# Homework 4

Henry Qiu, Goktug Cengiz, Mehmet Fatih Cagil

May 2019

# 1 Read the PCA_quetaltecaen data

Data is imported as a data-frame object, named *qData*. It is treated as contingency table where columns are qualities of "cultural identity" and rows are qualities of "geographical" categorical variables.

We used Pearson chi-square test to test whether a significant association exists between row and column qualities. Degrees of freedom is equal to 49 ((8-1)*(8-1)). The null hypothesis says that there is no significant association exists between row and column qualities. Obtained result is 7.2584 for X-squared statistics, which is not in the interval of [31.6, 70.2] at 5% risk level. This result leads us to reject the null hypothesis; we can say that there is a significant association of row and column variables.

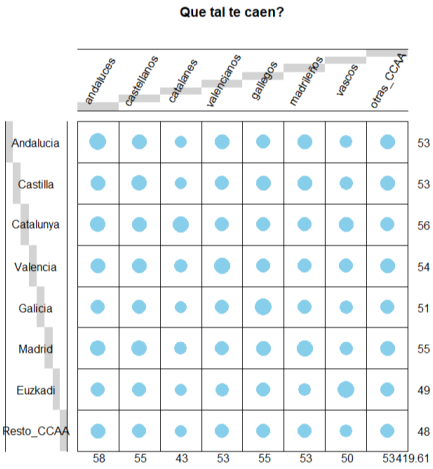
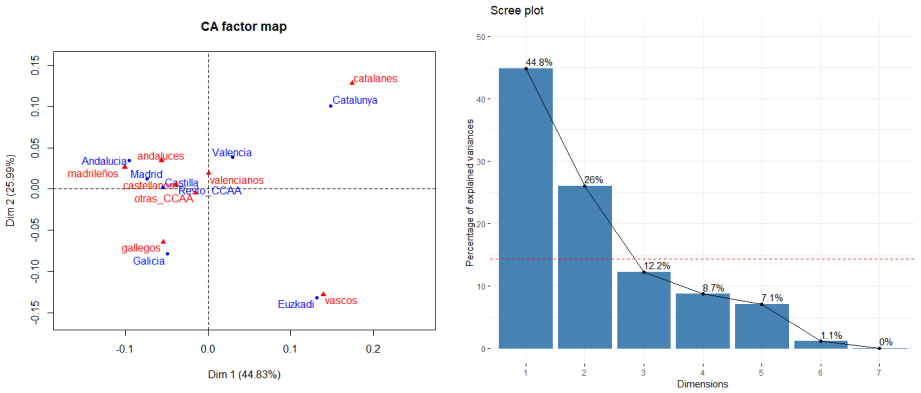



Figure 1: Balloon Plot of the data

Balloon plot in *Figure 1* illustrates the importance of cells by scaling the blue point size with each cell's count number and showing each cell's X-squared residuals remain small. Residuals are widely positive on diagonal, therefore we can anticipate that diagonal may be overloaded.

# 2 Perform a CA of data

We did perform the CA using FactoMiner package. We see that number of significant dimensions is 3. Cumulative variance percentage of first 3 dimensions is 83% which is greater than our threshold, 80.




(a) CA Factor Map. (b) Scree plot.

Figure 2: Figures for original data.

When we investigate the scree plot, first three dimension explains the 83%(44.8+26+12.2) of the variation, which is higher than 80%. Redline represents the average eigenvalue. At the factor map, which is the factor map of first factorial plane, row and column profiles are printed together as blue and red respectively. Distances between same color points occurs with respect to X-squared values. We observe that Catalunya, Basque Country and Galicia on one hand and their populations on the other hand are separated form the rest of the regions and populations. Basques and Catalans are most distant from one another along PC2, conflicting with their proximity along PC1.Also, Galicians are away from the rest. Galicia is a region that is characterized by a strong cultural identity and a language; other regions share castillan as their *prima lingua*.

