Faculty of Applied Sciences

High-dimensional data analysis

MATH2021-1

Project 1: Exploratory data analysis and correlation analysis

Authors :
Delcour Florian - s181063
Makedonsky Aliocha - s171197

Instructors :
Haesbroeck Gentiane

October 2021

1 Statistical analysis

1. https://archive.ics.uci.edu/ml/datasets/Wine

   The dataset we chose contains information about chemical analysis of three different kinds of wine, which were all grown in Italy. The data was collected on 178 wines by an institute of pharmaceutical, food analysis and technology in July 1991. The original dataset contained around 30 variables but for some reasons, the author gives us a dataset with only 14 variables, and we removed the variable "category" which is the category in which the wine is ranked, but for the purpose of this project it wasn't really interesting, since it doesn't represent a numeric value. The 13 others variables are constituents found in each of the three types of wine. For all the variables, here is a small description :

   - Alcohol : the percentage of alcohol in the wine
   - Malic acid : one of the main acids found in the acidity of grapes (in g/L)
   - Ash : inorganic matter that remains after evaporation and incineration (in g/L)
   - Ash Alkalinity : the alkalinity of ash determines how basic (as opposed to acidic) the ash in a wine is.
   - Magnesium : magnesium is a metal that affects the flavor of wine. (in mg/L)
   - Total phenols : include a large group of several hundred chemical compounds that affect the taste, color and mouthfeel of wine (g/L)
   - Flavanoids (type of phenol) : natural product well known for their beneficial effects on health thanks to its anti-oxidative, anti-inflammatory, anti-mutagenic and anti-carcinogenic properties
   - Non-flavanoids phenols (type of phenol) : participate in copigmentation, oxidation and juice browning of wine (g/L)
   - Proanthocyanins (type of phenol) : capability to bind salivary proteins, these condensed tannins strongly influence the perceived astringency of the wine (g/L)
   - Color intensity: refers to the degree of color shade
   - Hue: refers to the vividness of the color and the degree of warmth and coldness.
   - OD280/OD315 of diluted wines: protein content measurements.
   - Proline: an amino acid present in wines

   The missingness rate is null because there is no missing value, so we don't have to handle this.

   Based on the description above, we expect that total phenols is strongly correlated to flavanoids, non-flavanoids phenols and proanthocyanins which are all types of phenol. Malic acid and ash alkalinity may be correlated too since the second determines the basicity or acidity of the wine. The percentage of alcohol may influence the color intensity or the hue of the win. It's difficult to link other variables because it's essentially about chemistry.

2. Here are the histograms of all variables of our dataset, few are normals except Alcanity of ash but thanks to the log-transform we are able to improve it a bit.

