Multivariate Analysis

Joan Prat Sicart

April 22th, 2019

Practice MCA and Clustering

The aim of this practice is to study the "mca_car" dataset, which contains information about the cars and their characteristics found in specialized magazines. The final goal will be to find a **model to predict the price of cars as function of its characteristics**. First we will perform a visualization of the information contained in the dataset using the **Multiple Correspondence Analysis**, then we will perform a **Clustering** of the cars. Note that the data has been previously preprocessed to have it in categorical form.

Read the dataset

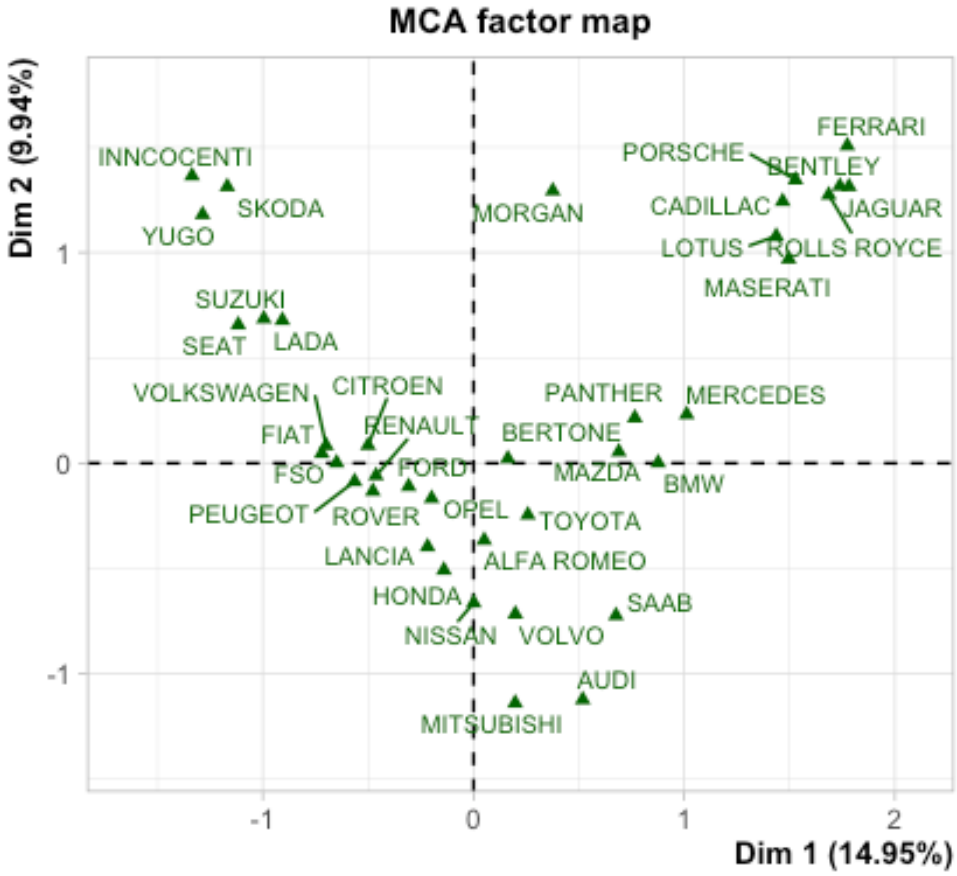
The first step consists in loading the new dataset with R, but first of all we are going to change the row number 416: 'SEAT-Ibiza Special 0;9' to 'SEAT-Ibiza Special 0.9' since the ";" is giving problems when loading the dataset.

```r
# Read the data as dataframe
path_MCA = "/Users/joanpratsicart/Documents/miri_DS/3_quatri/MVA/lab6/mca_car
.csv"
Data_MCA = read.csv(path_MCA, header = TRUE, sep=";")
# Use the identifier (1st column) as the name of each row
row.names(Data_MCA) = Data_MCA[,1]
# Keep the remaining 19 dimensions
Data_MCA = Data_MCA[c(2:20)]
# Capitalize the name of the columns
names(Data_MCA) = toupper(names(Data_MCA))
```

Obtain a visualization of the information contained in the data, by performing a Multiple Correspondence Analysis

once the data has been loaded, it's going to be performed the Multiple Correspondence Analysis, however, to do so we won't consider the brand and price variables, so are going to be defined as supplementary variables.

```r
library("FactoMineR")
# Perform the MCA
mca.Data = MCA(Data_MCA, quanti.sup = c(17), quali.sup = c(18), graph = FALSE)
plot(mca.Data, invisible=c("ind","quanti.sup", "var"),cex=0.7)
```



