University of Liège

High-dimensional Data Analysis

Supervised classification using logistic regression and multivariate ranks

---

Master 1 in Data Science & Engineering

Authors
Tom Crasset
Antoine Louis

Professors
G. Haesbroeck

Academic year 2018-2019

1 Data

The data used for the prediction model is the same as the one used for the exploratory analysis. Quoting ourselves from the previous project :

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.

2 Supervised classification with logistic regression

2.1 Adequacy of the classification

Before doing any prediction, the conclusion from the exploratory analysis will be taken into account to determine, a priori, which features are adequate at describing the cancer diagnosis. Let's take back the statistical summary table of the quantitative 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 1 – Statistical summary of the quantitative variables

Table 1 doesn't give any relevant information about the features that could be used for the classification. By using the correlation matrix presented in Figure 1, we gain information on which variables are correlated. 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*. When doing a classification, one doesn't need to include variables that are heavily correlated, as they bring the same information to the table with respect to the classification. However, this doesn't specify which features to use, it rather informs us on what features not to use together.

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

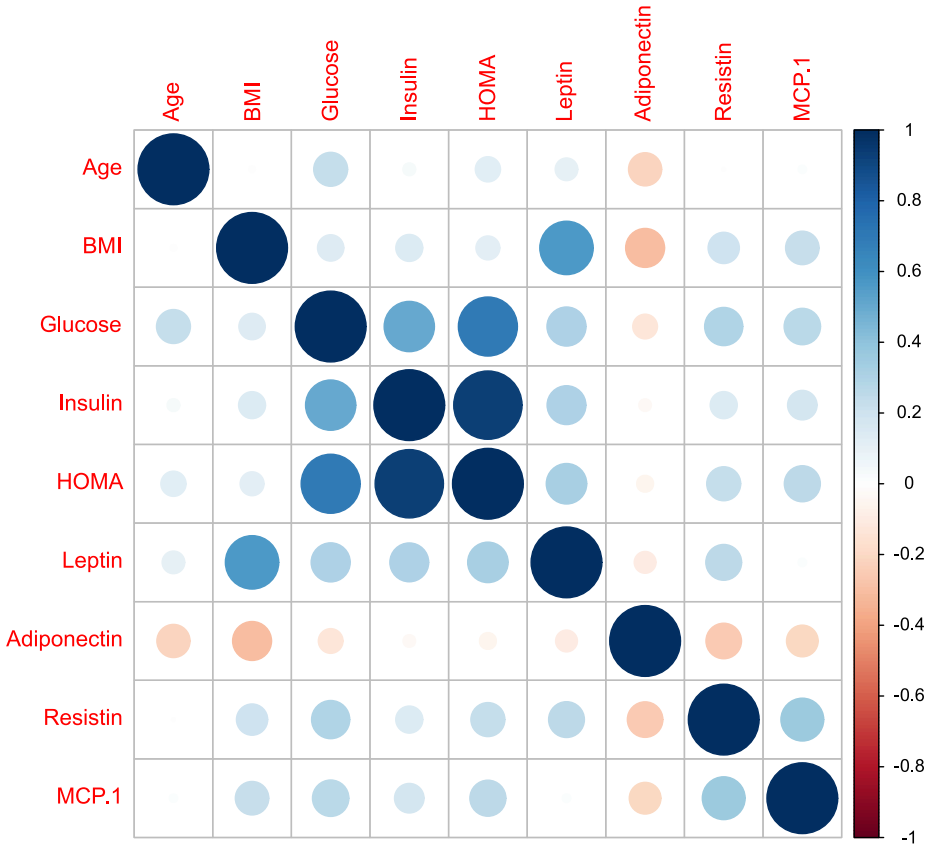

FIGURE 1 – Visualization of the correlation matrix of the quantitative data

Then, let's take back the boxplots we draw in the exploratory analysis project. Quoting ourselves :

In Figure 2, boxplots of the exploratory 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.[2] *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. Editor's note : However, this may be the cumulative result of older individuals being more inclined to join the study, as they fear being at risk due to their age, and of younger individuals who feel symptoms of breast cancer being more likely to participate in the study than young healthy individuals.

