Multiple Correspondence Analysis and Clustering

Author: Cedric Bhihe Date: 2018.05.14

Given data about cars and their characteristics, the goal of the analysis to be performed is to find a model to predict the price of cars as a function of their characteristics.

1. Read the file "`mca_car.csv`". The data has been previously pre-processed to have it in categorical form. Perform a visualization of the information contained in the dataset, then a clustering analysis.

The `cars` data set contains 490 observations, with no missing. Each observation is considered active and consists of 19 attributes of which 18 are categorical or ordinal variables and 1 is continuous `precio`. We consider `precio`, `precio-categ` and `marca` as 2 response variables and 1 descriptive variable respectively. We will treat all three as supplementary.

One way to conduct preliminary data visualization is to inspect 2D tables such as 'marca' and 'precio-categ' to derive product segmentation information per brand (car make). The plot of Fig.1 shows normalized row counts for each brand, where the normalization factor is the row marginal. *(Maximum cumulative barplot values slightly in excess of or below 100% are due to rounding errors.)*

Fig.1: Product segmentation per brand

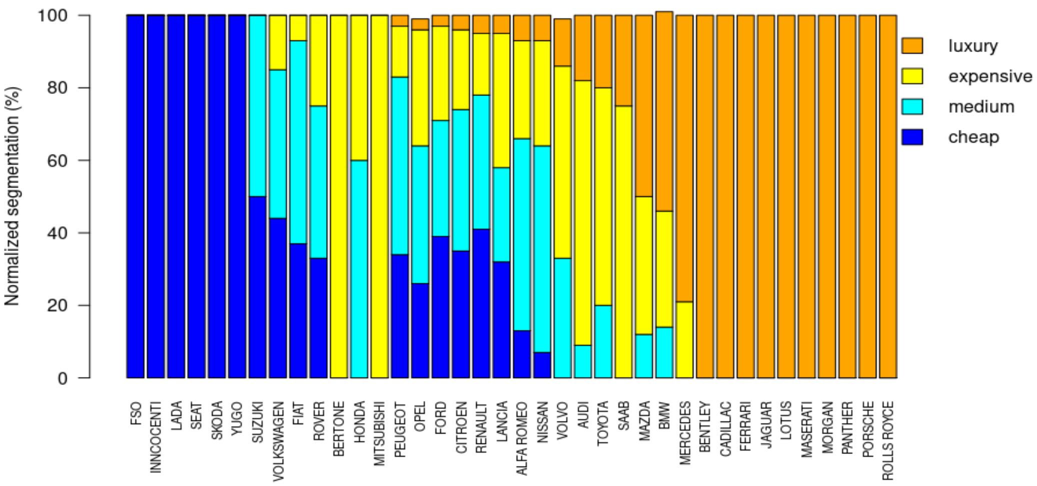

Table 1 (right) shows how mean cylinder displacement increases and mean gas consumption decreases as product category increases. This marked scissor effect could be principally attributed to higher-end vehicles relying on hybrid or electric technology. Only 4 car makes are shown here for more clarity.

Car make	Product Segment	Mean(cylDisp) (liter)	Mean gasConsum (liter/100km)
ALFA-ROMEO	cheap	1.000000	4.500
CITROEN	cheap	1.250000	3.500
ALFA-ROMEO	medium	2.500000	4.375
CITROEN	medium	2.777778	3.000
VOLVO	medium	2.200000	3.200
ALFA-ROMEO	expensive	4.250000	2.250
CITROEN	expensive	3.800000	3.200
VOLVO	expensive	2.750000	2.250
ALFA-ROMEO	luxury	5.000000	1.000
BENTLEY	luxury	5.000000	2.000
CITROEN	luxury	5.000000	2.000
VOLVO	luxury	4.000000	2.000

Table 1: Summary statistics of data subsets

2. With the obtained data frame perform a Multiple Correspondence Analysis.

We consider variables 18 'marca' and 19 'precio-categ' as supplementary categorical variables, while variable 17 'precio' is kept as a supplementary continuous variable. This means they do not take part of the MCA and their coordinates will be predicted. They are the reponse variables. All other variables are kept as active variables (i.e. there is no junk category). We also choose to keep (`ncp=`) 7 explanatory dimensions and not to ventilate any category (`level.ventil=0`), no matter how small its level of representation in the final result may be. The MCA is conducted on the Indicator matrix rather than the Burt table. Figures 2 and 3 below summarize results.