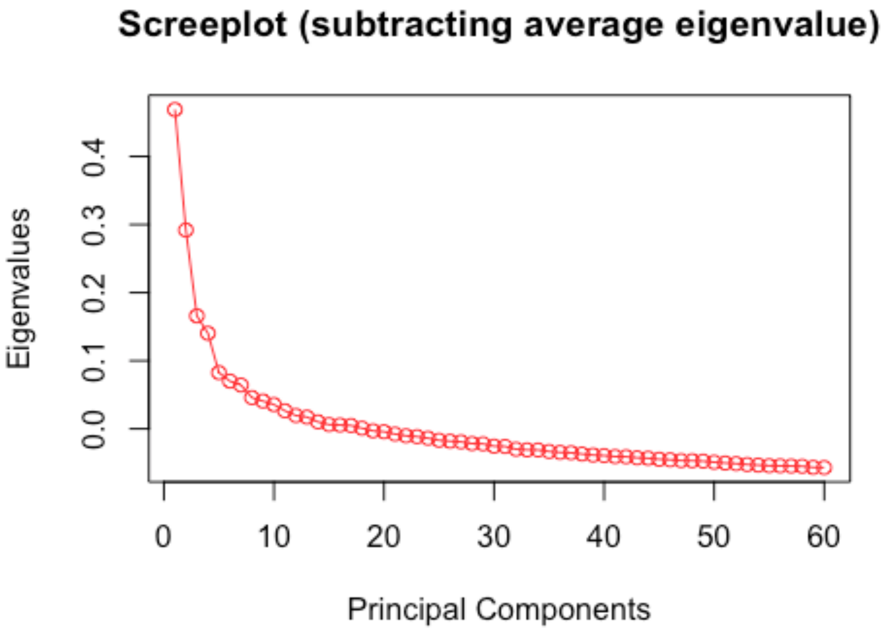
Interpret the first two obtained factors

As it can be appreciated the MCA is no very reliable, since the **two first dimensions account for a total variance of 24.89%**, however, it's possible to clearly detect the **"Guttman effect"**, sincd as it can be appreciated in the image above the distribution of the cloud of points has a shape of a “V” and from which is possible to extract the following knowledge: - **the first axis opposes the values of the brands**, that is why in the left there are the cheap cars and in the right we have the more expensive cars. -**The second axis opposes the intermediate values with the extreme ones**, since in the bottom there are moderated expensive cars like Audi, and super expensive cars like Ferrari in the top.

Decide the number of significant dimensions that you retain

As in the previous lab selecting the significant dimensions it’s a cornerstone of well performed hierarchical clustering, so first of all, it’s going to be performed a screeplot where we have subtracted the average eigenvalue.

```
# Average eigenvalue
average_eig = sum(mca.Data$eig[,1]) / length(mca.Data$eig[,1])
# Screeplot
plot(mca.Data$eig[,1] - average_eig, type = "o", xlab = 'Principal Components', ylab = '
Eigenvalues', main = 'Screeplot (subtracting average eigenvalue)', col = 'firebrick1')
```



There are some techniques available in the literature to decide how many significant dimensions should be retained (e.g., Kaiser rule, Last elbow rule, etc.).

Since the last elbow of the screeplot is quite clear, we decide to use this criterion and according to the screeplot this means taking the first **5 dimensions**.

