University of Liège

High-dimensional Data Analysis

Exploratory data analysis and principal component analysis

Master 1 in Data Science & Engineering

Authors
Tom Crasset
Antoine Louis

Professors
G. Haesbroeck

Academic year 2018-2019

1 Presentation of the data

The chosen data set corresponds to clinical features that were observed for 64 patients with breast cancer and 52 healthy controls. It has been collected from the **UCI Machine Learning Repository** [1].

The data set contains 9 attributes, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer. The nine attributes are anthropometric data which can be gathered in routine blood analysis. These are the following :

Attributes	Unit
Age	years
BMI	kg/m^2
Glucose	mg/dL
Insulin	$\mu U/mL$
HOMA	/
Leptin	ng/mL
Adiponectin	$\mu g/mL$
Resistin	ng/mL
MCP.1	pg/dL

TABLE 1 – Attributes of the data set

where the different abbreviations are used for :

— BMI : Body mass index
— HOMA : Homeostasis Model Assessment
— MCP-1 : Chemokine Monocyte Chemoattractant Protein 1

Breast cancer screening is an important process for early detection and to ensure a greater probability of having a good outcome in treatment. By collecting data in routine consultation, blood analysis given by some robust predictive models could make an important contribution to the screening process.

Given this data set, some correlations could be found between the different attributes and a patient affected by breast cancer.

[1] https ://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

2 Exploratory analysis of the data

2.1 Statistical summary of the variables

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Age	24.0	45.0	56.0	57.3	71.0	89.0
BMI	18.37	22.97	27.66	27.58	31.24	38.58
Glucose	60.00	85.75	92.00	97.79	102.00	201.00
Insulin	2.432	4.359	5.925	10.012	11.189	58.460
HOMA	0.4674	0.9180	1.3809	2.6950	2.8578	25.0503
Leptin	4.311	12.314	20.271	26.615	37.378	90.280
Adiponectin	1.656	5.474	8.353	10.181	11.816	38.040
Resistin	3.210	6.882	10.828	14.726	17.755	82.100
MCP.1	45.84	269.98	471.32	534.65	700.09	1698.44

TABLE 2 – Statistical summary of the quantitative variables

Table 2 contains the statistical summary of the variables. Moreover, the data contains one binary variable, the *Classification* variable where they are :

— 52 instances of *class 1* ("Healthy controls")

— 64 instances of *class 2* ("Cancerous patients")