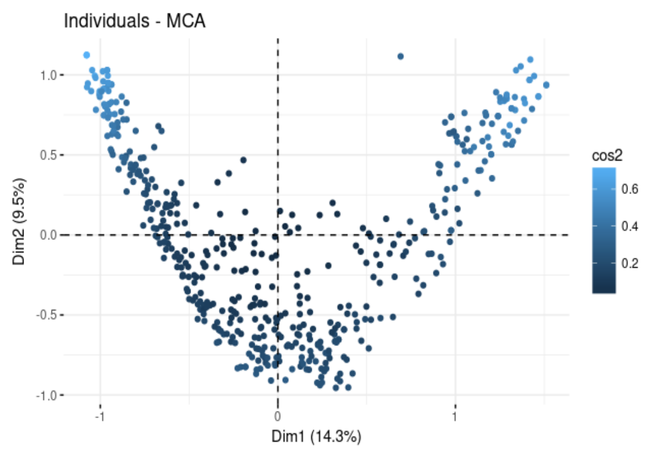

***Fig. 2a**: Scatterplot of individuals, with color coded quality of representation (cos²) on PC1-2. Lighter blue means better representation than a darker shade.*

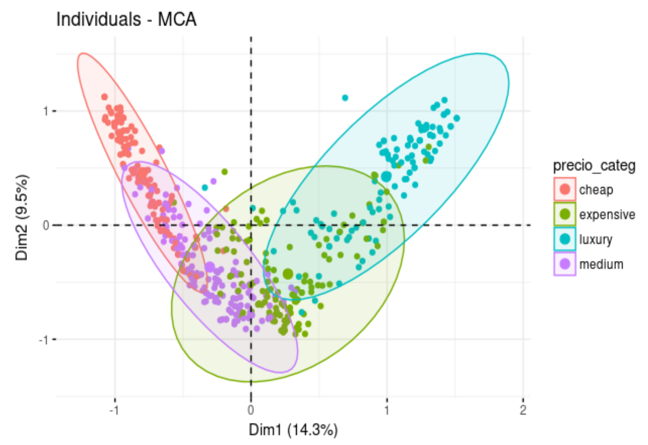

***Fig. 2b**: Scatterplot of individuals, with color coded grouping according to the categorical variable "precio_categ". Four partly overlapping groups are clearly visible.*

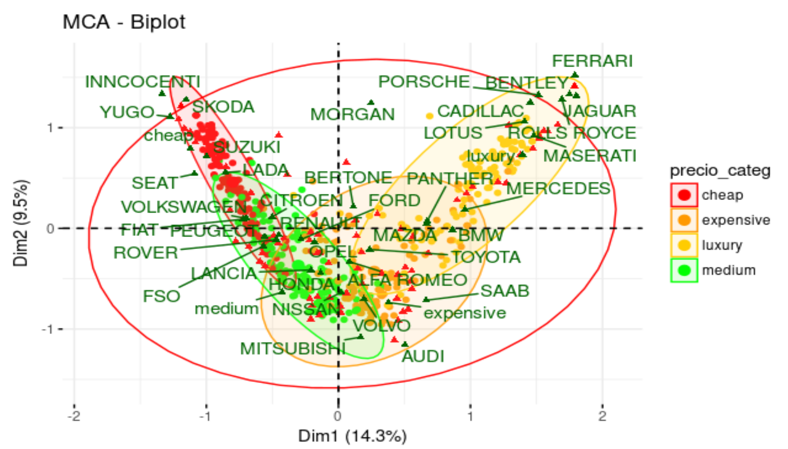

***Fig. 3**: Biplot of individuals (color-coded id. legend), supplementary categorical variables, "marca" and "precio_categ" (green), and variables' categories (red).*