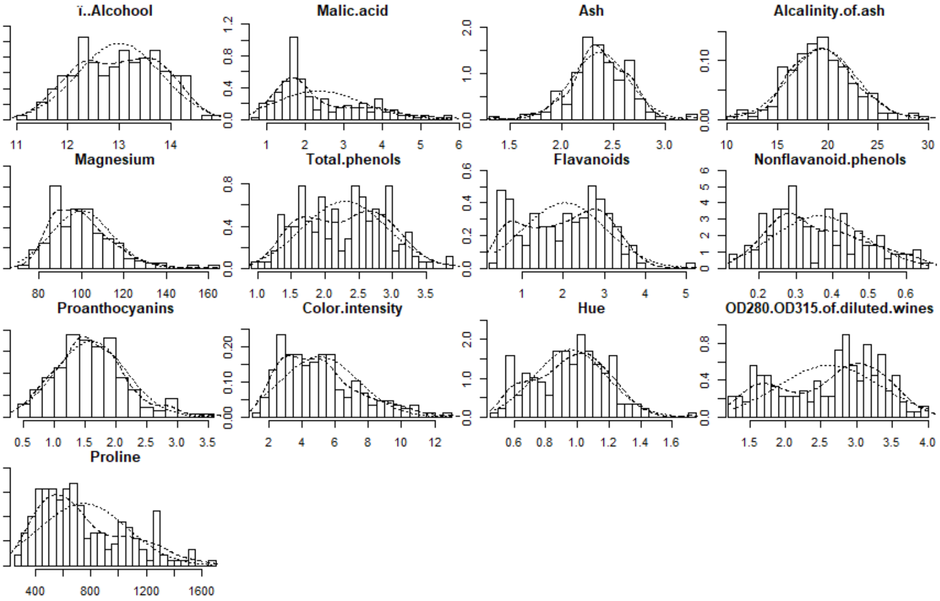

Figure 1: Histograms of all the variables

Here are the histograms after the application of the log-transform to all the variables whose distributions weren't quite normal (i.e. all variables except "Alcanity of ash"). After the log-transform, "Proanthocyanins" and "Color intensity" also follow a normal distribution, and the majority of the others tend to this normality too, but some are far from normality (e.g. Flavanoids, Malic acid, and Proline).

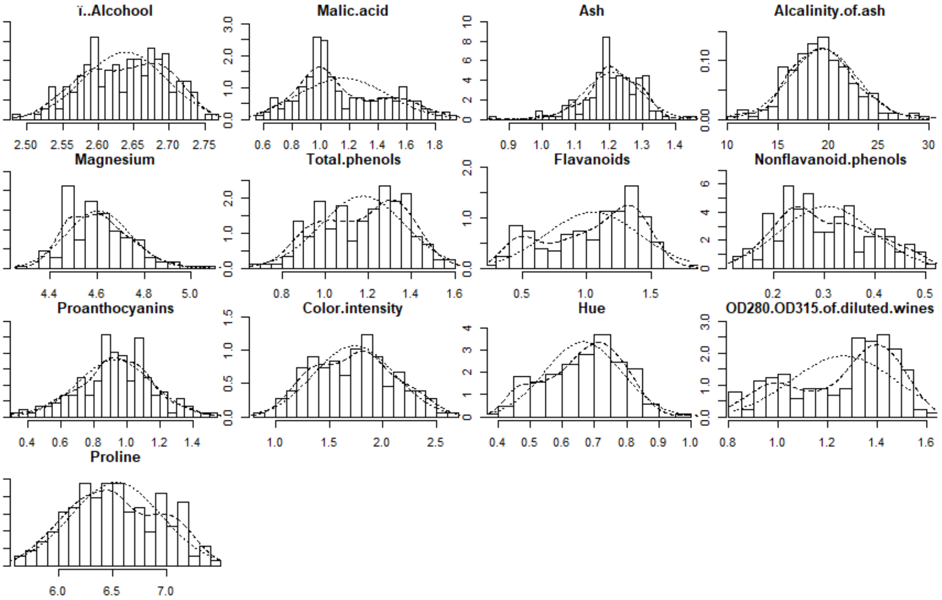

Figure 2: Histograms of all the variables after log-transform

3. For the robust distances we use the MCD estimator with parameter h. It represents the subsample of h observations of the original dataset (n samples) which minimises the determinant of the covariance matrix computed on these h points. We select a coverage of 75% of the sample size n which allows to compute the distances without the bad part of the dataset.

   The cutoff (represented by the red dotted lines on the graph) is determined by the χ^2 distribution with 13 degrees of freedom (number of quantitative variables we have) and a confidence level of 95%. An outlier is considered when the point exceed either the robust cutoff, either the Mahalanobis cutoff either both.

   We computed that for the classic Mahalanobis distance and we obtain 16 outliers, while we have 36 outliers for the robust one. This result is expected because with a non robust estimator, extreme outliers impact the mean and the covariance matrix and thus the classical Mahalanobis distance. So 20 Mahalanobis outliers which corresponds to 55% of the robust outliers are "masked" by the extreme outliers. Robust distances are less sensitive to those extreme outliers so we have less outliers.

   The right lower panel of the graph is empty, it confirms that no outlier is detected thanks to the classical Mahalanobis distance and not to the robust distance.