2.2 Correlation structure of the quantitative data

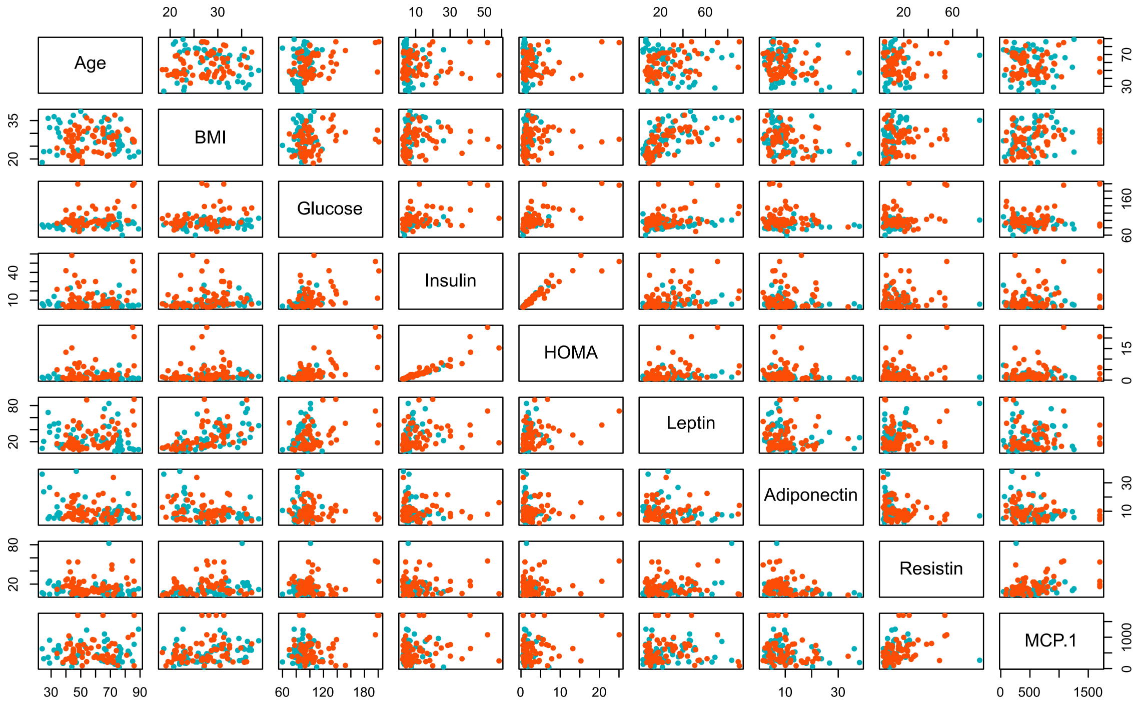

FIGURE 1 – Scatter plot matrix showing the distribution of the quantitative variables. Blue : Healthy, Red : Cancerous

In Figure 1, one can see the relation between the different quantitative variables of the dataset. As can be seen, there is an obvious correlation between *Insulin* and *HOMA*. Further correlations can not be seen directly from looking at this bivariate scatter plot so we have to resort to a more visual approach, for example a visualization of the correlation matrix, presented in Figure 2.

