Assignment 6 of Multivariate Analysis (MVA)

Pol Serra i Lidón
Qiaorui Xiang

1 Introduction

This work about Multivariate Analysis is focused on the concept of **Multiple Correspondence Analysis** (MCA). In concrete, the MCA analysis of the *mca_car.csv* data file, that contains information regarding some characteristics of different cars.

2 Practice

2.1 Read the data

The *mca_cars.csv* data file is read with *read.csv* function.

2.2 Perform MCA

MCA is performed upon the dataset. As we can observe in Figure 1, the first 2 dimensions represent the 24% of the cumulative variance.

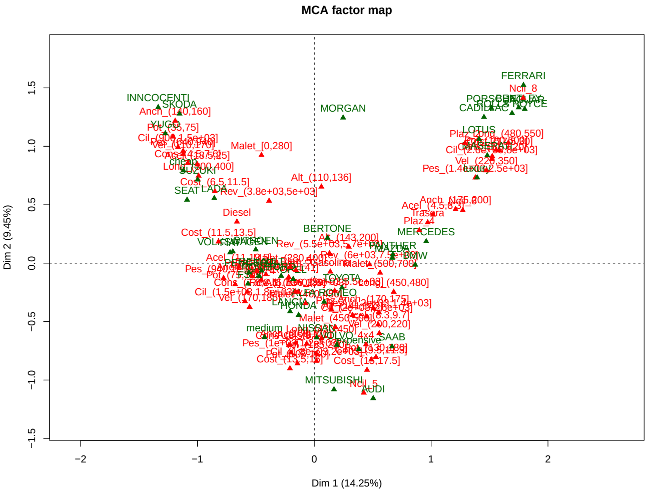


Figure 1: Resuting factor map from the MCA.

2.3 Interpretation of first two factors

In order to interpret the first two factors obtained in the MCA, we have made use of the function *dimdesc*. In Table 1, there are the first and last 4 category factors in terms of P-value, regarding both the first and the second factors.

As the first category factors in dimension 1 are high cost, luxury, cylinders, and high power, and the four last ones are low velocity, low number of cylinders (4 is low, 8 is high), and cheap price, we can conclude that the first

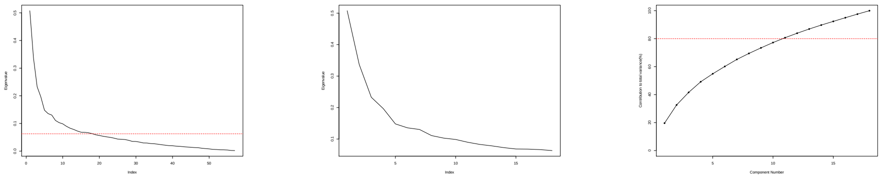
First factor 1	Last factor 1	First factor 2	Last factor 2
Cost_(17.5,30]	Vel_(110,170]	Pot_(35,75]	Pot_(105,130]
luxury	Delantera	Vel_(110,170]	Cil_(1.8e+03,2e+03]
Cil_(2.6e+03,8e+03]	cheap	Anch_(140,160]	Cost_(15,17.5]
Pot_(180,500]	Ncil_4	Pes_(640,940]	Pes_(1e+03,1.2e+03]

Table 1: First and last 4 category factors, on both the first and the second dimensions.

dimension in MCA represent how expensive and powerful is each car. The second factor that does not represent even the 10% of the variance, lacks of a separate interpretation from factor 1.

2.4 Decide number of significant dimensions

As it can be observed in Figure 2, we have taken as candidate dimensions all the dimensions above the average, 18 in this case. Then we select those dimensions that explain 80% of variance among the candidates. Finally, we only consider 11 dimensions as significant dimensions.