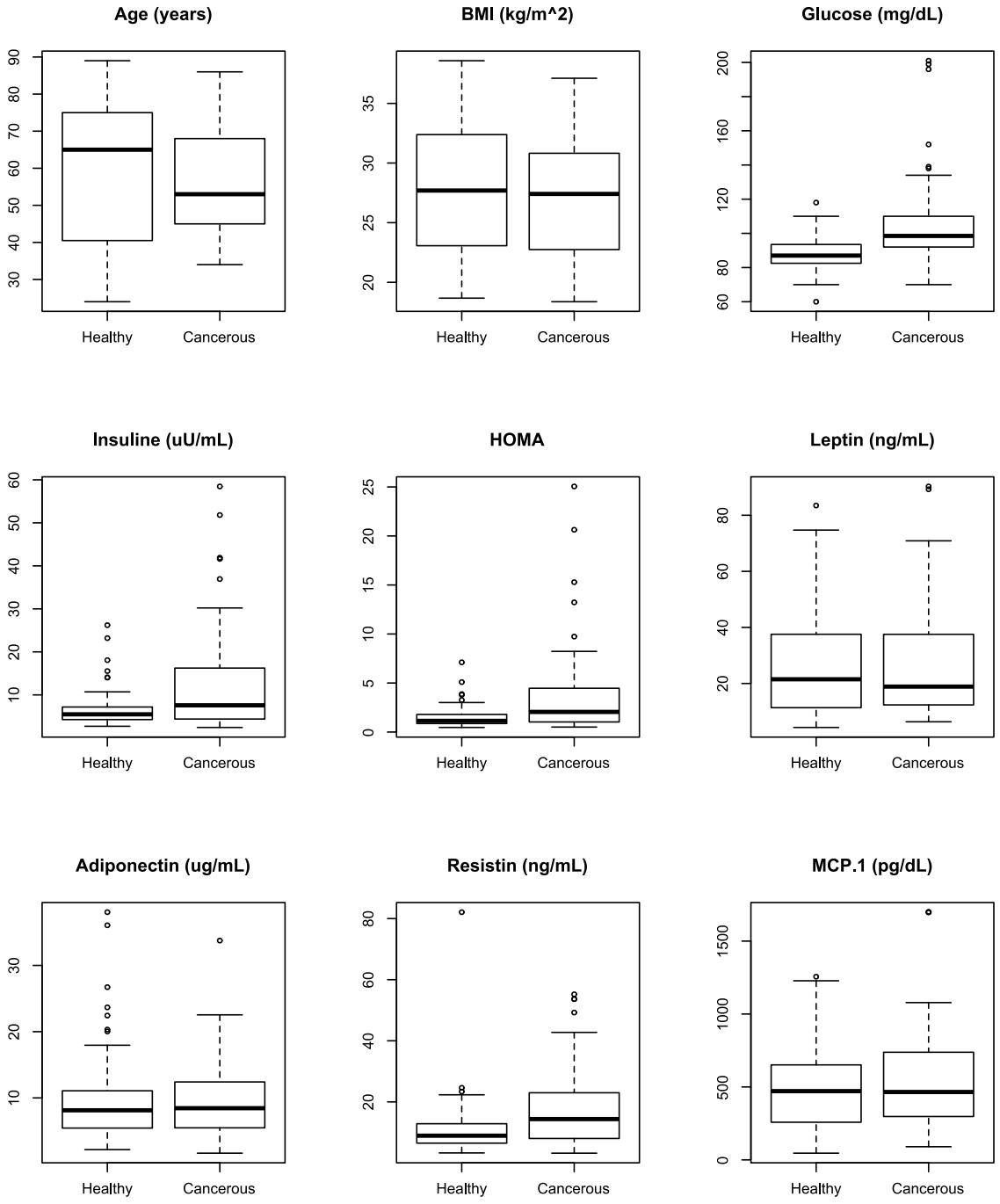

FIGURE 2 – Boxplots of the quantitative variables depending on the qualitative one

2.2 Logistic model

To select the variables needed to correctly classify the diagnosis of the patient, we used the `stepAIC` method in R to compute the (generalized) Akaike Information Criterion, with stepwise change of the variables used, both backwards and forward. After a rapid computation, the variables *BMI*, *Glucose*, *Insulin* and *Resistin* where deemed to minimize the AIC. However, while using this method we had the following error message : *1 :*

glm.fit : fitted probabilities numerically 0 or 1 occurred. This ocurs when the data is linearly separable. Thus, we decided to use a regularized regression using LASSO to make it