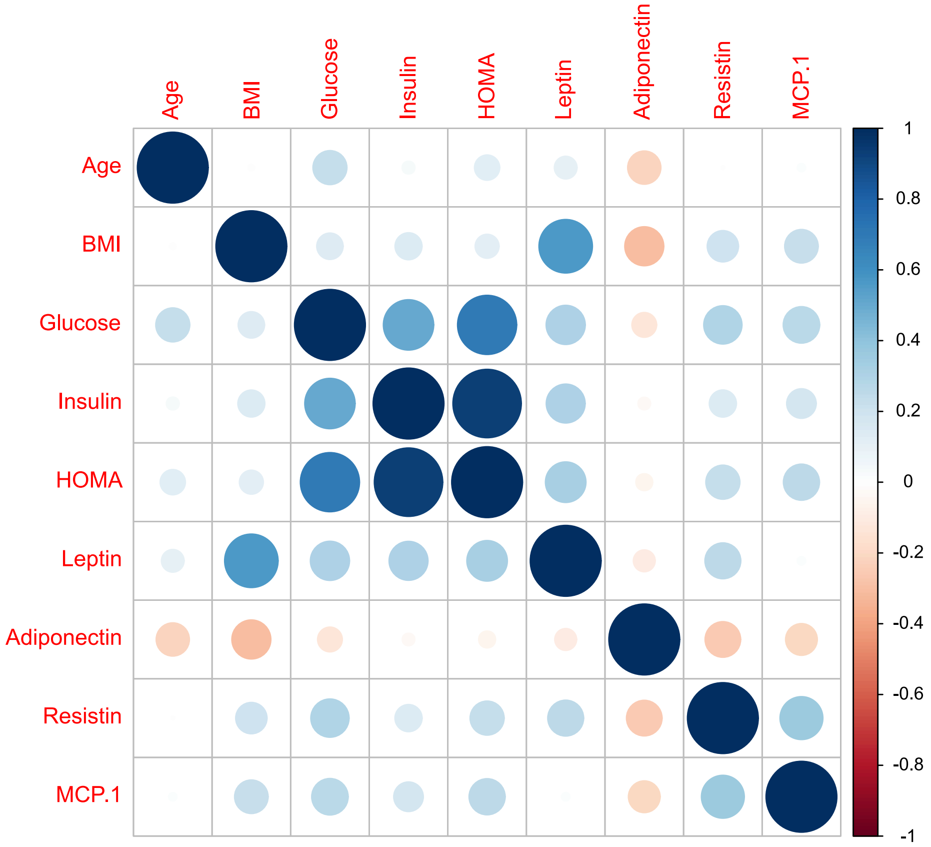

FIGURE 2 – Visualization of the correlation matrix of the quantitative data

In Figure 2, the correlations are much easier to see than in the scatter plot. The obvious positive correlation between *Insulin* and *HOMA* is confirmed in this figure and we gain knowledge about other noticeable correlations, such as between *Glucose* and *HOMA* or between *Leptin* and *BMI*.