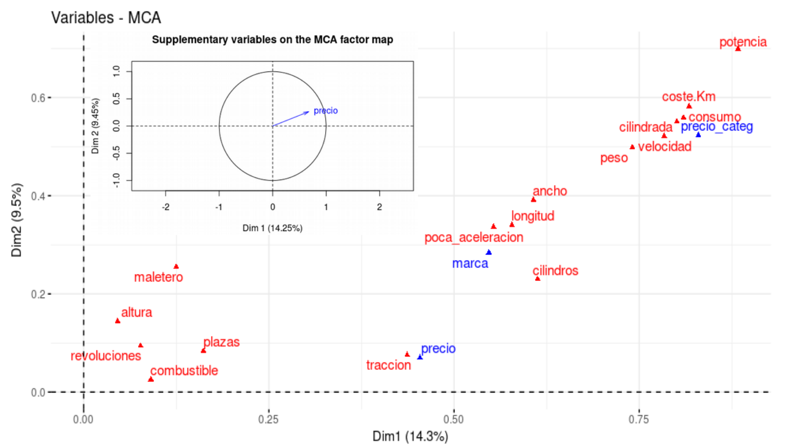

***Fig. 4**:Active variables' scatter plot (red), and supplementary variables "marca", "precio_categ" (categorical), and "precio" (continuous) in blue.*

3. Interpret the first two obtained factors.

- The first factorial plane, represents about 24% of the variability in the data set.

- A Guttman effect is visible in Figures 2 and 3. It points to diagonal overloading of the RFT (Relative Frequencies Table). The crescent shaped scatter plot of individuals and categorical variable "precio_categ" indicate that cheap, entry level vehicles are in the second quadrant, mid-range vehicles ("medium" to "expensive") are in the 3rd and 4th quadrants while up-scale, high-end cars are in the first quadrant.

- Figure 3 also demonstrates how modalities of the categorical variables "marca" and "precio_categ" are the pseudo-barycenters of the individuals exhibiting the given modality.

- In Figure 4, distance of variable points from the origin is a measure of factor quality. Points farther from the origin indicate variables which account for more inertia and are better represented in that plane than variables closer to the origin. It is the case of "potencia", "coste km" and "consumo", as compared with "revoluciones", "maletero", "altura", "plazas".

- The closer to a PC axis any segment linking the axes' origin and the variable points, the greater the correlation between the variable and PC. Hence "potencia", "coste km", "consumo" appear slightly more correlated with PC1 than with PC2 and are of good quality, whereas "traccion" appears preferentially correlated with PC1 but is of mediocre quality. A discussion of correlation for the group close to the origin (“revoluciones”, “maletero”, “altura”, “plazas”) is obviated by the fact that variables in that group are not well represented in PC1-2.

■ As for individuals between themselves, variables in close proximity to one another have a similar profile and explanatory power. Thus:
(a) the categorical variable “precio_categ” seems to be well predicted by the group of variables in the upper right corner of Figure 4.
(b) The continuous variable “precio” is somewhat isolated and close to “traccion”. We interpret this as “precio” not being as good an ***estimator of vehicles*** as “precio-categ” as the price tag of a car is highly sensitive to a number of extraneous parameters such as time of year, car-dealership and location, buyer, vehicle options, whereas its price category is not.
(c) “marca” appears to be somewhat predictable based on design criteria, such as “cilindros”, “longitud”, “ancho”, and “poca_aceleracion”, the latter being the result of deliberate engine, gear-box and power-consumption profiling, often based on marketing.

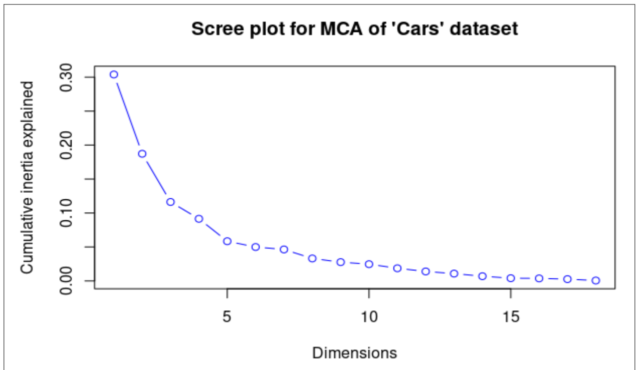

Fig. 5*:Cumulative inertia as explained by eigenvalues.*

4. Decide the number of significant dimensions that you retain (by subtracting the average eigenvalue and represent the new obtained eigenvalues in a new screeplot).

Using the elbow rule on Figure 5 (right), we choose to retain nd=6 significant dimensions, which coincides with 80.7% of inertia being explained.