# 3 Contribution of each cell to the total inertia, Percentage of inertia due to the diagonal cells

| | andaluces | castellanos | catalanes | valencianos | gallegos | madrileños | vascos | otras_CCAA |
|---|---|---|---|---|---|---|---|---|
| Andalucia | 1.303498e-03 | 4.739381e-06 | 3.659717e-04 | 3.084034e-08 | 3.368955e-05 | 4.433239e-05 | 4.663950e-04 | 4.361343e-06 |
| Castilla | 1.550959e-06 | 2.516498e-04 | 1.580075e-04 | 4.374119e-06 | 8.937748e-07 | 3.411794e-05 | 1.430193e-04 | 1.139230e-05 |
| Catalunya | 1.590274e-05 | 8.625272e-05 | 3.846138e-03 | 7.931462e-05 | 2.308695e-04 | 2.581110e-04 | 1.564309e-05 | 3.095999e-05 |
| Valencia | 1.656530e-04 | 7.303155e-06 | 1.012375e-04 | 9.868932e-04 | 7.539885e-05 | 4.565323e-05 | 5.430294e-05 | 2.403469e-06 |
| Galicia | 9.871097e-05 | 4.394589e-05 | 2.970143e-04 | 2.715242e-05 | 1.859812e-03 | 3.756521e-06 | 2.327578e-06 | 7.829120e-06 |
| Madrid | 8.210152e-06 | 8.902189e-05 | 1.520766e-04 | 1.483750e-04 | 1.710377e-05 | 9.646440e-04 | 1.392706e-04 | 1.782190e-06 |
| Euzkadi | 1.120936e-04 | 4.212389e-05 | 1.602183e-05 | 3.308056e-05 | 7.111767e-05 | 5.204745e-04 | 3.611128e-03 | 3.542396e-08 |
| Resto_CCAA | 2.905985e-05 | 2.782769e-06 | 8.099904e-05 | 8.175150e-06 | 4.822098e-07 | 9.236540e-06 | 6.037135e-05 | 9.757530e-06 |

Figure 3: Total Inertia table

Total inertia is **0.01729803** and diagonal's contribution is **74.19%**. Gutmann effect, which is an overloaded diagonal effect, exists.

# 4 Nullification of overloaded diagonal elements in terms of inertia

We used a convergence criterion of $\varepsilon = 1e^{-4}$. Algorithm is explained below.

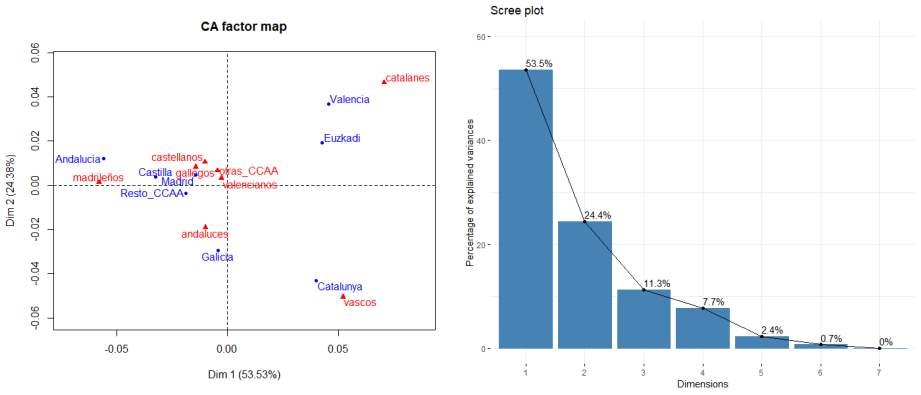
- Initialize total count N, relative frequencies tables $f_{i,j}$ and inertia matrix
- Calculate row weights $f_{i.}$ and column weights $f_{.j}$
- Impute new diagonal at the count table
- Recompute new total count, inertia matrix and total inertia
- If absolute difference between *new inertia* and *old inertia* $>= \varepsilon$, go to 2nd step

| | andaluces | castellanos | catalanes | valencianos | gallegos | madrileños | vascos | otras_CCAA |
|---|---|---|---|---|---|---|---|---|
| Andalucia | 1.795704e-08 | 4.568475e-08 | 9.469690e-05 | 2.536410e-05 | 1.017592e-06 | 1.475727e-04 | 1.726201e-04 | 5.922702e-06 |
| Castilla | 5.946106e-07 | 6.579982e-08 | 3.518200e-05 | 5.603027e-07 | 1.746085e-05 | 5.467831e-05 | 4.100794e-05 | 1.535041e-08 |
| Catalunya | 2.639605e-05 | 1.669288e-05 | 1.653618e-07 | 1.013686e-06 | 3.312910e-05 | 7.329674e-05 | 2.958001e-04 | 5.932073e-06 |
| Valencia | 7.215353e-05 | 2.212422e-06 | 4.196512e-04 | 9.052688e-10 | 9.544417e-06 | 8.452728e-06 | 9.379084e-07 | 5.999876e-06 |
| Galicia | 1.039368e-05 | 1.031151e-05 | 4.773939e-05 | 7.840742e-07 | 4.972199e-08 | 1.850870e-05 | 8.356879e-05 | 1.854005e-06 |
| Madrid | 3.348218e-06 | 1.176160e-04 | 1.110114e-05 | 7.391455e-05 | 2.566611e-06 | 7.398639e-10 | 1.388856e-05 | 1.247870e-07 |
| Euzkadi | 2.084750e-06 | 3.349735e-07 | 8.835364e-05 | 7.483029e-06 | 2.984638e-06 | 2.346470e-04 | 1.499002e-07 | 1.796651e-05 |
| Resto_CCAA | 2.768955e-05 | 3.304666e-06 | 1.876364e-05 | 4.598647e-06 | 3.933453e-08 | 5.458740e-06 | 1.832184e-05 | 1.808862e-07 |

