MVA - Project 6

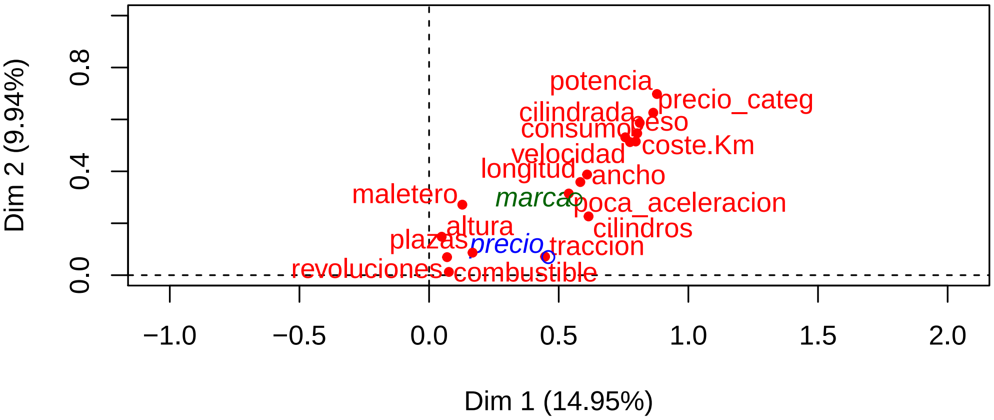
Carles Garriga Estrade i Balbina Virgili Rocosa

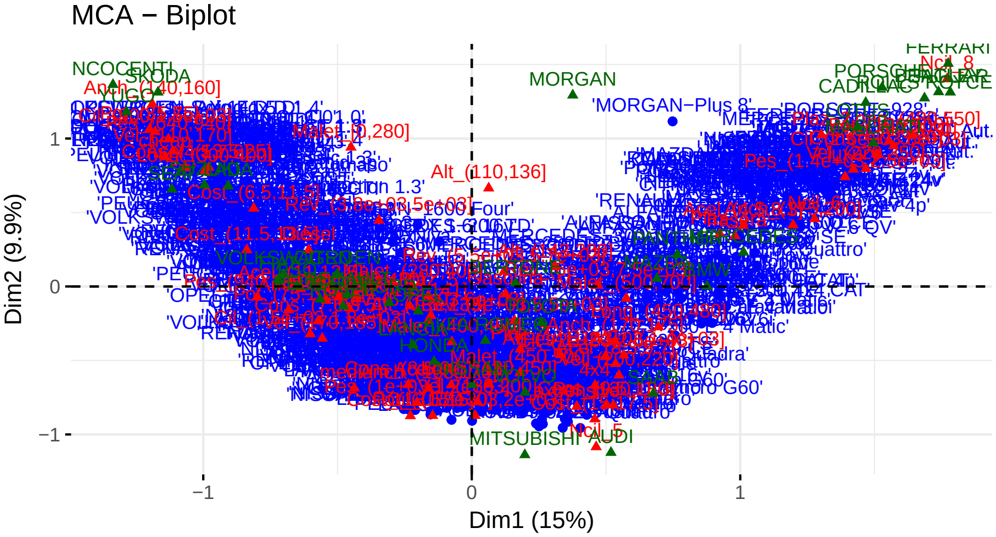
05/14/2018

All the code created for us to develop this assignment can be found on LAB06.R file. This file is located in the same folder of this document.

Exercise 1 First of all, the mca_car dataset has been read from the file. In order to be able to correctly manipulate it during the following exercises, missing values analysis has been performed. After checking the quantity of missing values of the dataset, non has been found so we have not need to apply any method to fill them.

Exercise 2 and 3 Then, we have performed a multiple correspondence analysis of the loaded dataset. To do it, we have used the MCA function of FactoMineR library and we have set brand as qualitative supplementary variable and price as quantitative supplementary variable. The number of dimensions kept in the result is modified to the number of variables of our dataset, 19. To be able to interpret the results obtained, we show the biplot and variables on the first factorial plane.





With the results obtained, first we can see that the accumulative variance of the first factorial plane is just the aprox. 25%, which is a low an non-really representative value. Then, looking into the showed biplot, we see the rows represented in red, the columns represented in green and the brands, which is defined as a qualitative supplementary variable, is represented in green. We also realize that the individuals represented on the first factorial plane form a shape of v. This parabolic shape of the cloud is known as Guttman effect, which denot a loaded diagonal. It does not mean that exists a quadratic relation between the principal variables but exists a linear relation between columns and rows.