(a) All the dimensions

(b) The 18 dimensions above the average

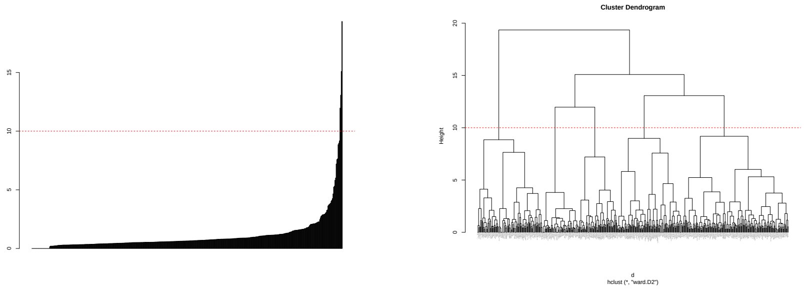
(c) Contribution to total variance

Figure 2: Scree plot representing the eigenvalue of each dimension

2.5 Hierarchical clustering and consolidation

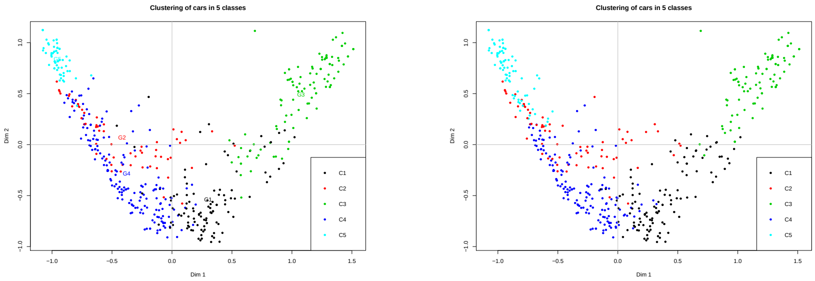
In order to perform the hierarchical clustering and consolidation, we computed the aggregated distance between factors, and analyzed it together with the dendogram. After observing the highest jump in Figure 3 (a), we have divided the observation in 5 clusters, as it can be observed in Figure 3 (b).

The clustering and consolidation are realized, and the results can be observed in Figure 4.



(a) Aggregated distance (b) Dendogram

Figure 3: Aggregated distance and dendogram.



(a) Clustering (b) Consolidated Clustering

Figure 4: Clustering and consolidated clustering of the car models.

2.6 Function *catdes*

After applying the *catdes* function to the clustered data, we observe the principal characteristics of each clusters in Figure 5.

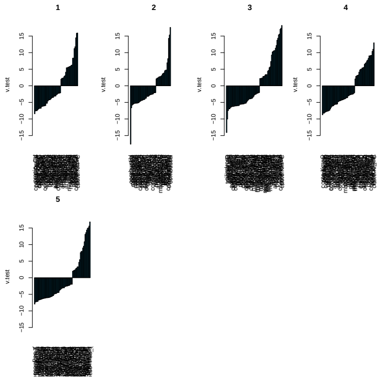


Figure 5: Catdes plot of the 5 obtained clusters

Furthermore, the number of cars in each cluster that are present in each possible price variable range can be observed in Table 2.

	Cheap	**Medium**	**Expensive**	**Luxury**	
Cluster 1	0	10	74	23	107
Cluster 2	13	30	20	8	71
Cluster 3	0	0	10	76	86
Cluster 4	35	94	17	1	147
Cluster 5	75	4	0	0	79
Total	123	121	108	138	490

Table 2: Distribution of Cluster in terms of price category

We could name the clusters from left to right as:

1. **Junk cars**: The ones in *Cluster 5.* Almost all of them are cheap cars with very poor features.They could be portrayed as old vehicles, the type of car to drive for students, or people with very low income.

2. **Poor cars**: The cars in *Cluster 2.* These cars are situated between the prices of cheap, medium and even expensive, which do not tend to present specially good features. *Seat* or *Skoda* are brands that we consider to be representative of this cluster, and they can consider them as cars for young people.

3. **Average cars**: The ones in *Cluster 4.* This cluster agglutinates the cars that are in the average in all the aspects. They do not present special or distinctive characteristics, and the majority of them are familiar cars for medium-high class workers. The majority of them is in the medium price range, like *Volskwagen* or *Citroen.*

4. **Executive cars**: The ones in *Cluster 1.* This cars tend to present a high cost, and come with high quality characteristics. *Mercedes* and *BMW* are brands that we consider very representative of this cluster.

5. **Affluent cars**: The ones in *Cluster 3.* They tend to be in the luxury price range. Almost all of these cars are very expensive, and are not possible to afford for the majority. They present the best features between all of them. *Ferrari* or *Porsche*, are very representative brands of this cluster of luxurious cars.