2.3 Cross-validation

Cross validation was done with regularized regression using LASSO with a leave-one-out strategy to find the best coefficients as well as changing the regularization parameter λ. The misclassification error can be found in Table 2. Looking at this error tells us that determining if a patient has breast cancer using our model is slightly better than tossing a coin, so nearly random. We weren't able to decrease that error in any way. That shows that evaluating the presence of breast cancer using only bloodtests is only a preliminary measure and should not be the determining factor for diagnosis.

MSE	0.448 ± 0.0129

TABLE 2 – Lowest misclassification error with $\lambda = 0.0225$

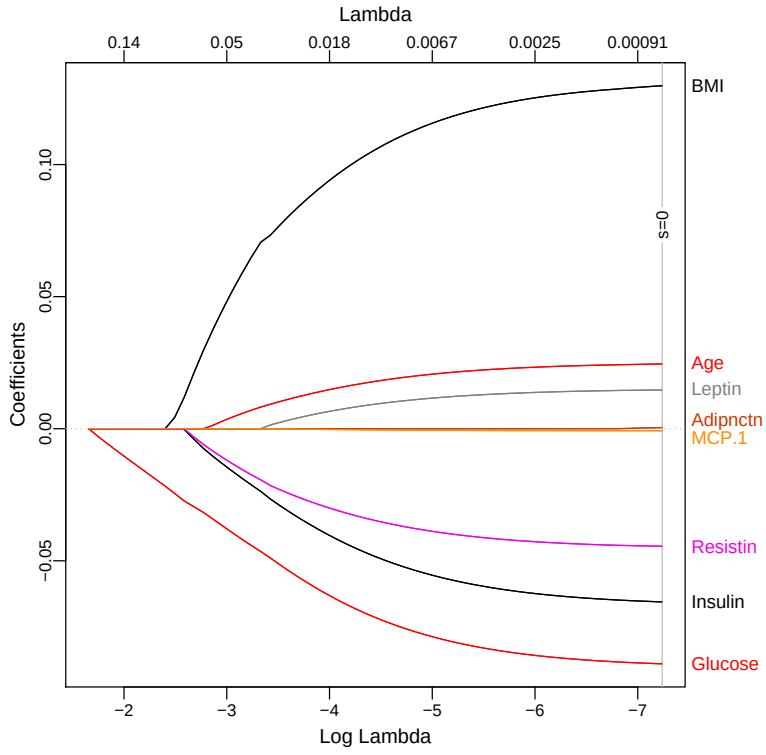

FIGURE 3 – Evolution of the coefficients of exploratory variables with varying lambda

	Weight
Intercept	2.856
Age	0.013
BMI	**0.0875**
Glucose	**-0.0586**
Insulin	**-0.0360**
HOMA	0
Leptin	0.005
Adiponectin	0
Resistin	**-0.0274**
MCP.1	-0.0002

TABLE 3 – Coefficients and their weight for the best logistic regression model after cross validation ($\lambda = 0.0225$) Highlighted in bold are the most significant coefficients

Figure 3 and Table 3 show the evolution of the coefficients with the varying regularization parameter and the coefficients of the best model respectively. The bold coefficients are the most relevant ones, and the other features (apart from Age) are really not significant, with a weight of magnitude 10^{-4}. Thus, we find the same significant features as the one that came out using the Akaike information criterion and during the exploratory analysis phase.