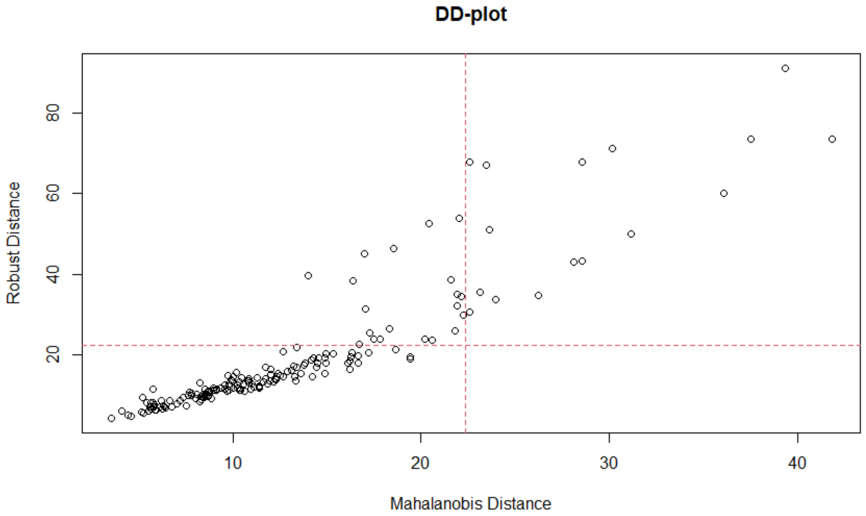

Figure 3: DD-plot of the classical and robust Mahalanobis distances

4. Here are the plots of the classical and the robust (estimated by the MDC estimator) correlation matrices. The first one is based on the whole sample size (n) and the second one is based on 75% of the sample size. Generally, the correlations are increased in the robust one compared to the classical one (with some exceptions), that can be explained by the fact that for the MCD, the best subset of the dataset is taken, meaning without the outliers. So this generally reinforces the correlations that were in the classic correlation matrix.