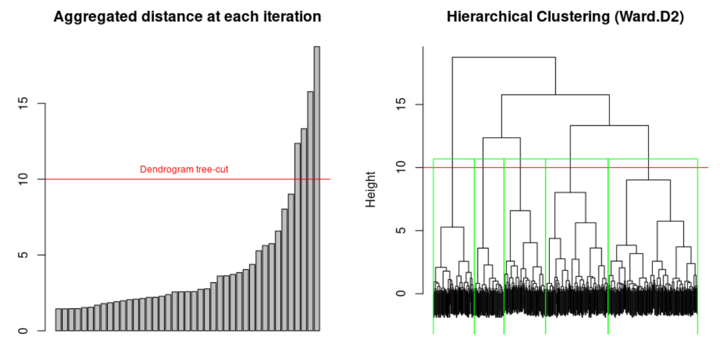

Fig. 6a*:Aggregated distance at each iteration (left) and HC tree cutoff at 5 classes (right).*

5. Perform a hierarchical clustering with the significant factors, decide the number of final classes to obtain and perform a consolidation operation of the clustering.

After building a distance matrix from the data set of individuals’ projections on the PC space (restricted to 6 dimensions), we perform a hierarchical clustering based on the Ward.D2 distance (Figures 6a).

We represent the color-coded clustering of observations in Figure 6b below. In addition to the four known modalities “cheap”, “medium”,”expensive” and “luxury” roughly equivalent to clusters #1, 3, 4 and 5, it appears that cluster #2 mainly consists of Diesel engine vehicles.

Cluster sizes <u>before consolidation</u> are (from cluster 1 to 5): 116, 77, 76, 166, 55, and the quality index is *Ib=60.97*.

<u>After k-means consolidation</u> the quality index increases to *Ib=64.28*. The sizes of consolidated clusters are now (from cluster 1 to 5): 114, 68, 82, 146, 80. The new scatter plot of clustered observations’ projections on PC1-2 (Figure 7) is very similar to that of Figure 6b, prior to k-means consolidation.

They exhibit a few misclassified observations as demonstrated by the silhouette graph of Figure 8.

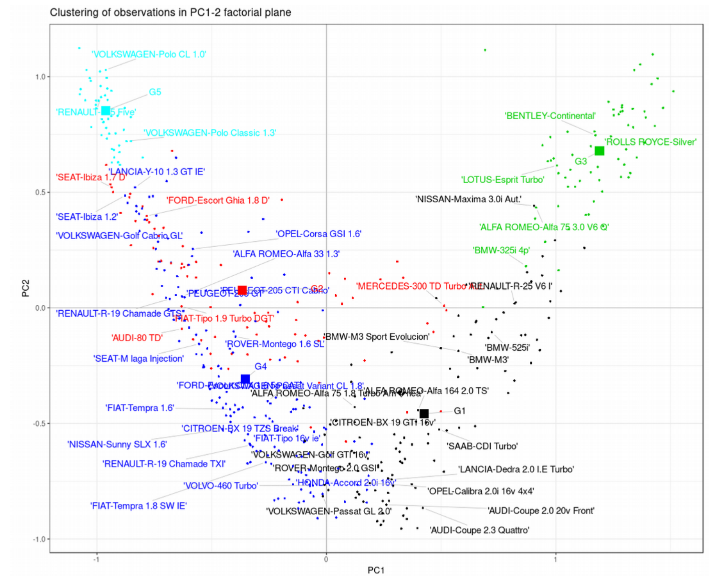

Fig. 6b*: Color-coded clustering of observations with centroids, according to HC, before consolidation. 10% of all labels are shown.*

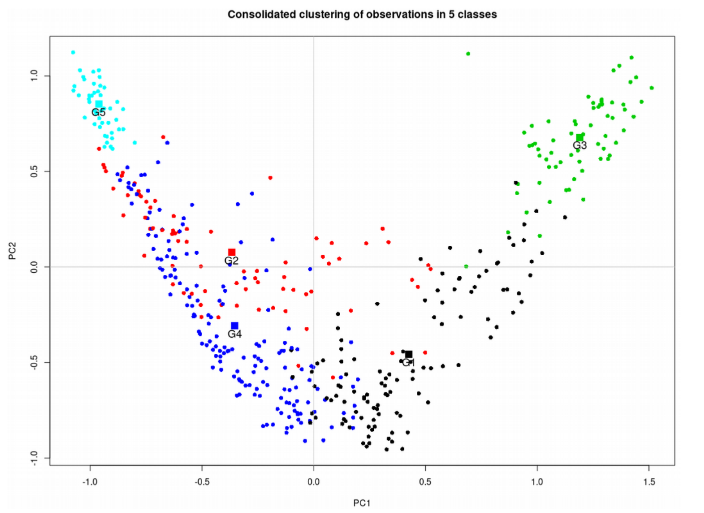

Fig. 7: Color-coded clustering of observations with centroids G1 to G5, according to HC, after k-means consolidation.