2.3 Impact of the qualitative variable

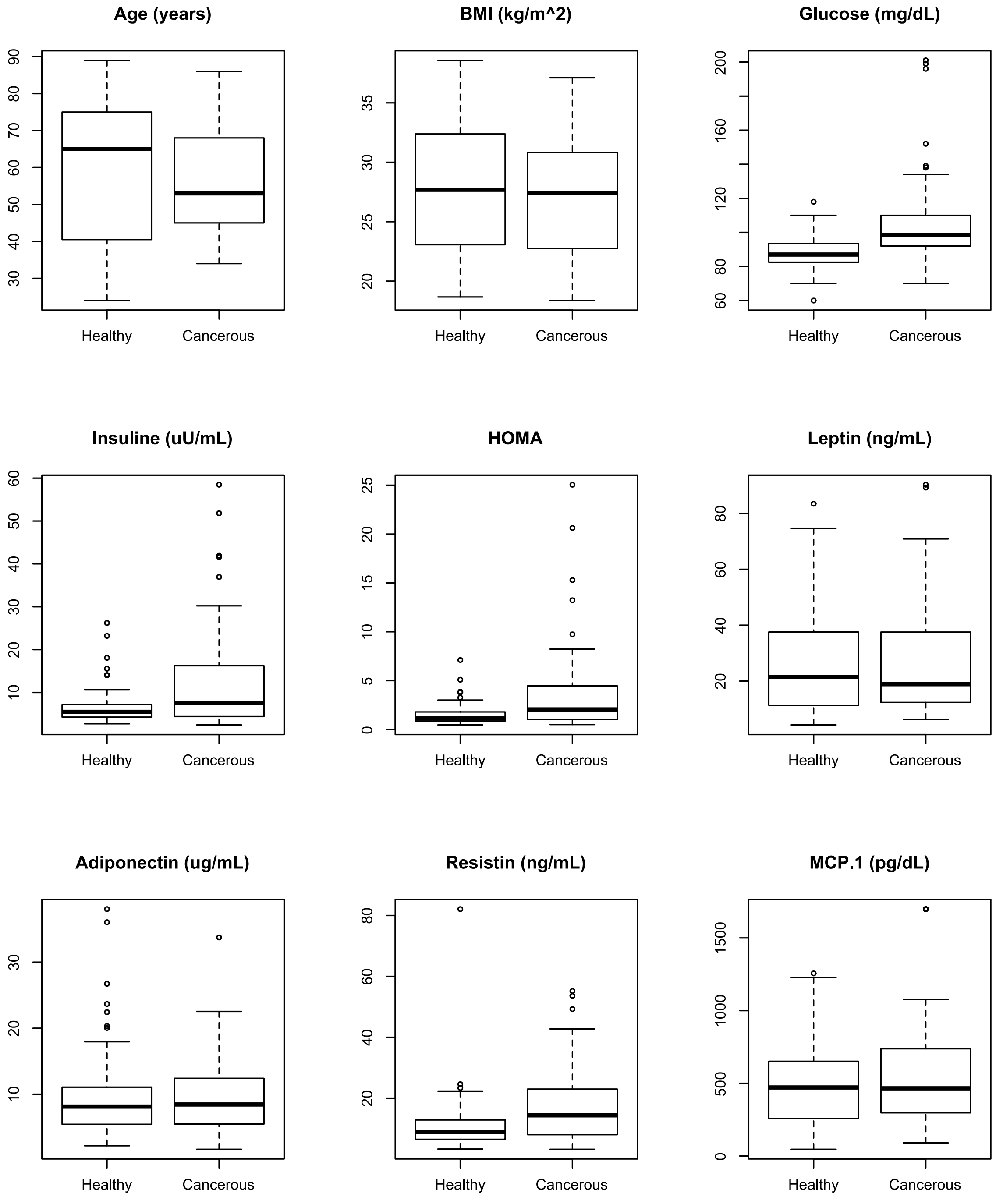

FIGURE 3 – Boxplots of the quantitative variables depending on the qualitative one

In Figure 3, boxplots of the quantitative variables are presented, depending on the qualitative *Classification* variable. From these boxplots, it seems that the age of healthy patients is higher by more or less 15 years than the one of cancerous patients. You might also notice that cancerous patients have a bigger concentration of *glucose*, *insulin* and *resistin*. Finally, one could say that the *BMI*, *leptin*, *adiponectin* and *MCP.1* rates don't seem to influence that much the fact that a patient is cancerous or not.

2.4 Robust multivariate approach

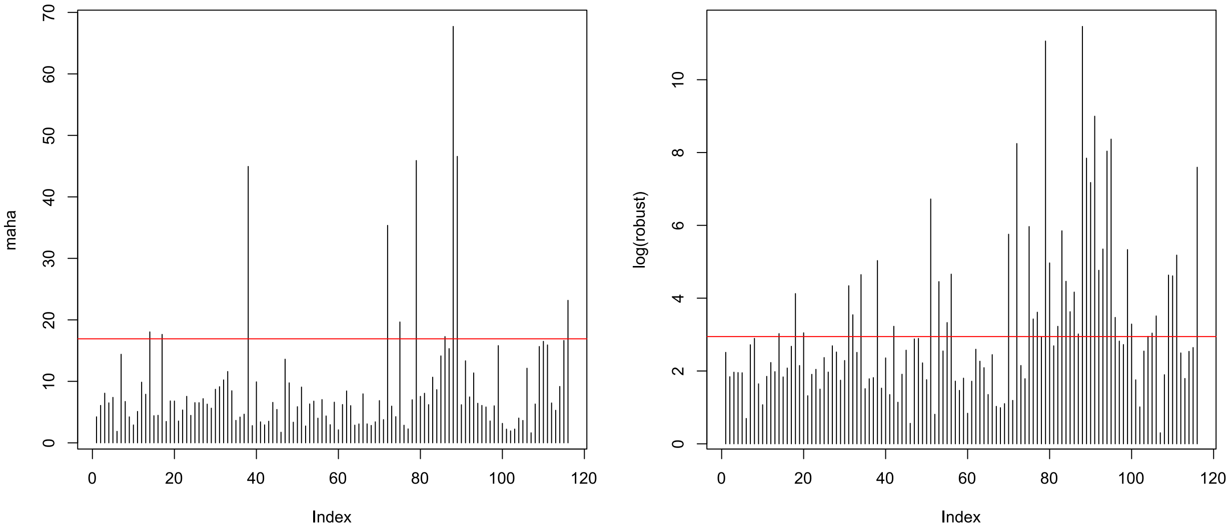

FIGURE 4 – Plots of the mahalanobis distances with a non robust estimator (left) and of the log of mahalanobis distances with a robust (MCD) estimator (right)

In Figure 4, one can see the plots of the mahalanobis distances with a standard estimator (left) and with a robust estimator (right). Notice that a log scale is used for the robust estimator. The red line was chosen using a chi-square distribution with a 0.975 confidence interval with 9 degrees of freedom.