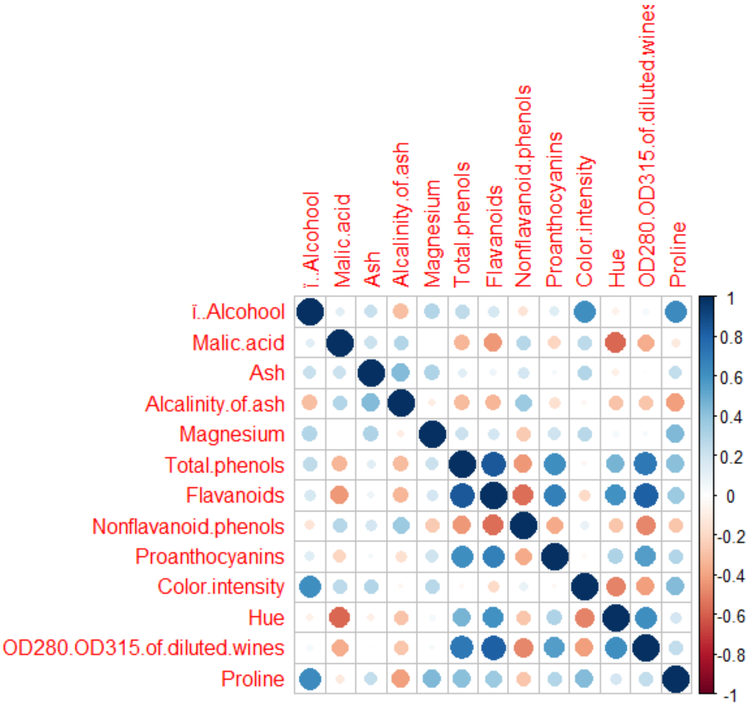

Figure 4: Classic correlation matrix

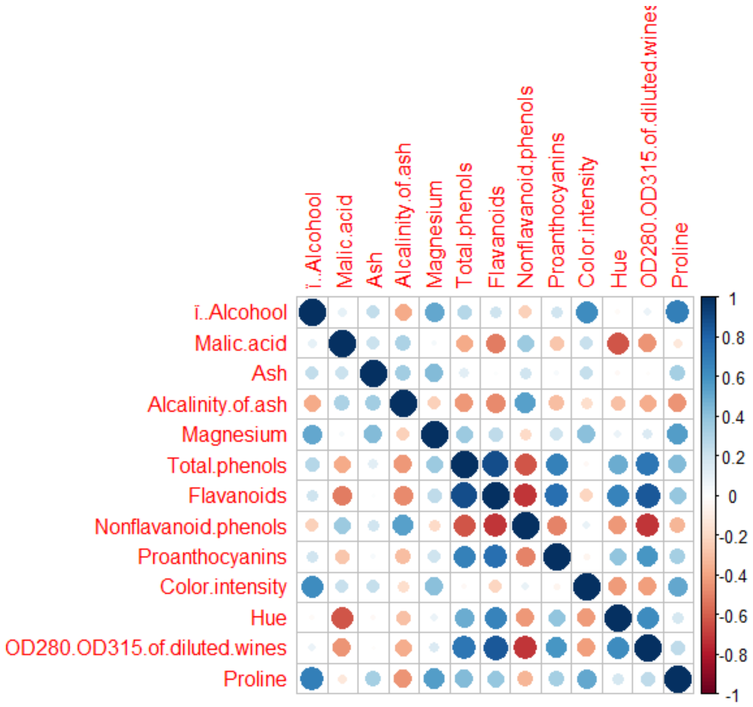

Figure 5: Robust correlation matrix

The most correlated couples are (using the correlation matrix estimated with the MCD estimator, so the robust one) :

- total phenols - flavanoids
- flavanoids - OD280/OD315 of diluted wines
- flavanoids - proanthocyanins

Their correlation factor is positive so the element of the couples are proportionnal, and if we plot the couples (one element on the x axis and the other on the y axis) we should see a linear relationship with a positive slope. Here are the graphs :