Furthermore, by looking both graphics retrieved, the cars found on the top right corner of the biplot are the ones with higher potencia, precio_categ, cilindrada, peso, consumo, coste.Km and velocidad. It can be corroborated with the qualitative supplementary variable which shows brands such as Ferrari, Porche or Cadillac on that zone. On contrast, cars considered of low range are found on the top left corner, such as Skoda, Yugo or Innocenti. Also, the cars found on the central left bottom are the ones with more revoluciones, plazas, combustible, altura and maletero, such as Mitsubichi. And, finally, the cars considered as medium cars can be found on the central right bottom of the biplot.

Exercise 4 Now, the number of significant dimensions need to be chosen. One of the problems associated with MCA is the number of eigenvalues. In order to achieve a certain inertia (e.g: 90%), a lot of dimensions will be taken. However, we have reduced the dimensionality performing the following steps:

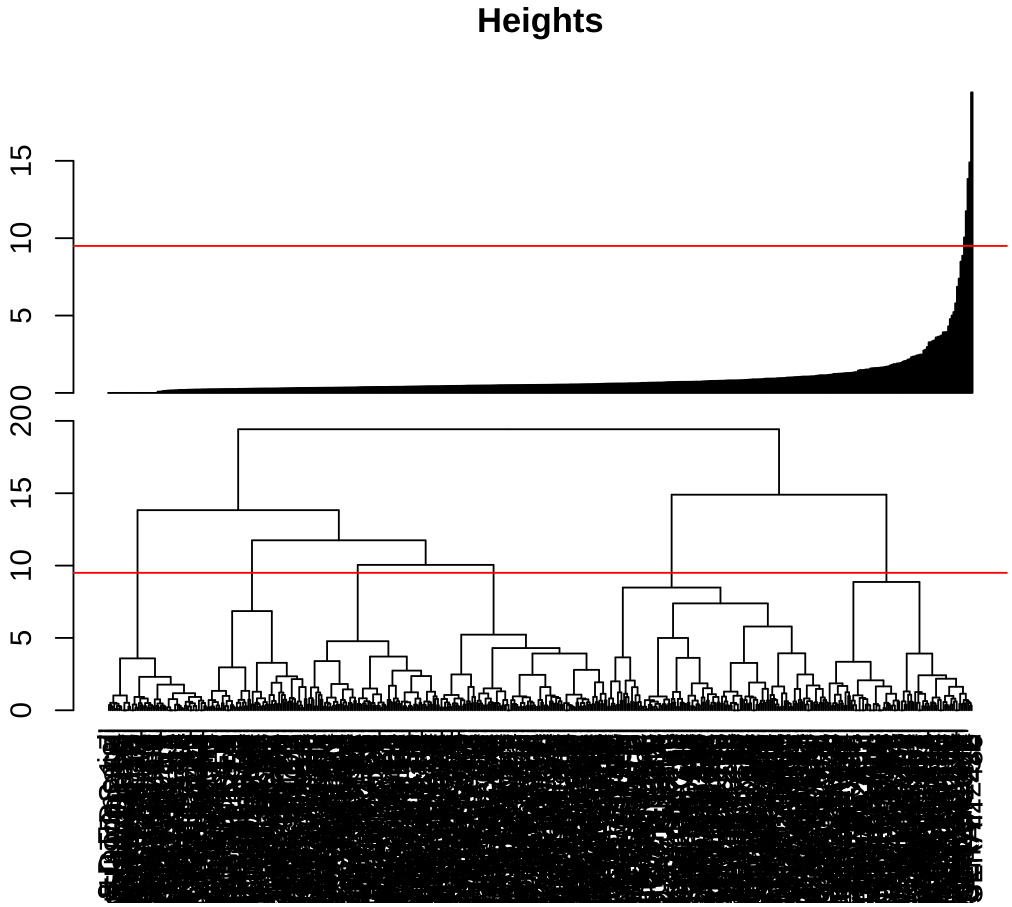
1.- Computing the mean value of the eigenvalues.

2.- Filtering the eigenvalues, extracting those greather than the mean.

3.- Computing the new eigen values by subtracting the mean to its value.

We have decided that the number of significant dimensions must be the first nine dimensions, by taking a look at the screeplot and having computed the number of dimensions necessary to ensure the 90% of the inertia (by using the cumulative percentage).

Exercise 5 Once the new number of significant dimensions is decided, a hierarchical clustering has been performed.