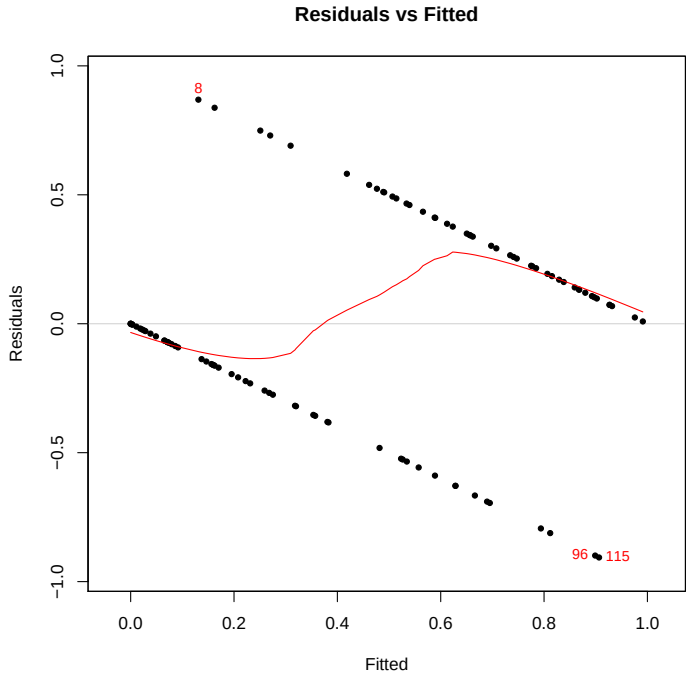

FIGURE 4 – Residuals with respect to the fitted values of the logistic regression model

The residuals plotted against the fitted points is a good way to see if the model has any suspicious behavior. As can be seen on Figure 4, the residuals are distributed randomly over the whole range. This is a good sign that the model is somewhat correct and not completely flawed. However, this is not enough to determine the goodness of the model.

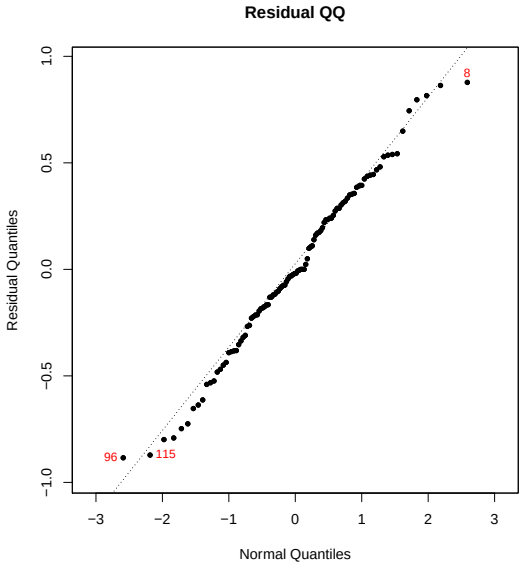

FIGURE 5 – Residual quantiles in relation to the normal quantiles

The QQPlot on Figure 5 is to check for normality of the residuals and as we can see, they are distributed pretty well in the middle range, but they are not very normal anymore at the lower end.

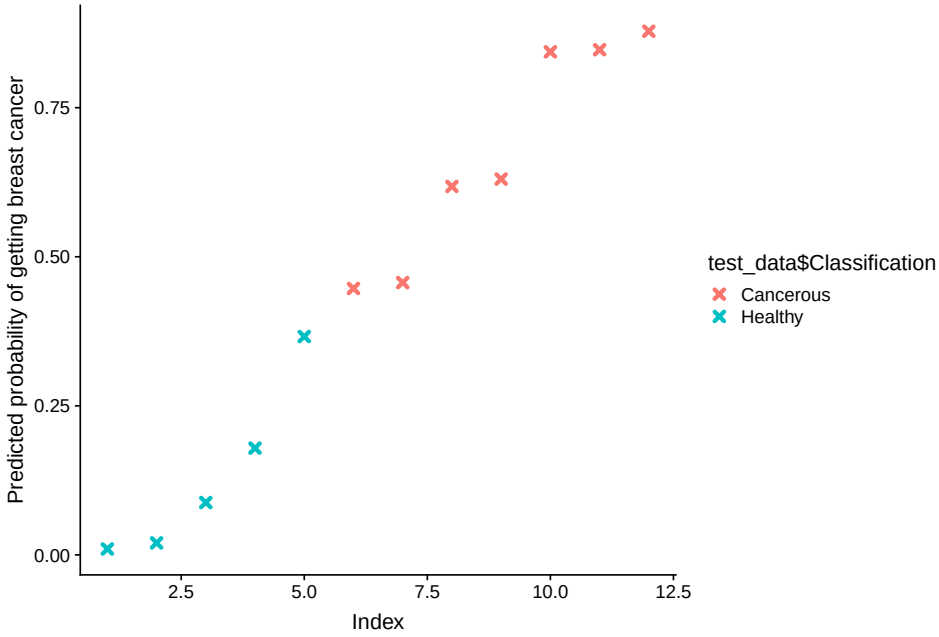

FIGURE 6 – Prediction of the model

To conclude, a prediction was done on 10% of the data to see the goodness of the model. As one was see, just under the 0.5 probability value, two cancerous patients were deamed healthy, the rest were correctly classified. However, that is because the threshold for the classification is a strict boundary at 0.5 probability.