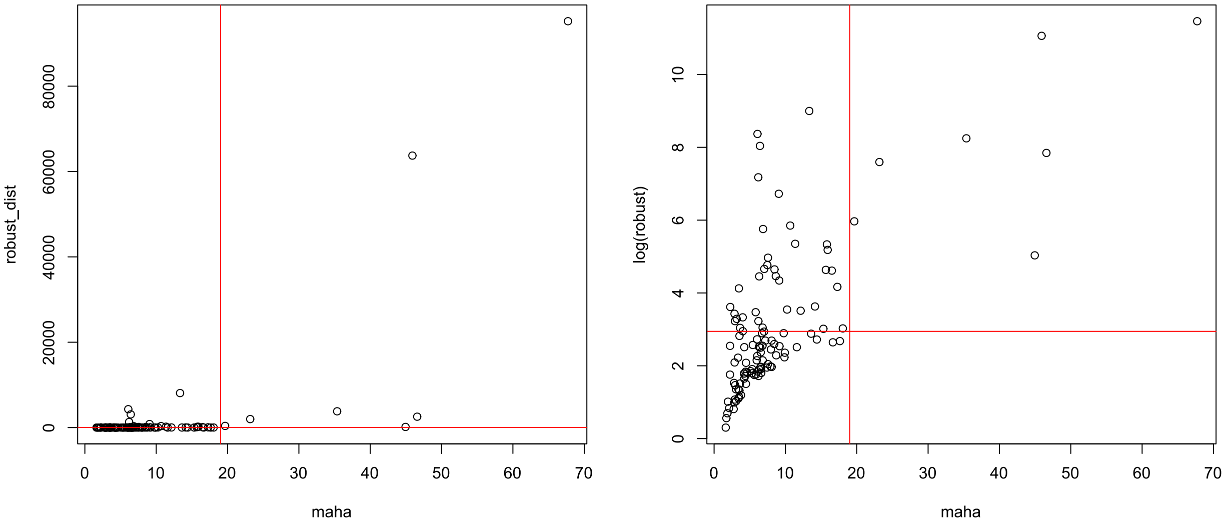

FIGURE 5 – Distance-distance plots of of the robust distances with the classical Mahalanobis distances

In Figure 5 are presented two DDplots in order to compare the robust distances with the classical Mahalanobis distances. Once again, the graph (left) is not easily interpretable because the scales are way too far off, so we chose to apply a log function to scale it better (right).

In both plots and especially the right one, 7 outliers clearly stand out. Also to be noted is the many outliers that were not labeled as such by the standard mahalanobis distance because they were masked by the extreme outliers. These are the points that are in the upper left quadrant of the graph.

3 Principal component analysis

The principal component analysis is an useful to reduce the number of features while maintaining a high amount of information.

