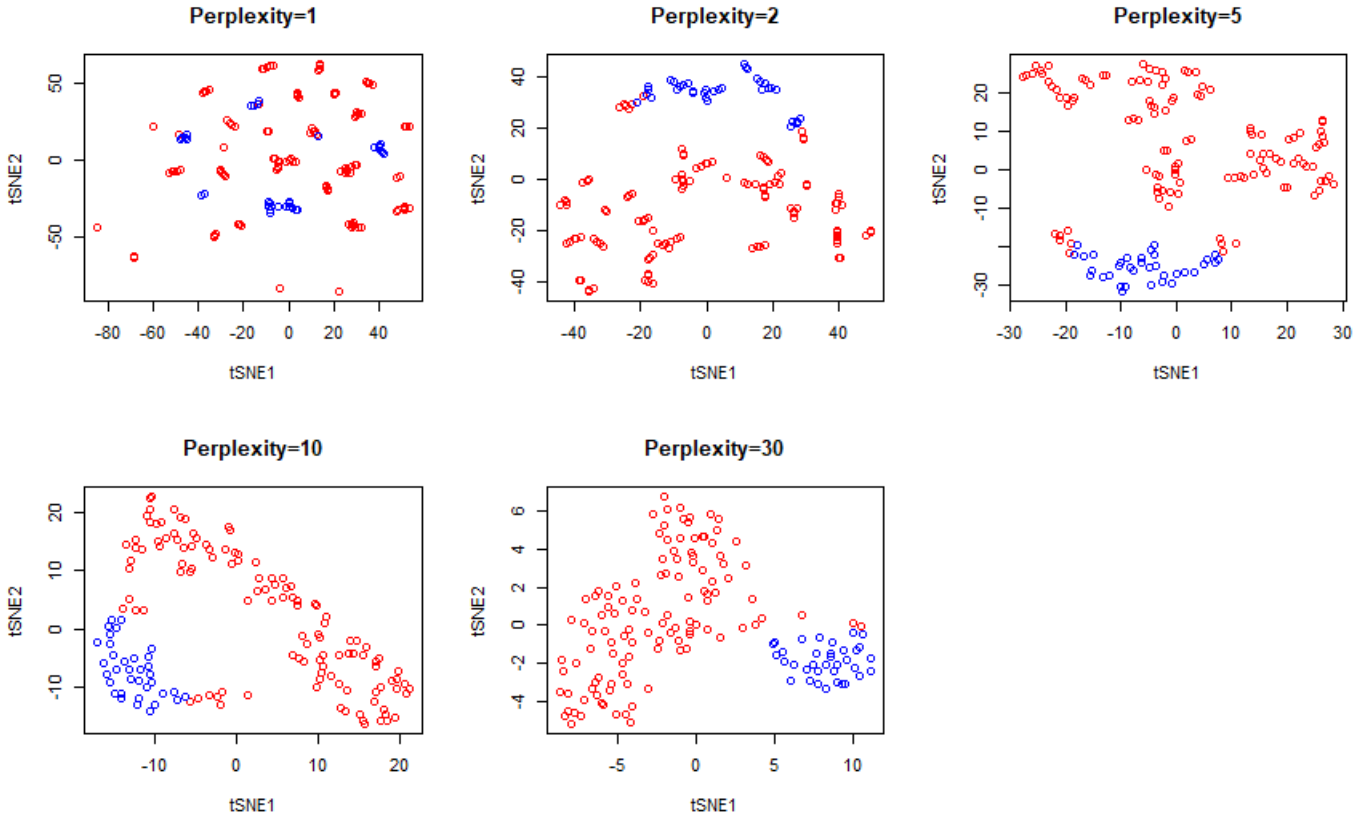

Figure 3: `tSNE` with different perplexities

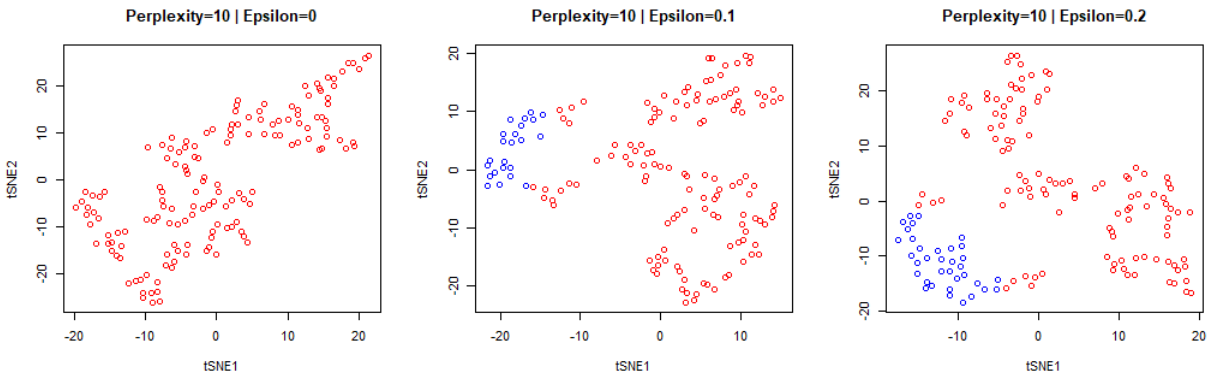

Figure 4: tSNE 2D map, `pca=FALSE` and `standardize=FALSE`

Plotting the scores of the data on those 2 principal components, we obtain the Figure 5.
Indeed we clearly observe two different groups, with the red samples being the "good" observations, and the blue one the initial outliers. Once more we observe some red samples close to the blue cloud, they are the outliers from the "good" observations. Unlike `tSNE`, we see that the blue dots are not con-

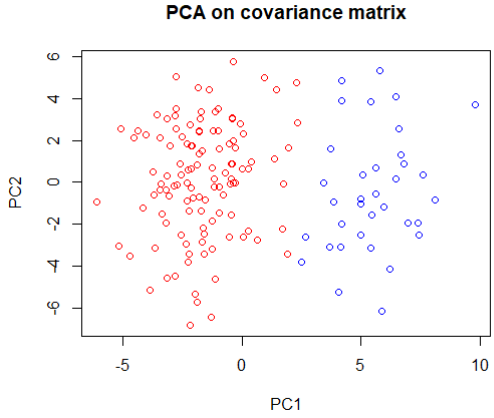

Figure 5: Two principal components, $\epsilon = 0.2$

centrated in a region.

(c) The aim of `tSNE` is to try to get the same distribution in the lower dimensional representation. Moreover, `tSNE` approach focuses on retaining the local structure of the data in the map. It is therefore helpful to separate the two groups to detect outliers. Unlike `tSNE`, `PCA` tries to map all the clusters as a whole due to which local structures might get lost. These can be seen on Figures 4 and 5. In this case, `tSNE` seems to be a better tool to reduce our data for the purpose of detecting outliers.