3 Multivariate ranks

3.1 Multivariate ranks with Mahalanobis depths

Using all the quantitative variables of our data, it is asked to compute the ranks of the observations using the Mahalanobis depth function, that is based on an outlyingness measure : the Mahalanobis distance between the given point and the center of the data. To compute it, two choices are possible : either classically estimated by means of the sample mean vector and covariance matrix or robustly estimated with the MCD estimator. The former is easy to compute but the corresponding depth may be sensitive to outliers. Choosing the latter ensures obtaining a depth that is more robust to outliers in the data.

The data is represented on the first principal plane in Figure 7 in some colors varying from green to red in respects with their depth value, green corresponding to a small depth and red to a big depth.

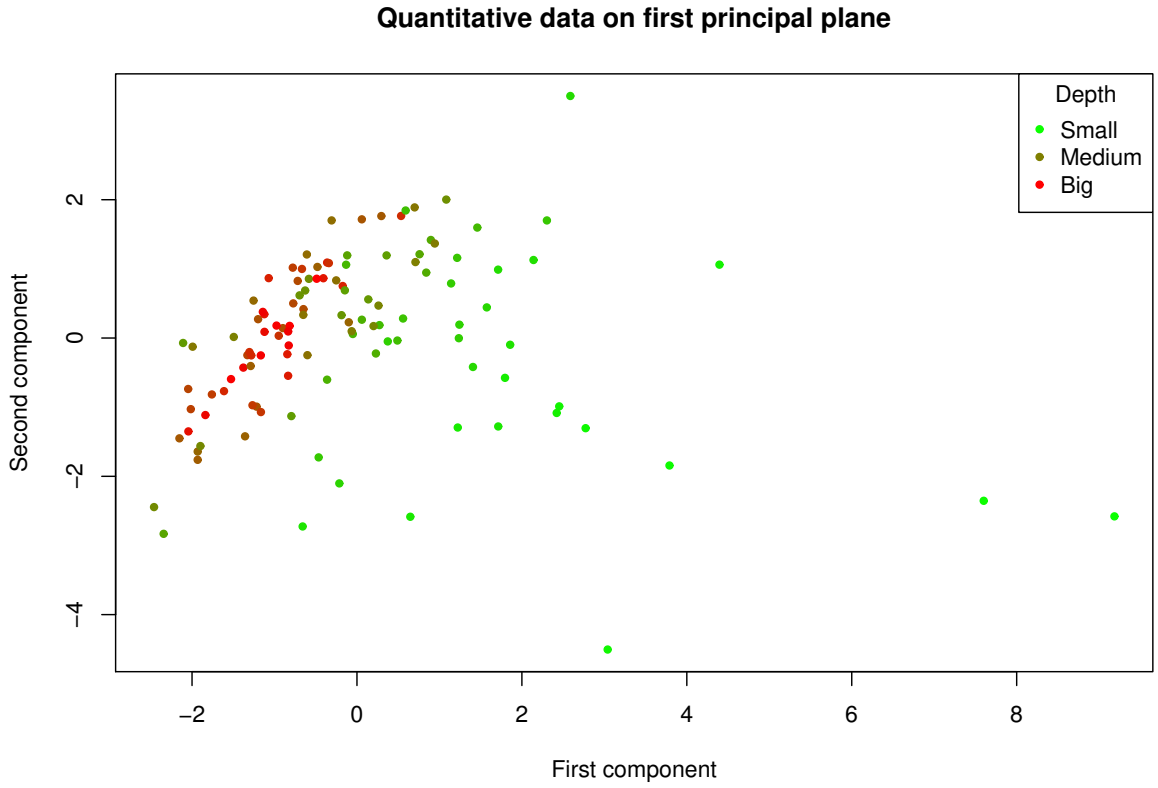

FIGURE 7 – Data on first principal plane with respect to their ranks

Let's now have a look at the extreme cases of our observations and check their specificities. The summary table of the depth variable for all the observations is represented in Table 4.