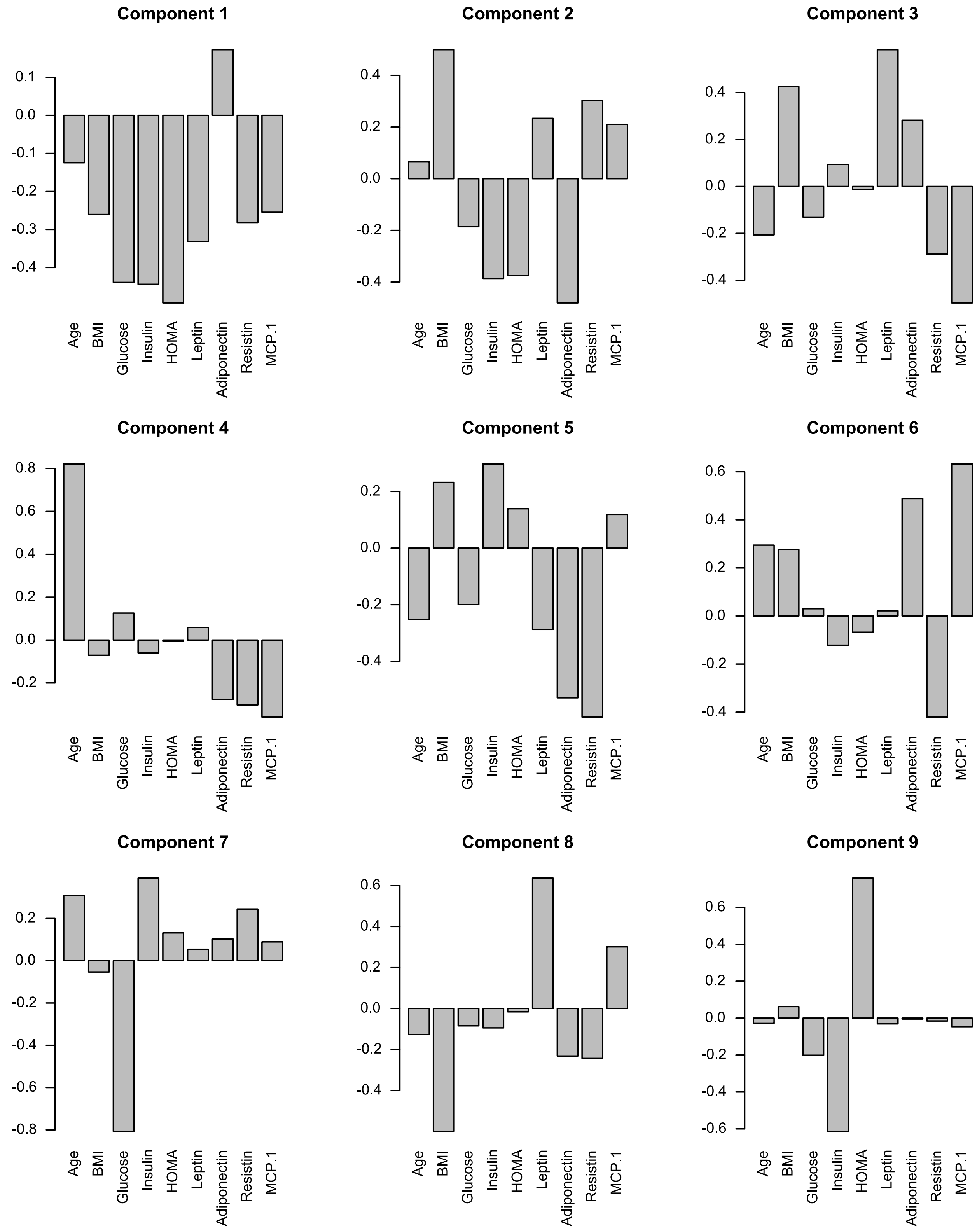

FIGURE 6 – Barplots representing the loadings

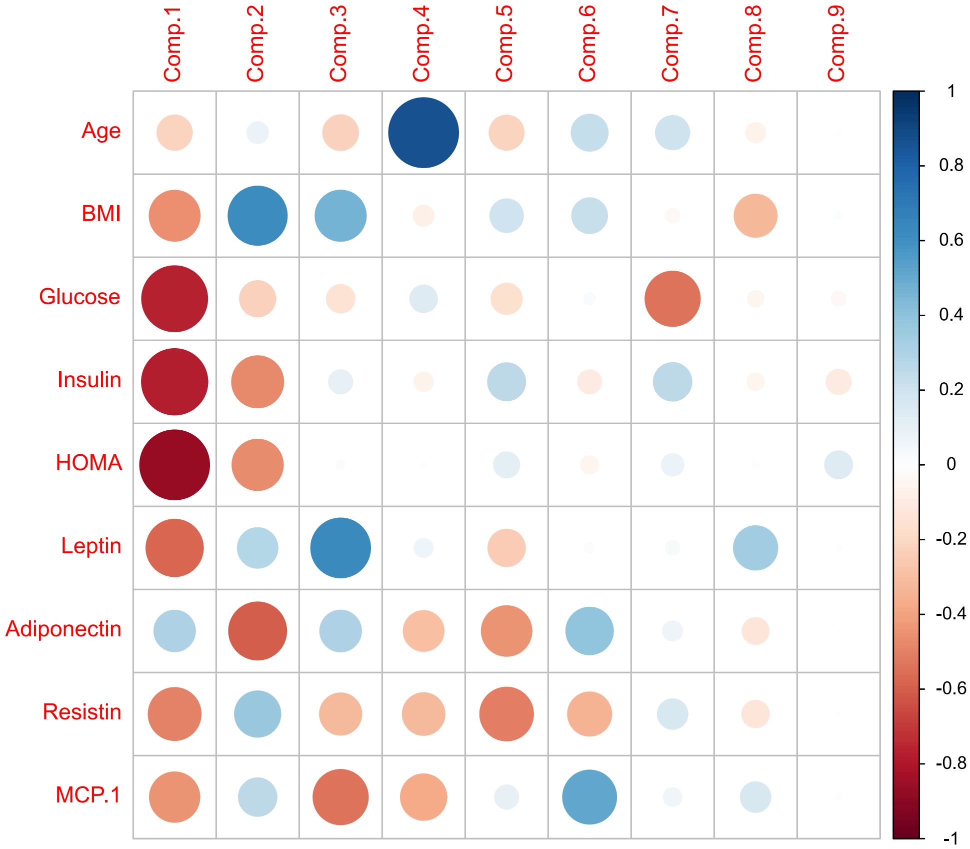

FIGURE 7 – Correlation matrix between the principal components and the data features

Figure 6 shows the loadings of the different features relative to each component while Figure 7 shows a visualization of the correlation matrix between the principal components and the data variables.

The analysis was done with a correlation matrix because, while almost all our features have the same units of measurement, they were often orders of magnitudes apart. The difference in both options is astounding and can be seen from the two scree plots in the Figure 8.