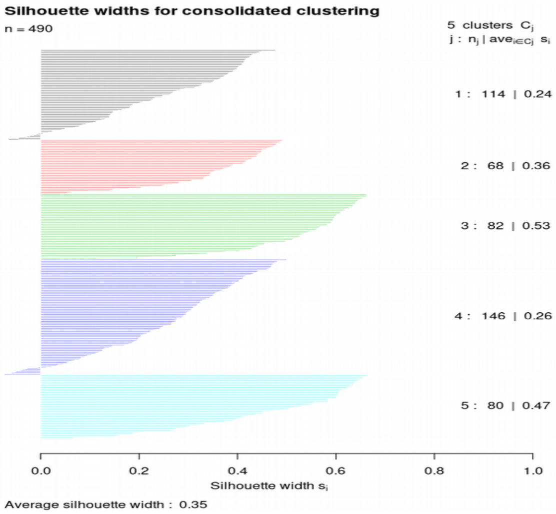

Fig. 8: Silhouette graph of the consolidated clustering. A few points in clusters 1 and 4 are misclassified.

6. Using the function catdes() interpret and name the obtained clusters and represent them in the first factorial display.

Cluster are represented in Figure 7. Based on the clustering information previously obtained, we test each active categorical variable for independence from others (chi-square test of Table 2).

For each variable we compute the v-test statistics based on a 0.01 significance threshold. The null hypothesis, H_0, for categorical variables, is that the variable's mean in any given group is equal to the variable's mean for the entire data set.. Tables 3a and 3b summarize results for 6 most significant p.values for clusters 1 and 2 respectively. Values of v.test stand for the number of std-dev below or above the overall mean values for the sample data. Values of of |v.test| >2 are significant and mean that the category mean is distant form the global mean by at least 2 standard deviations. Tables 3 include the 6 most significant categories for clusters 1 and 2. Doing the same for Cluster 3 to 5 affords us a glimpse into the semantics of clustering:

Cluster 1: Higher end, well motorized vehicles with an elevated cost per km.
Cluster 2: Diesel engine vehicles with more modest performances compare to Cluster 1 and a lower cost per km.
Cluster 3: Top end vehicles with high powered engine, high cost per km, elevated gas consumption, high top velocity.
Cluster 4: Mid range, more affordable (cost per km + gas consumption) vehicles, with average engine size and performance.
Cluster 5: Entry level, lighter-weight and samller size vehicles with small engines and accordingly modest performances.

Chi-square independence	p.value	DF
coste.Km	6.3e-217	16
potencia	3.0e-194	16
consumo	4.0e-167	16
cilindrada	2.2e-147	16
velocidad	5.1e-139	16
peso	2.3e-118	16

***Table 2**: Top 6 categorical variables sorted by p.values for the chi-square independence test.*

Cluster1 catgories	p-value	v.test
coste.Km = Cost_(15,17.5]	2.64e-58	16.2
Potencia = Pot_(130,180]	4.73e-56	15.8
Consumo = Cons_(9.5,11.3]	1.13e-48	14.7
Velocidad = Vel_(200,220]	6.42e-29	11.2
Peso = Pes_(1.2e+03,1.4e+03]	1.19e-16	8.3
poca_aceleracion = Acel_(8.3,9.7]	9.55e-16	8.0

Cluster2 categories	p-value	v.test
Combustible = Diesel	3.95e-65	17.0
Revoluciones = Rev_(3.8e+03,5e+03]	4.82e-50	14.9
coste.Km = Cost_(6.5,11.5]	9.41e-48	14.5
Consumo = Cons_(4.5,7.6]	1.96e-17	8.5
poca_aceleracion = Acel_(13.5,25]	5.31e-13	7.2
Velocidad = Vel_(110,170]	6.21e-07	5.0

***Tables 3**: v.test and p-values for top most categories for (a) Cluster 1, (b) Cluster 2, ….*