	Depth values
Min	0.0000104
5^{th} Quantile	0.0003745273
1st Quartile	0.0275956
Median	0.0731731
Mean	0.0949595
3rd Quartile	0.1416265
95^{th} Quantile	0.2577666014
Max	0.4090883

TABLE 4 – Summary table of the depth variable on all observations

From Table 4, one can see that that the depth values of the observations points below the 5^{th} quantile are extremely small compared to those of the first quartile for example, and so these points can be considered as our extreme cases. The summary table of these observations is given at Table 5.

Comparing these values with the one from the summary table over all observations, presented in Table 1, one can see that the values of the variables for the extreme cases tend to be very large. For example, the *insulin* rate is extremely high with a median of 41.75 whereas the median is 5.925 for all observations, and the *HOMA* variable has a median of 14.256 for the extreme cases whereas it is 1.3809 for all observations. The extreme cases seem to represent the outliers of our observations.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Age	40	45.25	54.50	60.67	78.75	86
BMI	24.74	26.92	29.26	28.94	31.13	32.46
Glucose	106	128.8	132.5	149.3	180.5	201
Insulin	24.89	33.00	41.75	41.47	49.33	58.46
HOMA	8.226	10.609	14.256	15.539	19.294	25.050
Leptin	18.16	32.74	40.12	41.33	46.33	70.88
Adiponectin	5.357	6.596	8.153	9.120	10.197	16.100
Resistin	5.310	7.201	14.528	19.953	22.666	55.215
MCP.1	244.8	456.6	647.3	785.4	972.9	1698.4
Depth	1.044e-5	4.244e-5	1.767e-4	1.601e-4	2.524e-4	3.216e-4

TABLE 5 – Summary table of all quantitative variables for 5^{th} quantile observations

A similar analysis could be made for the deepest cases. These can b considered as the observations having a depth value greater than the one of the 95^{th} quantile. The summary table of these observations is given at Table 6.

Here, the values of the deepest cases are much smaller and the median values of each variable seem pretty close to those for all observations, meaning that the variables values stay pretty close to each other in the second quartile (before the median). One can also notice that the intervals between the min value and the max value of each variable is quite tight compared to the original one of all the observations. For example, the *glucose* rate for the deepest cases is contained between 85 and 100, which is quite precise given the original interval of $[60, 200]$ over all the observations.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Age	45	52.25	63	60.5	65.5	77
BMI	22.2	22.71	24.38	25.12	27.25	29.38
Glucose	85.00	90.00	91.00	91.17	92.00	98.00
Insulin	3.226	3.756	4.646	4.624	5.453	6.042
HOMA	0.7321	0.8327	1.0033	1.0413	1.2676	1.3779
Leptin	6.832	10.446	12.965	15.220	21.321	24.846
Adiponectin	4.784	6.896	8.703	8.958	10.865	13.680
Resistin	6.705	10.412	11.235	11.493	13.378	15.556
MCP.1	225.9	419.2	509.6	494.4	598.6	704.0
Depth	0.2586	0.2853	0.3104	0.3210	0.3483	0.4091

TABLE 6 – Summary table of all quantitative variables for 95^{th} quantile observations

Notice that these observations can not lead to any conclusions because the set of considered points is very small (5% of a dataset of approximately 100 points).

3.2 Univariate ranks with center-outward procedure

To see which variables are the most interesting in our dataset, a comparison between the values of the different variables for the deepest cases and the extreme ones could be made, looking for the biggest differences between these two. Comparing Table 5 and Table 6, one can notice that almost all variables have (sometimes significantly) smaller values for the deepest cases than for the extreme cases, with an outstanding difference for *Insulin*, *Leptin* and *HOMA*. However, considering *HOMA* as one of the two most representative variables would be a mistake as this variable contains a lot of outliers. So the *Insulin* and *Leptin* variables were the one we decided to take as the two most representative variables.

To compute the univariate ranks of each one of these two variables, a simple median-centering procedure can be used, substracting the median value of each variable to all the values and considering that as their 1D-ranks.