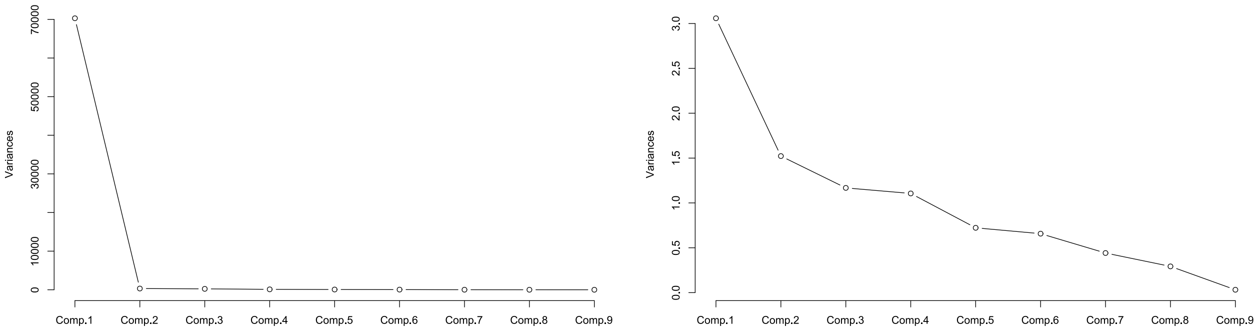

FIGURE 8 – Scree plot for the PCA using covariance (left) and correlation (right)

4 Further analysis

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Cumulative proportion of variance	0.34	0.51	0.64	0.76

Table 3 – Cumulative proportions of the variance of the components

Table 3 contains the cumulative proportions of the variance from the data set with each principal component added. We chose to cut off at 75% of the total variance, which is maybe still a bit much and we could have stopped earlier and taken only 3 or 2 components.