Figure 4: New Inertia table

New total inertia is 0.0023963 and diagonal cells account for 0.03%

# 5 New CA with modified diagonal




(a) CA Factor Map.

(b) Scree plot.

Figure 5: Figures for original data.

Factor map summarizes the result of the new CA with Gutmann effect corrected. First 3 dimensions represents almost 90% of the inertia while first 2 represents 78%. The first factorial plane represents inertia better(78% now, 71% before).

| Sum of Squared Cosines | | |
|---|---|---|
| Name | Before | After |
| Andalucia | 0.5793340 | 0.9264637 |
| Castilla | 0.6289046 | 0.9292892 |
| Catalunya | 0.9351664 | 0.9810883 |
| Valencia | 0.2113604 | 0.8509638 |
| Galicia | 0.4451455 | 0.6138461 |
| Madrid | 0.4876898 | 0.1330491 |
| Euzkadi | 0.9216689 | 0.6973444 |
| Resto_CCAA | 0.8224603 | 0.5675070 |

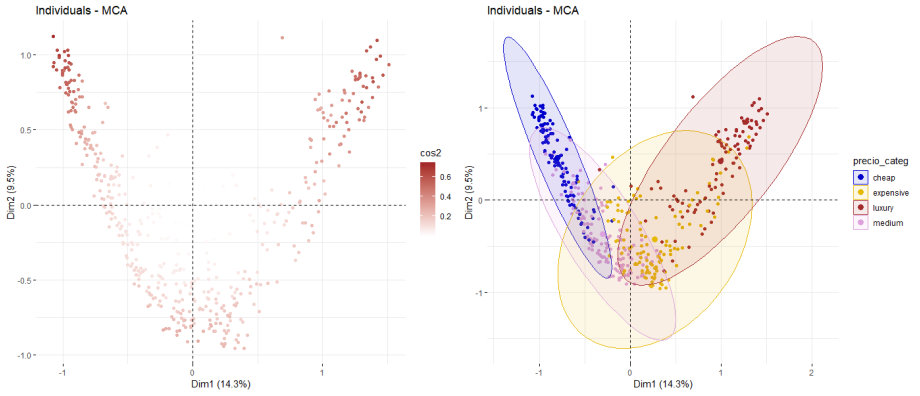
The tables shows the sum of squared cosines for the row profiles in the first factorial plane before and after the modification at diagonal.

# 6 Read the mca_car

The data contains 490 instances without any missing values. Each instance contains 19 attributes and 18 of them are categorical or ordinal and 1 is continuous, *precio*. We considered *precio* and *precio-categ* as response variables and *marca* as descriptive variable. All 3 will be treated as supplementary.

# 7 Multiple Correspondence Analysis

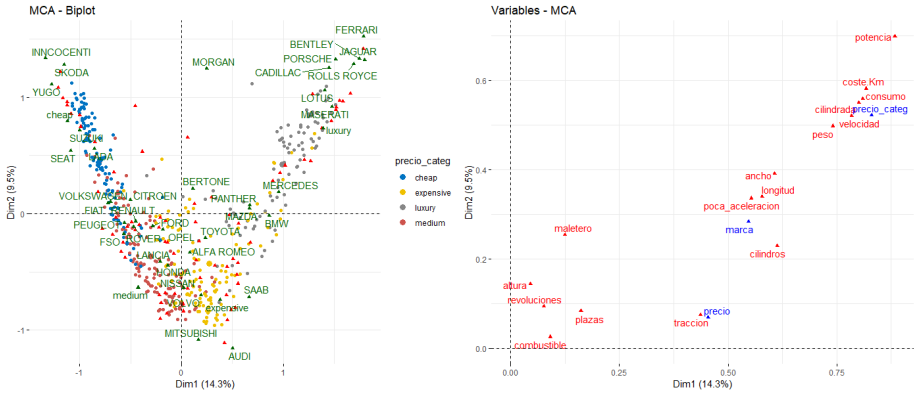
We treated *precio* as supplementary continuous variable, *marca* and *precio-categ* as supplementary categorical variables. They will not be included at the MCA and their coordinates will be predicted. All other variables are active variables. *Figure 6* and *Figure 7* gives information about the result.