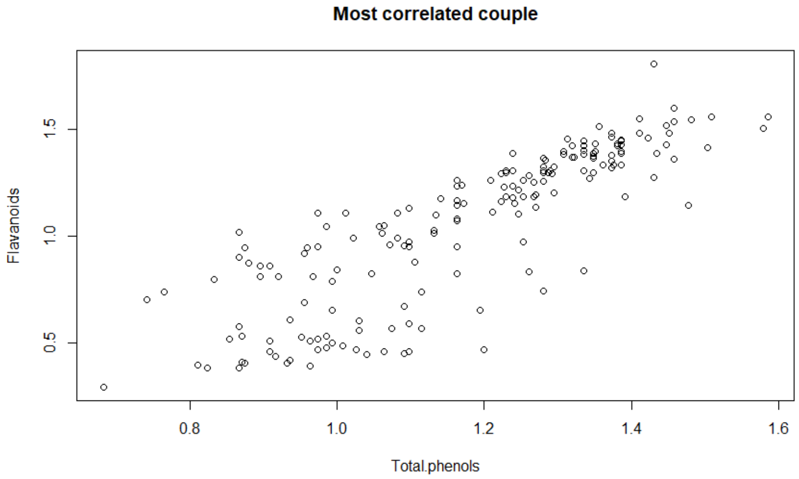

Figure 6: Total phenols - flavanoids

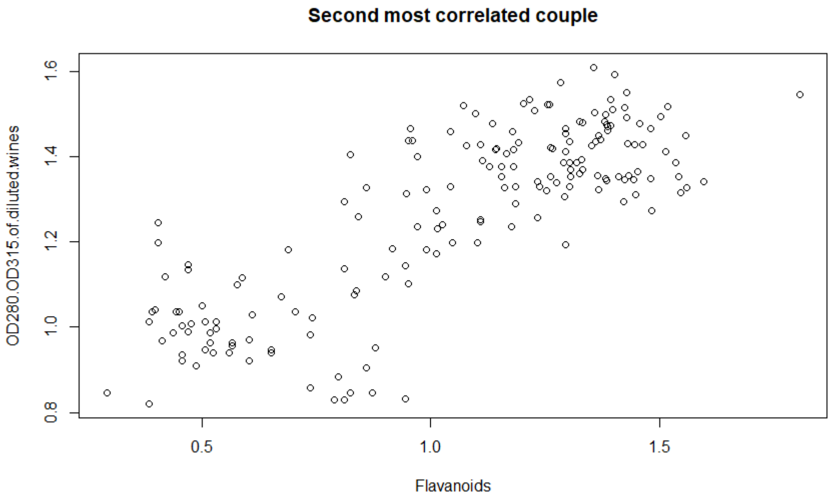

Figure 7: Flavanoids - OD280/OD315 of diluted wines

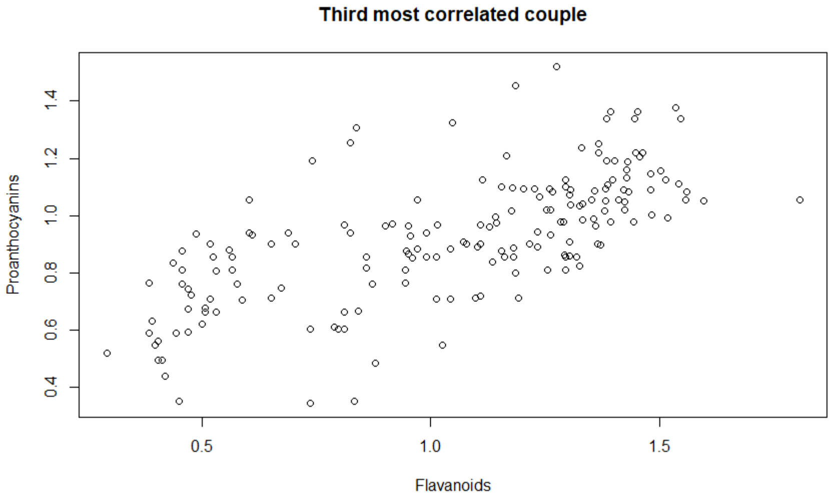

Figure 8: Flavanoids - Proanthocyanins

Indeed, the most linear relationship is between total phenols and flavanoids but even for the 2 others we see a relationship quite linear.
The strong relationship between total phenols and flavanoids was expected as explained in the first question. We can add that about 80%-90% of total phenols is flavanoids in red wine. They are also present in white wine but in a less quantity. This may explain the correlation and the fact that non-flavonoids are not in sufficient quantity to see a very strong correlation..
What we weren't expecting was the correlation between OD280/OD315 from diluted wines and flavonoids, but after some research we found out that flavonoids have the property of binding to proteins. So the correlation makes sense because

OD280/325 is a measure of protein content.
The correlation between flavanoids and proanthocyanins can be explained by the fact that proanthocyanins is just a type of flavanoids so the link is logic eventhough we didn't think about it.

Here is the scatter plot of the most negative correlated couple, nonflavanoids phenols - OD280/OD315 of diluted wines (which seems logical since nonflavanoids phenols and flavanoids are opposed (second biggest negative correlation) and flavanoids - OD280/OD315 of diluted wines is the second most positively correlated couple). We see a weak linear relationship, because even though it's the most negatively correlated couple, the correlation is still not really strong, with a negative slope because the correlation factor is negative, the two variables are inversely proportional.