Perform a hierarchical clustering with the significant factors

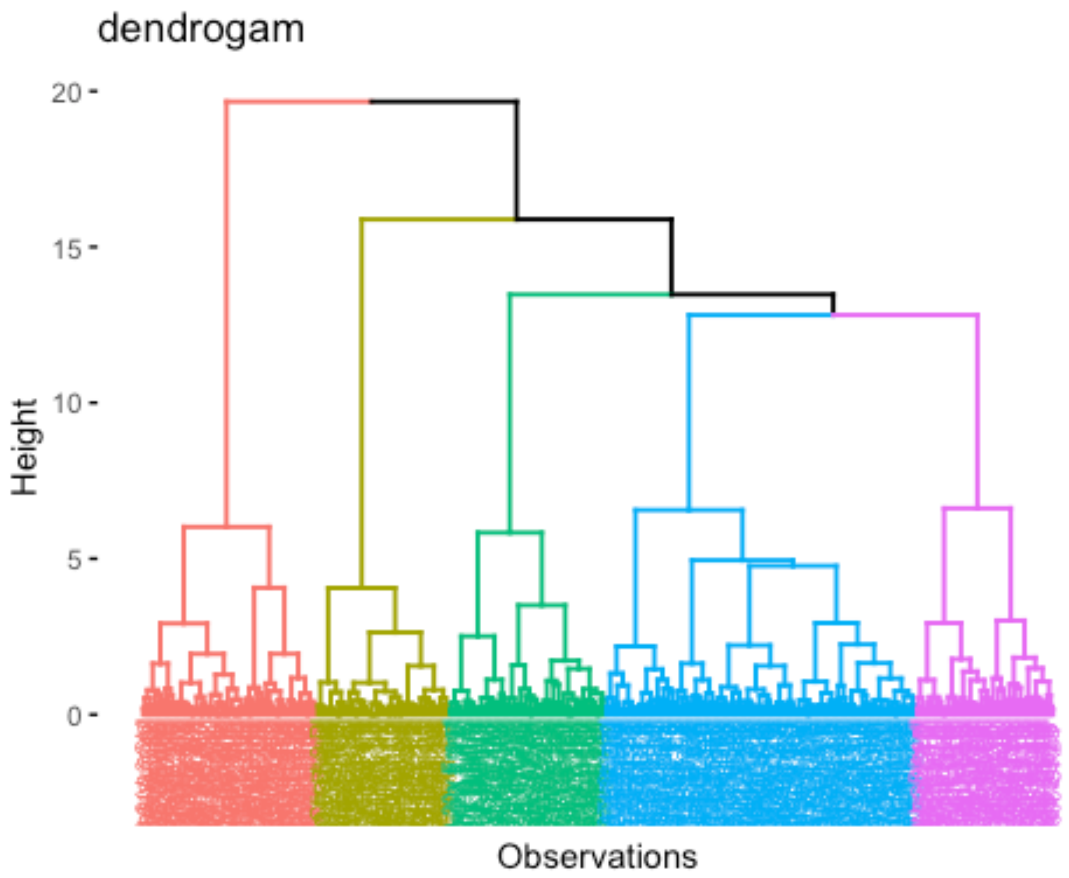
First of all, it's going to be done a **hierarchical clustering** with 5 factors as it has been jusfitified previously.

```
library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3
WBa

# Number of significant dimensions
n_dimensions = 5
# Save the coordinates in the first n_dimensions
coordinates = mca.Data$ind$coord[,1:n_dimensions]
# Hierarchical clustering using Ward method
hierarchical = hclust(dist(coordinates), method="ward.D2")
# Plot the dendrogam with colors
fviz_dend(hierarchical, cex = 0.5, k = 5, xlab = "Observations", main = "dendrogam")
```



The second phase will consists in **selecting the number of classes**. and as it can be appreciated in the dendrogam, it should be considered dividing the data set in **5 clusters**.

```
# Save the coordinates in the first n_dimensions
coordinates = mca.Data$ind$coord[,1:n_dimensions]
# Cut the hierarchical tree
clusters = cutree(hierarchical, 5)
# Create a dataframe with the original data + the cluster
df = data.frame(coordinates, CLUSTER = as.factor(clusters))
```

Once teh the number of classes has been selected, the **centroids** of the resulting clusters will be calculated.

```
# Calculate the centroids with hierarchical
centroids = aggregate(coordinates, list(df$CLUSTER), mean)
```

-And finally, in the last step it's going to ve performed the **k-means** algorithm taking as seeds the centroids previously calculated.

```r
# k-means taking as seeds the centroids calculated with hierarchical
k_means = kmeans(coordinates, centers = centroids[,-1])
# Update the dataframe with the coordinates + the k-means cluster
df = data.frame(coordinates, CLUSTER = as.factor(k_means$cluster))
```

Interpretation of the clusters

In order to **interpret the clusters** obtained in the previous steps, it's going to be represented in the **first factorial plane (Dim 1 and Dim 2)**. (using the default significance threshold of 0.05 to characterize the clusters).