(a) Scatterplot of individuals. Dark red means better representation than lighter.

(b) Scree plot.

Figure 6: MCA Figures




(a) Biplot of individuals

(b) Scatterplot, active variables are in red, supplementary are in blue

Figure 7: MCA Figures

## 8 Interpretation of the first factorial plane

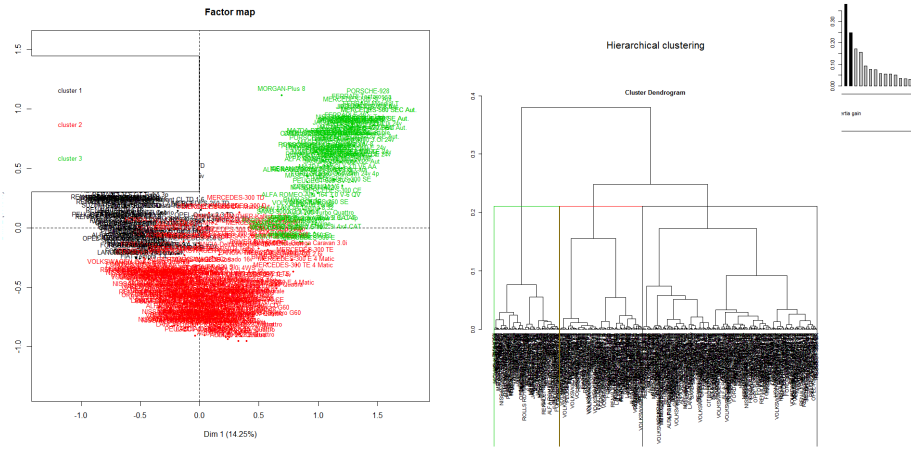
The first factorial plane resprents about 24% of the variability in the dataset. We can see the Guttman effect on *Figure 6* and *Figure 7(a)*. They signal a diagonal overloading. According to the *Figure 6*, entry level, cheap vehicles are heavily in the 2nd quadrant while medium to expensive vehicles are at 3rd and 4th quadrant; high end cars are at the 1st quadrant. *Figure 7(a)* demonstrates how modalities of the categorical variables *marca* and *precio_categ* are pseudo-barycenters of the individuals. In *Figure 7(b)*, distance of variable points form the origin is a measure of factor quality.Points further from the origin denotes variables which clarify for more inertia and are better represented in that plane than variables closer to the origin.

## 9 Number of significant dimensions

To obtain the significant dimensions we first obtain the eigenvalues calculated when we compute the MCA. Withe these eigenvalues we do the difference with the mean of these eigenvalues, only those which are greater than the mean are significant. We can see that the number of dimensions is very high, there are 18 dimensions over the mean. We can reduce this retaining only the dimensions that contributes more information. Over these dimensions, as it is shown in the code, we compute the percentage of eigenvalues over the total of the eigenvalues of these dimensions, and we finally select the dimensions that contributes in total more than 80%. We finally obtain the conclusion that the number of significant dimensions will be 11, where 80.79740% of information significant information is explained.

## 10 Perform a hierarchical clustering

Once we have the significant dimension we perform the hierarchical clustering. But before we do this, we compute again the MCA with the number of significant dimensions obtained. Performing the hierarchical clustering will give us as result the individuals classified in different clusters. In the following charts we can see the number of clusters obtained are 3.




(a) Plot of individuals marked as different clusters.

(b) Dendogram obtained after performing hierarchical clustering.

Figure 8: Hierarchical clustering

# 11 Cluster interpretation and representation in the first factorial plane

With the catdes function we can see the relations between the cluster variable and the categorical variables. In the results we can observe that the variables that most influence in the decision of clustering an individual are power, the cost per kilometer, and the consumption of the car.

| Link between the cluster variable and the categorical variables (chi-square test) | | |
|---|---|---|
| | p.value | df |
| potencia | 6.787323e-149 | 8 |
| coste.Km | 9.876154e-144 | 8 |
| consumo | 5.276185e-129 | 8 |
| velocidad | 5.998802e-121 | 8 |
| cilindrada | 1.236725e-102 | 8 |

In the tables below, we can see the variables that much determined each cluster. Thanks to this, we have more information about what describes more the cluster. In the case of the first cluster, it is mostly determined by the variable speed. In the case of the second cluster, it is mostly determined by the variable powet. In the case of the third cluster, it is mostly determined by the variable cost per kilometer. We can conclude that the first cluster corresponds to the fastest cars, the second cluster corresponds to the cars that costs most per kilometer.