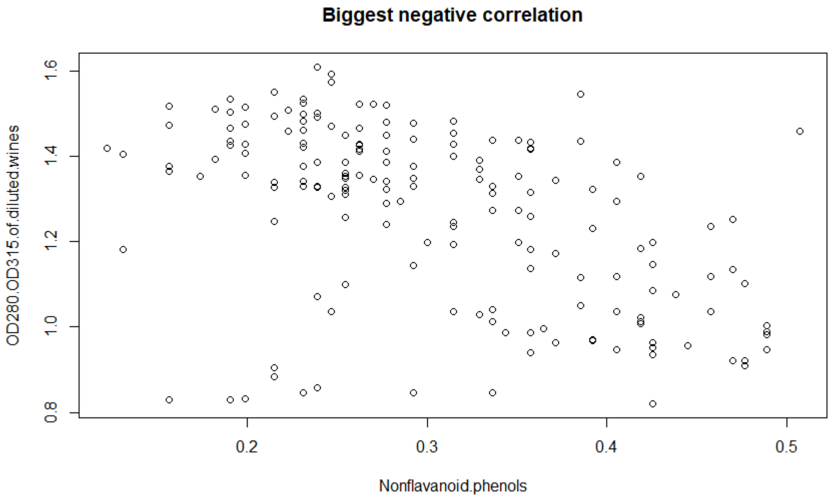

Figure 9: Nonflavanoids phenols - OD280/OD315 of diluted wines

5. First we are going to scale the variables to standardize them so they aren't influenced by the units of measure, and we are goingto work with the robust covariance matrix (estimated with the MCD estimator). We want to make appear at least the positive correlations we spotted in the previous question, and some others that we didn't spot because their correlation factor was smaller in absolute value, but they still have a link, showed thanks to the inverse of the covariance matrix. For that purpose we chose, for the penalization constant, $\lambda = 0.6$ after having tried some smaller and bigger values, it appears that this one was the best for us. The results fit pretty exactly the correlations we had, we have no correlation that disappears with this way of computing. Some others appear, because they had a big correlation factor, but still less than those that we spotted. The surprise is that some links stays when the link flavanoids - proanthocyanins disappear (e.g. for $\lambda = 0.7$ we still have a link between alcohol and proline)

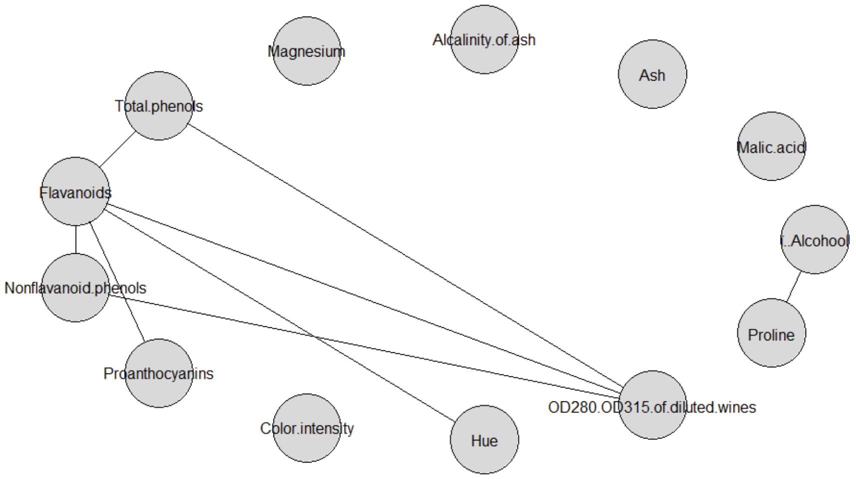

Figure 10: Graph of the edges

Here is the plot of the evolution of the number of edges on the graph with regard to the value of the penalization constant λ. Smaller the penalization constant is, more edges we have since we draw edges between the variables with a bigger inverse covariance than a certain threshold ($1e-16$ here, under that value the numbers are rounded to 0), and the penalization constant reduce this inverse covariance since it gives importance to the penalization function. For $\lambda = 0.9$ or 1, we see that there are no more edges drawn, since the only value that are above the threshold are the diagonal one, so the inverse covariance of a variable with itself.

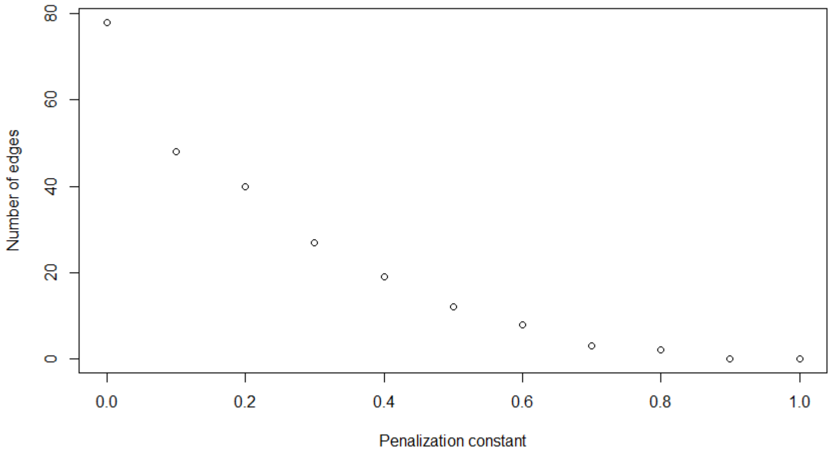

Figure 11: Number of edges with regard to the penalization constant