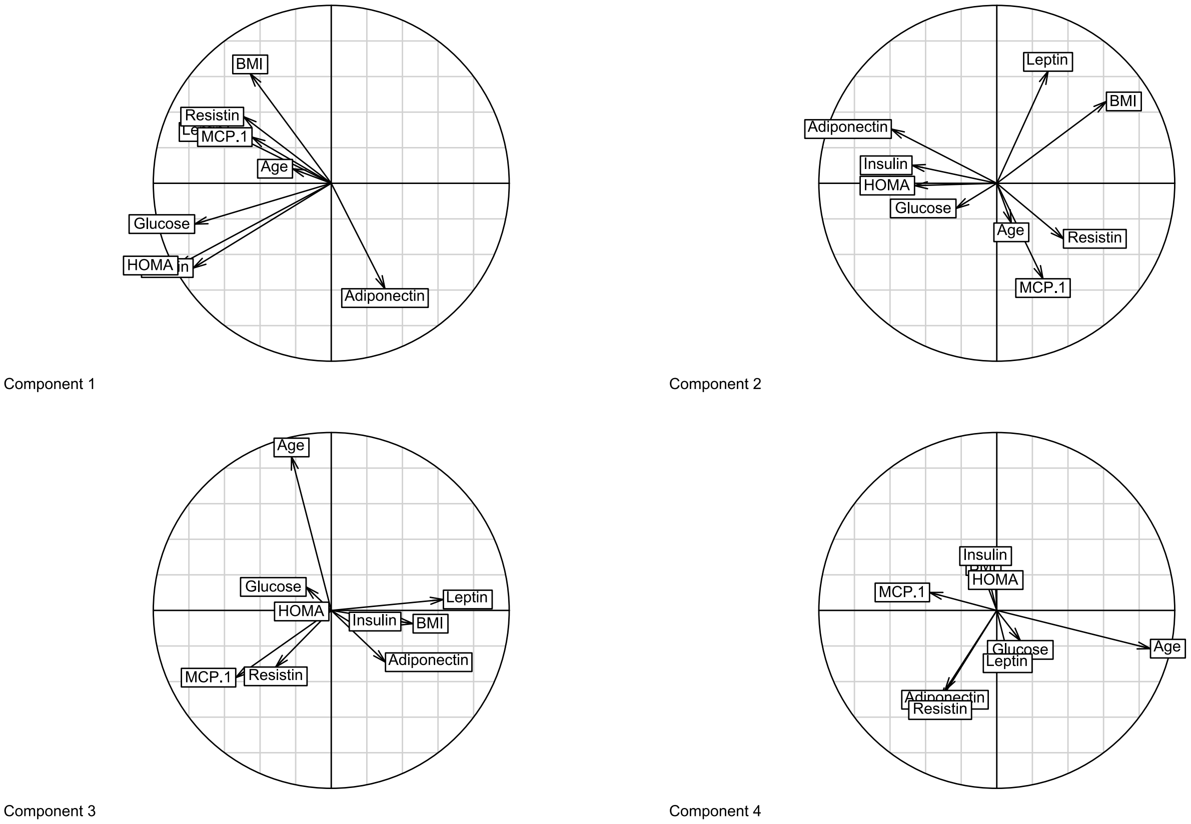

Figure 9 – Correlation circles of the 4 first principal components

To put an image in front of numbers, Figure 9 represents the correlation (magnitude and sign) of the multiple data features of each principal component up to the fourth one using a simple but powerful tool, the correlation circle. This goes hand in hand and shows the same data as Figure 6 and 7 but in a different way.

5 Conclusion

High dimensional data analysis is now easier than ever before. Using a few quick tools, we were able to gain a lot of knowledge on this data set. Especially, we found that concentrations of a few selected molecules in the bloodstream could be an indicator for breast cancer. This would be a cost effective, quick and non invasive procedure for the patients to get a first look at whether they might be at risk.

We found that cancerous patients have a tendency to have higher concentrations of *glucose*, *insulin* and *resistin*, which is close to the result that came out of this paper : "*Support vector machines models using Glucose, Resistin, Age and BMI as predictors allowed predicting the presence of breast cancer in women [...]*"[2].

[2] Using Resistin, glucose, age and BMI to predict the presence of breast cancer, Miguel Patrício et al., 2018