2. Due to the dimension reduction, we use the χ^2 cutoff with 2 degrees of freedom. After the dimension reduction we observe that the percentage of true positives increases clearly, which is a good thing since it shows that the initial outliers are better detected, the results are excellent with robust distances.
The percentage of good observations tagged as outliers decreases, which is nice too. For the classical distances we even tend to 0%. `tSNE` thus make a great job.

Mean	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0.2$
Classical PTP	-	26.26%	1.52%
Robust PTP	-	85.4%	82.4%
Classical PFP	0.29%	0.07%	0.07%
Robust PFP	2.77%	1.11%	1.6%

Table 2: Average PTP and PFP for 100 simulations, in reduced space

With the boxplots under, we see that the PTP and PFP are scattered for some of the measures, which shows that the confidence of the classifier isn't optimal.

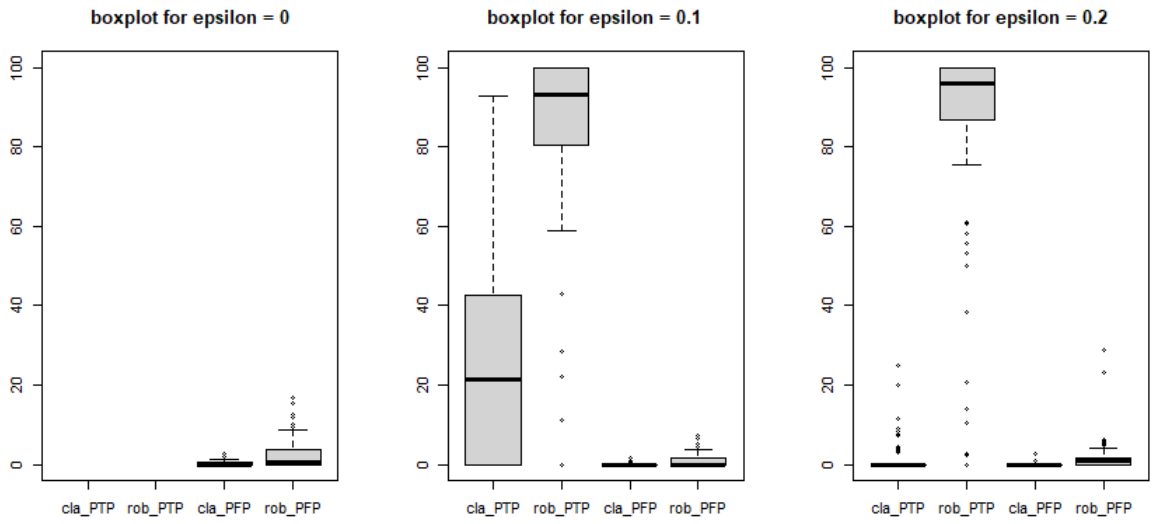

Figure 6: PTP and PFP boxplots with classical and robust detection, in reduced space