3.3 Bivariate ranks using bag plot

Using the two most interesting quantitative variables *Insulin* and *Leptin*, the data can be represented in 2D with a bag plot, as shown in Figure 8.

By looking at Figure 8, one can notice that the median has quite a small value, as for the majority of the points located in the bag (dark blue). The min points are then really close to the bag while the max points are more distant. Finally, one can see that there aren't many outliers (only 4).

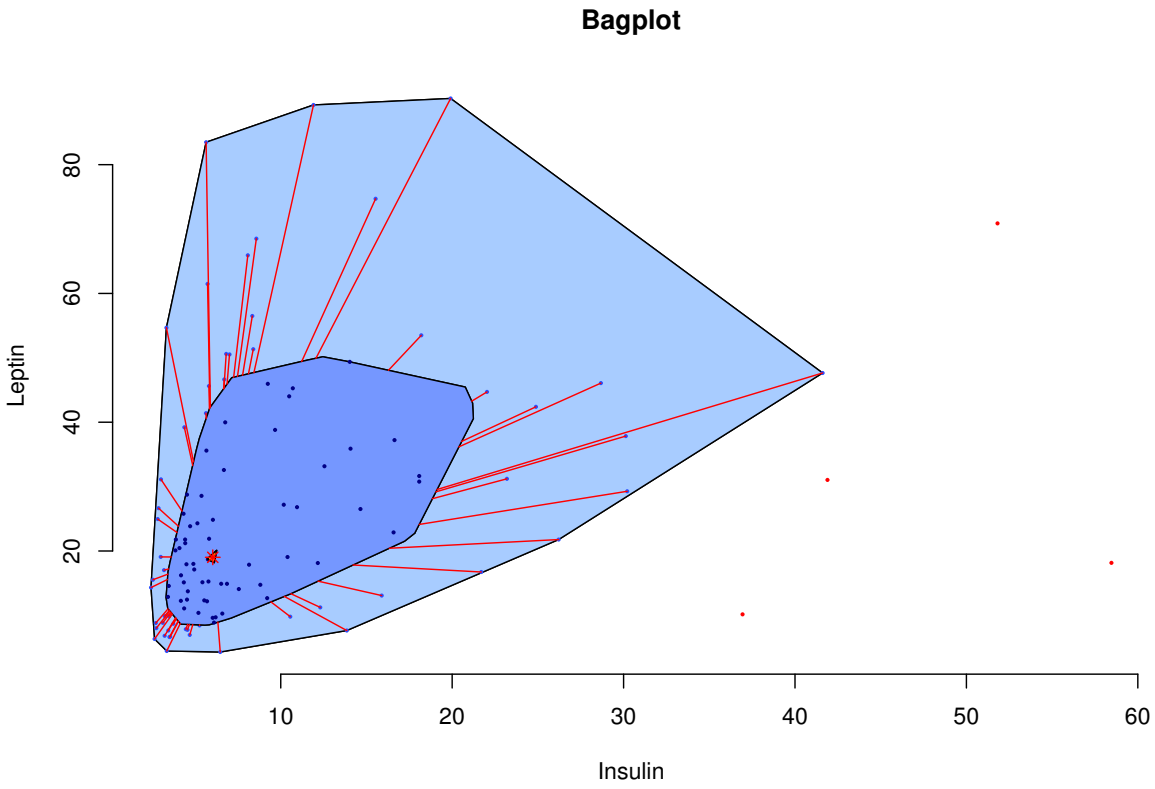

FIGURE 8 – Bag plot of the two most representative variables (Insulin and Leptin)

An interesting point would be to measure the discrepancy between this 2D-ranking and the multivariate ranking derived in Section 3.1, as well as with the 1D-rankings computed in Section 3.2. To do so, the Spearman's rank-order correlation could be used. Indeed, Spearman's correlation coefficient measures the strength and direction of association between two ranked variables. The results of the different correlations are considered in Table 7.

One can see that there is a moderate downhill linear relationship between the 2D-rankings and the two 1D-rankings of our representative variable. On the other hand, the correlation is weaker between the 2D-rankings and the multivariate Mahalanobis rankings.

A more graphical way to see some kind of correlation between these rank variables is by plotting the adequate scatter plots, presented in Figure 9.