| Description of each cluster by the categories | | | | | |
|---|---|---|---|---|---|
| Cluster1 | Cla/Mod | Mod/Cla | Global | p.value | v.test |
| velocidad=Vel$_{(}$110, 170] | 96.4601770 | 63.3720930 | 23.061224 | 8.946233e-58 | 16.022177 |
| potencia=Pot$_{(}$35, 75] | 100.0000000 | 55.8139535 | 19.591837 | 1.343544e-54 | 15.560826 |
| consumo=Cons$_{(}$4.5, 7.6] | 96.1538462 | 58.1395349 | 21.224490 | 3.594854e-51 | 15.047324 |
| peso=Pes$_{(}$640, 940] | 94.1176471 | 55.8139535 | 20.816327 | 5.105935e-46 | 14.240908 |
| cilindrada=Cil$_{(}$900, 1.5$e$+03] | 97.7011494 | 49.4186047 | 17.755102 | 1.176334e-43 | 13.855622 |

| Description of each cluster by the categories | | | | | |
|---|---|---|---|---|---|
| Cluster2 | Cla/Mod | Mod/Cla | Global | p.value | v.test |
| potencia=Pot$_{(}$105, 130] | 94.949495 | 42.922374 | 20.204082 | 1.140604e-32 | 11.903079 |
| coste.Km=Cost$_{(}$15, 17.5] | 89.622642 | 43.378995 | 21.632653 | 1.486702e-27 | 10.876813 |
| coste.Km=Cost$_{(}$13.5, 15] | 92.857143 | 35.616438 | 17.142857 | 1.822626e-24 | 10.208174 |
| velocidad=Vel$_{(}$185, 200] | 91.011236 | 36.986301 | 18.163265 | 5.770465e-24 | 10.095722 |
| consumo=Cons$_{(}$9.5, 11.3] | 86.274510 | 40.182648 | 20.816327 | 2.305308e-22 | 9.727345 |

| Description of each cluster by the categories | | | | | |
|---|---|---|---|---|---|
| Cluster3 | Cla/Mod | Mod/Cla | Global | p.value | v.test |
| coste.Km=Cost$_{(}$17.5, 30] | 96.7032967 | 88.888889 | 18.571429 | 1.975597e-80 | 18.992280 |
| consumo=Cons$_{(}$11.3, 20] | 92.3913043 | 85.858586 | 18.775510 | 3.662808e-71 | 17.836768 |
| potencia=Pot$_{(}$180, 500] | 93.2584270 | 83.838384 | 18.163265 | 1.682847e-69 | 17.621565 |
| cilindrada=Cil$_{(}$2.6$e$+03, 8$e$+03] | 85.1063830 | 80.808081 | 19.183673 | 3.895504e-58 | 16.073786 |
| velocidad=Vel$_{(}$220, 350] | 87.8048780 | 72.727273 | 16.734694 | 4.663623e-52 | 15.181861 |

Finally, we can see the representation of individuals in the first factorial plane. Where the red individuals corresponds to the cluster 1, the green individuals to the cluster 2, and the blue individuals to the cluster 3.

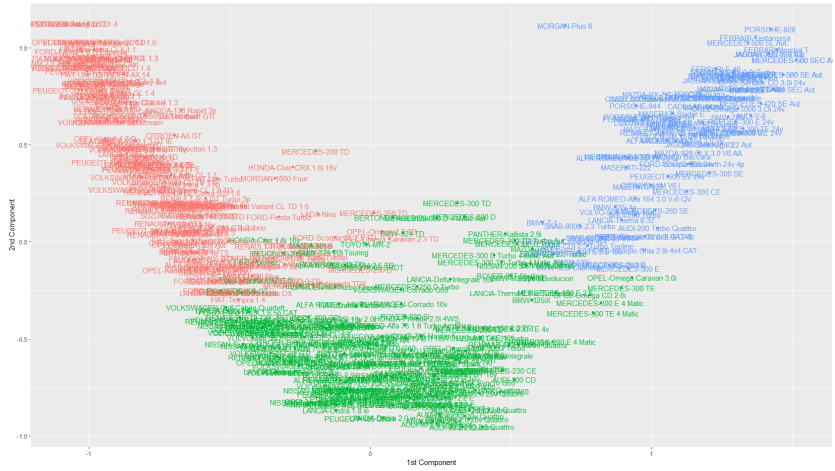



Figure 9: Representation of individuals in the first factorial plane.