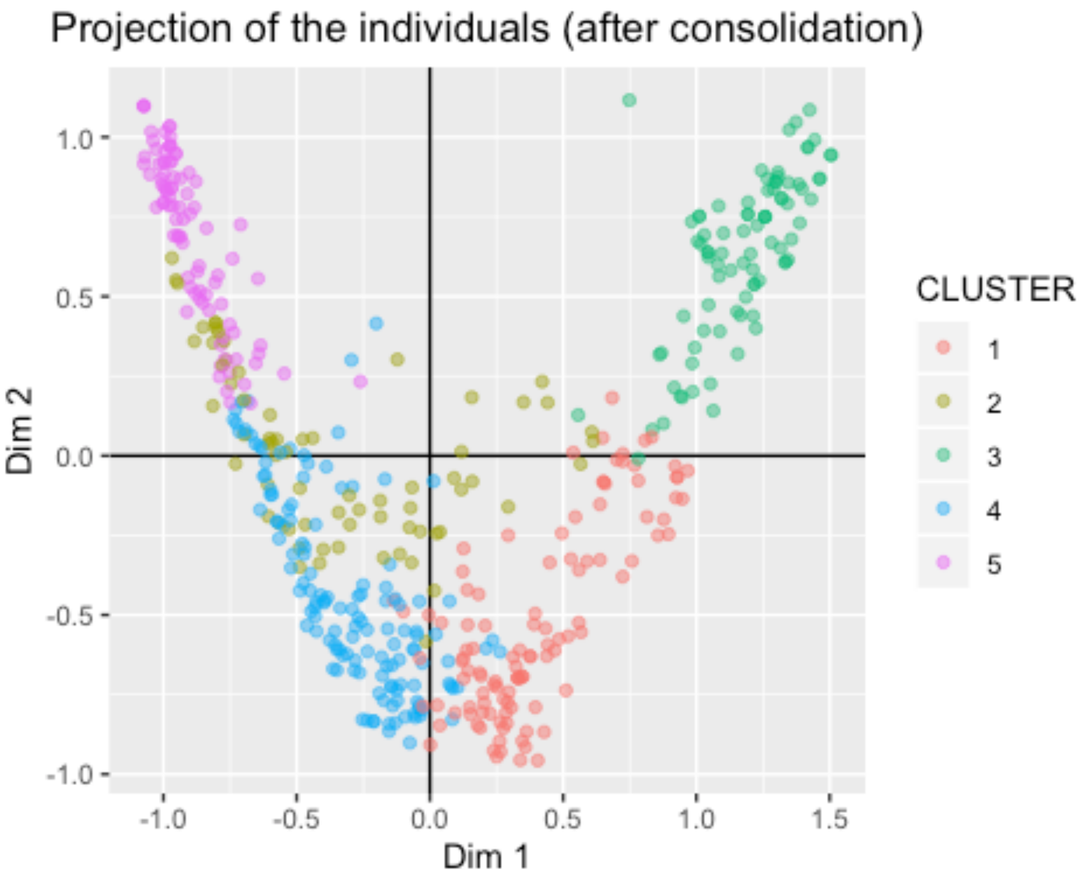
```r
# Use only the coordinates and clusters of Dim 1 and Dim 2 (first factorial plane)
df_2 = df[,-c(3:5)] # df was the data frame containing coordinates in 5 dimensions + clus
ter number
intepretation = catdes(df_2, num.var = length(df_2), proba = 0.05, row.w = NULL)
intepretation
```

According to the results, we obtain the following information:

- **Cluster 1** is positively correlated with dimension 1 and negatively with 2. These are **moderated expensive cars**.
- **Cluster 2** is negatively correlated with dimension 1. These are **intermediate cars**.
- **Cluster 3** is positively correlated with dimension 1 and 2. These are **super expensive cars**.
- **Cluster 4** is negatively correlated with dimension 1 and 2. These are **moderated cheap cars**.
- **Cluster 5** is negatively correlated with dimension 1 and positively with 2. These are **very cheap cars**.

Finally, to make it more visual it will the clusters will be plotted in different colors.

```r
# Plot the clusters with different colors
library(ggplot2)
ggplot(data = df, aes(x = Dim.1, y = Dim.2)) +
 geom_hline(yintercept = 0, colour = "black") +
 geom_vline(xintercept = 0, colour = "black") +
 geom_point(aes(colour = CLUSTER), alpha = 0.5) +
 labs(x = "Dim 1", y = "Dim 2") +
 ggtitle("Projection of the individuals (after consolidation)") +
 theme(plot.title = element_text(hjust = 0.5))
```



Conclusions

To conclude, in thsi homewokr it has been possible to observe how **Multiple Correspondence Analysis** can help for tables of individuals with qualitative variables and how MCA can be used to pre-treat data before doing classification.