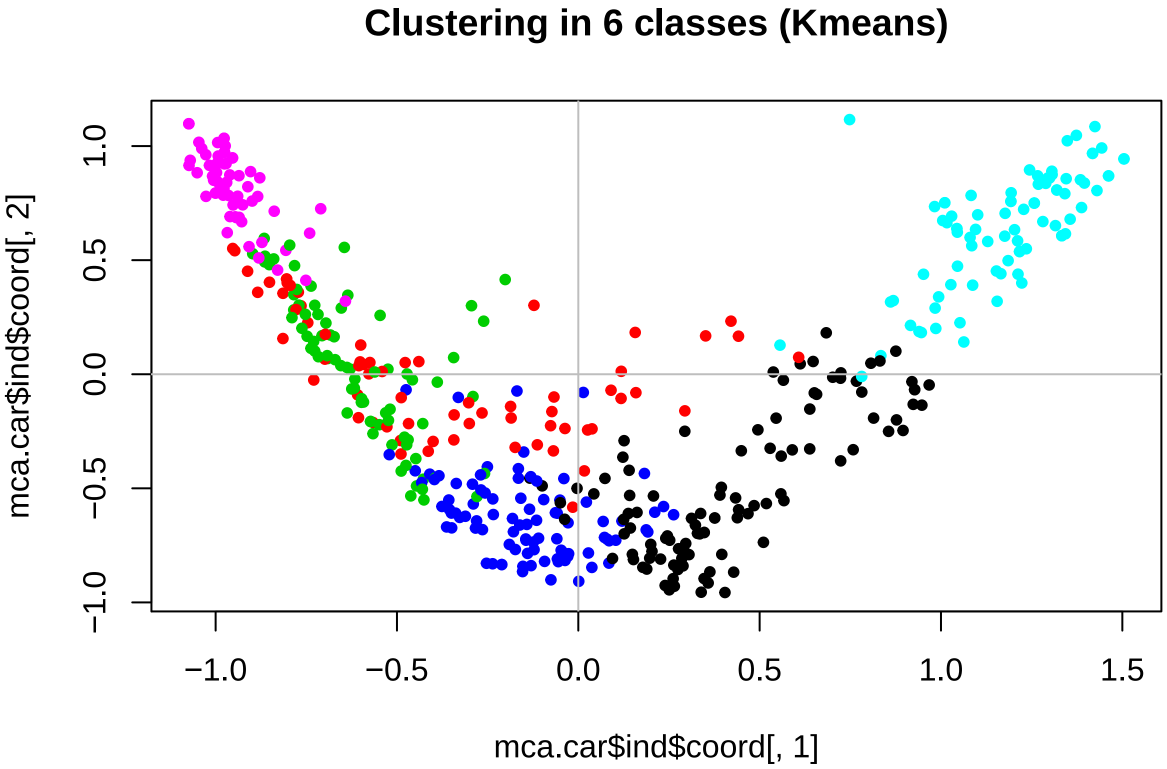
By looking at both the dendogram and the barplot of height, we have several partitions that could be feasable (2, 4, 5, 6 clusters and even more). But, we have decided to choose 6 as the most reasonable number of clusters.

After that, a consolidation has been applied by kmeans using the centers taken from the mca instance.

```
## [1] 55.29173
```

The computed quality of the partition of the data in 6 clusters is 55.29%.

Using the previous mca result, we have clustered the individuals in the first factorial plane by using the results from the kmeans generated clusters.



Exercise 6 After the clusters are defined, we need to interpret and represent them in the first factorial display using the function 'catdes' of the FactoMineR library. To do it, we need to pass to the function the defined clusters with the dataset used to define them. The number of response variables is set to 1.

With catdes function we want to realize which are the most significant categories on each clustered defined. This way, we are able to find out the profiling, which is the significant characteristics which make the group different than the other sets. For it, we first have a look on p.test calculated value. The p.test is less than 0.05 when one category is significantly linked to another categories, so we realize that all variables that catdes return are already below this stablized threshold. For this reason, we need to also have a look to v.test value, which is the one that gives us more information.

After ordering them by v.test value, we then state that if the v.test is positive, it means that the category is over-expressed for the category and if the v-test is negative it means that the category is under-expressed for the category.

With the results retrieved we can see that the first cluster is best expressed by cars considered as expensive with high speed and power, and high cost per Km and consumption.

The second one is best expressed by cars that use Diesel fuel, that are low reving and have a reduced cost per Km aswell as a low consumption.

The third cluster is best expressed by cars considered cheap with medium power (75, 105), consumption and cost per Km but consideribly medium speed (170 - 185 km/h).

The fourth cluster is best expressed by cars considered as medium price range, with higher power, cost per Km and speed than cars of third cluster.

The fifth cluster is best represented by cars considered as luxury which have the best features, such as power, speed and displacement but at the same time also have the highest cost per Km and consumption.

Finally, the sixth cluster is best represented by cars considered cheap, light cars with low power, low to medium speed and small displacement. Also, they have a width between 1.40m to 1.60.