		Correlation coefficient
2D-rankings	Mahalanobis-rankings	0.3871317
2D-rankings	1D-Insulin-rankings	-0.4560389
2D-rankings	1D-Leptin-rankings	-0.5220641

TABLE 7 – Correlation coefficients between the computed rankings

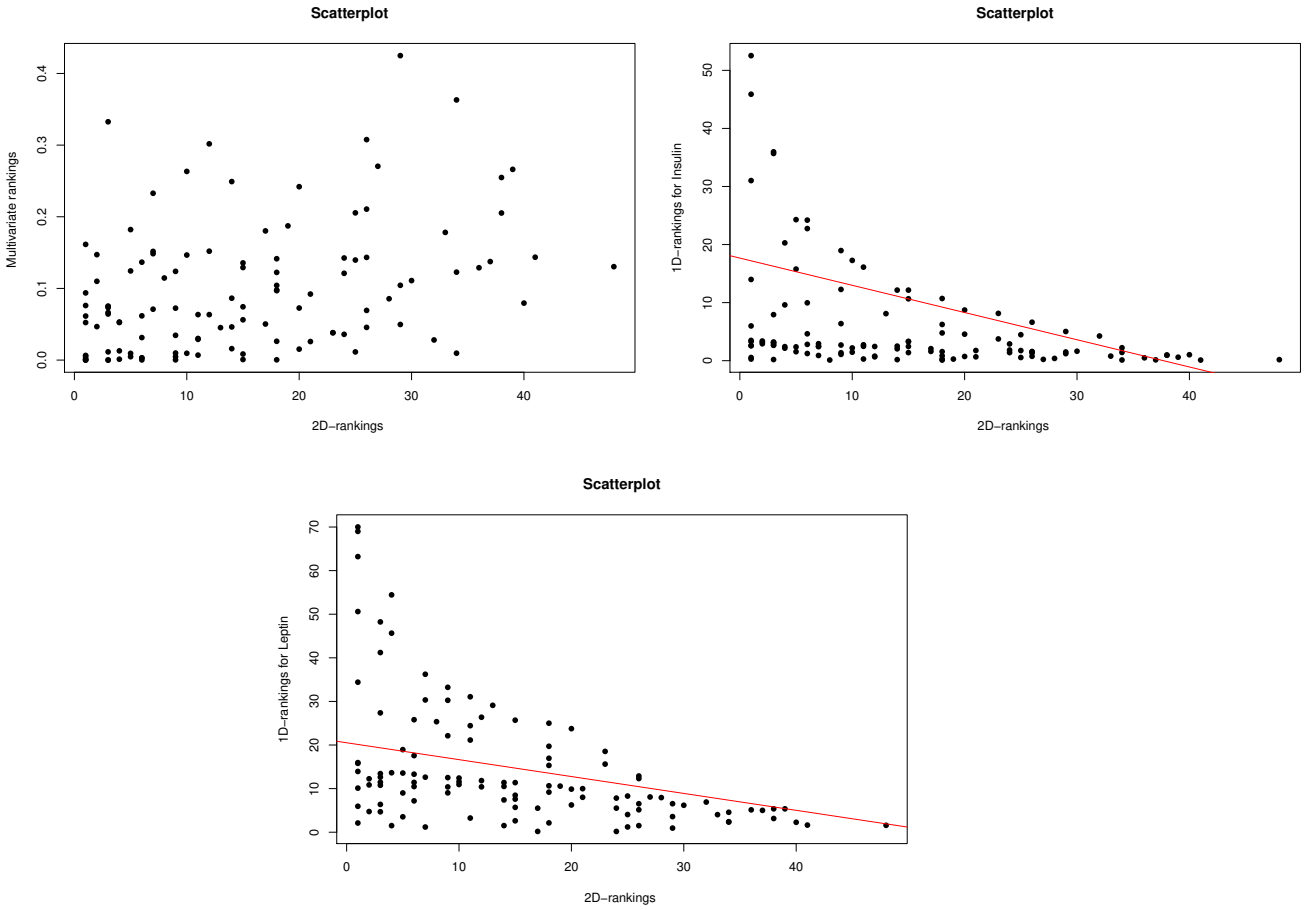

FIGURE 9 – Scatter plots