3 Supervised classification on the mixture

(a) The mixture being totally unbalanced, the logistic regression model wouldn't be a good choice to perform a supervised classification, for this method a more or less balanced mixture (ideally, the proportions of the different classes should be equal) is preferred.

(b) i. The value of the discriminant power with all the explanatory variables is 0.7929781 for $\epsilon = 0.2$, and 0.7541457 for $\epsilon = 0.1$. Taking $\epsilon = 0$ presents no interest since we work on a classifier, and with this value there is only one class.

ii. My development is done for $\epsilon = 0.2$, but everything holds for $\epsilon = 0.1$.
Globally, leaving any variable out doesn't make us lose a lot of power, excepted the three last variables that causes a bigger loss, which is logic since the outlier group is created in the distribution by playing on those 3 variables. After having tried to delete different number of variables, we observed that deleting 9 variables is the most appropriate number to have a discriminant power not so far from the initial one (we go from 0.7929781 to 0.7859185 which is acceptable) because deleting one more variable causes a more important loss of discriminant power. Indeed, we must make a compromise between losing discriminant power and simplifying the definition of the canonical variable. The 3 remaining variables are the 3 last for $\epsilon = 0.2$, and the 7th, the 10th and the 12th for $\epsilon = 0.1$. The fact that the three last are important is not a surprise as explained at the beginning of this paragraph, the 7th overtaking the 11th for $\epsilon = 0.1$ is more surprising, it may be because since we have only 10% of initial outliers, the trend in the variables shows up less.

iii. After having projected the data, we can classify an observation in some group

if its value is bigger or smaller than a certain threshold, which is obtained by training the model on a certain proportion of the data, on the variables that we kept.

(c) Table 3 shows extremely good results for the classification of the observations. The results are quite similar but slightly better for $\epsilon = 0.2$ than $\epsilon = 0.1$ (2% gain on PTP but 0.2% loss on PFP).

Mean	$\epsilon = 0.1$	$\epsilon = 0.2$
PTP	94.49%	96.79%
PFP	0.45%	0.64%

Table 3: Average PTP and PFP for 500 simulations, using LDA

It can be seen from Figure 7 that unlike all other detection techniques, the variance is low which means the classifier is confident.

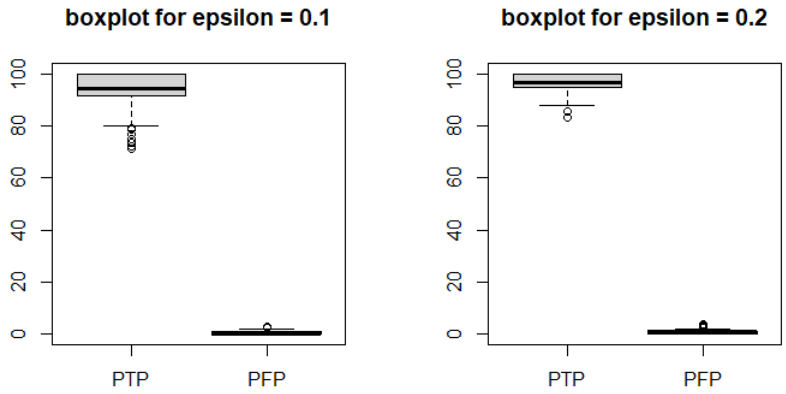

Figure 7: PTP and PFP boxplots using LDA

4 Conclusion

Information (contained in variance) from high-dimensional data is unorganized and dispersed. Dimension reduction techniques allowed us to reduce the size of our data by minimizing information loss. In addition, it allows to visualize the data in 2D, which was not possible with 12 variables. As a conclusion, we can say that with the different techniques of dimension reduction, ... we are able to classify accurately the two groups present in the mixture. Without the reduction it's much less accurate, because we have "parasite" variables, so variables that disturb the classification and the outlier detection, because their values don't allow to spot a particular trend that could help the processes. The best detection technique to perform classification and which provided the best results was the supervised classification using LDA.