MVA - Project 5

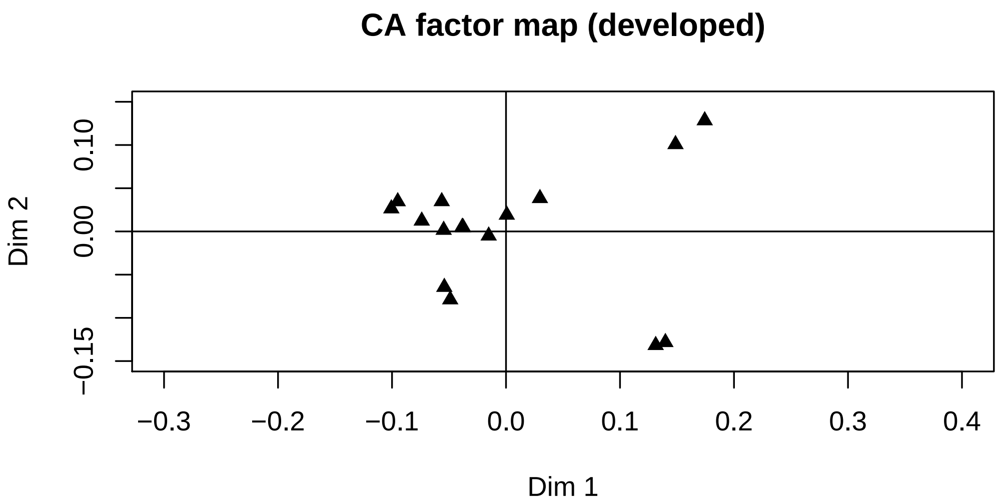
Carles Garriga Estrade i Balbina Virgili Rocosa

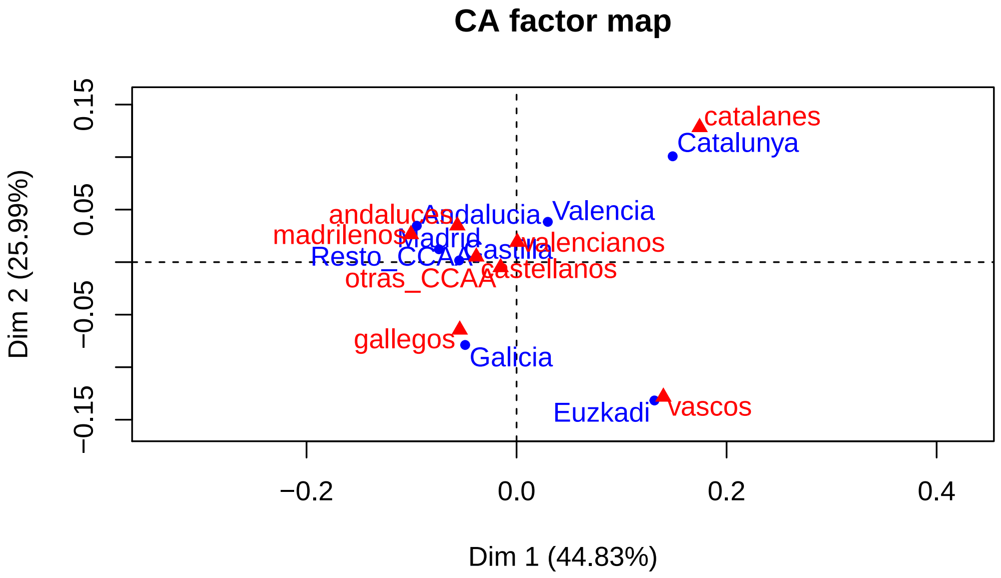
05/04/2018

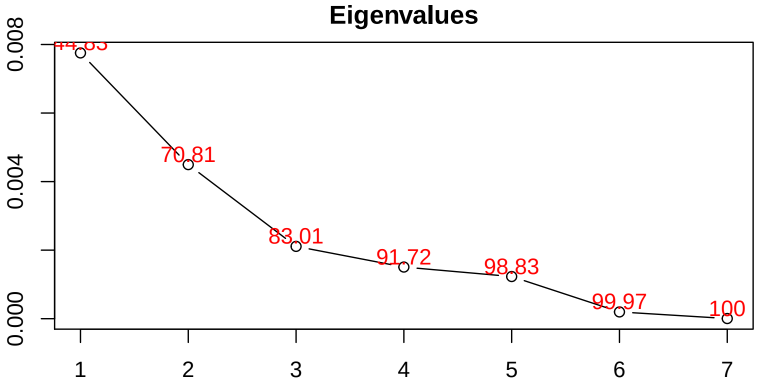
All the code created for us to develop this assignment can be found on LAB05.R file. This file is located in the same folder of this document.

Exercise 1 and 2 First of all, the PCA_quetaltecaen dataset has been read from the file. In order to be able to manipulate it, the first column with CCAA names has been removed and added as row names. This way, we obtain the contingency table and all conditions to apply CA are fulfilled with the given dataset (all cells have positive numbers, make sense to compare rows from row-profiles and...).

Then, we have performed the CA of the given dataset. For this second exercise, two approaches have been developed. Both, a manual implementation of CA as well as the CA function of 'FactoMineR' library are used in order to confirm that both provide the same correct results. The first factorial plain and the eigenvalues calculated are plotted below.







On the one hand, as specified during previous homeworks, there are several methods that can be used to define the number of significant dimensions. We have decided, once again, to determine those by taking the number of factors that explain up to a 90 percent of the variance. Therefore, the significant dimensions are the first four dimensions extracted in the CA.

On the other hand, deeply observing the first factorial plane, we need to be careful because we need to interpret the CA as a double PCA, one PCA of the row-profiles and another PCA for the column-profiles. We need to keep in mind that we cannot interpret the distance between row and column points directly and they need to be interpreted as pseudo-baricenters of the other set. That is why, we are able to interpret that Andalucia, Madrid, Castilla, Valencia and Resto_CCAA have given similar punctuation to other communities, while Galicia, Euzkadi and Catalunya have given different ones to any other community. Furthermore, similar results are obtained for the punctuation received on each community.

Exercise 3 Now, we have computed the contribution of each cell to the total inertia. After computing the matrix of relative frequencies, we can compute the total inertia of the cloud of points by using the formula $(f_{ij} - fi. * f.j)^2/(fi. * f.j)$

```
## [1] 74.19063
```

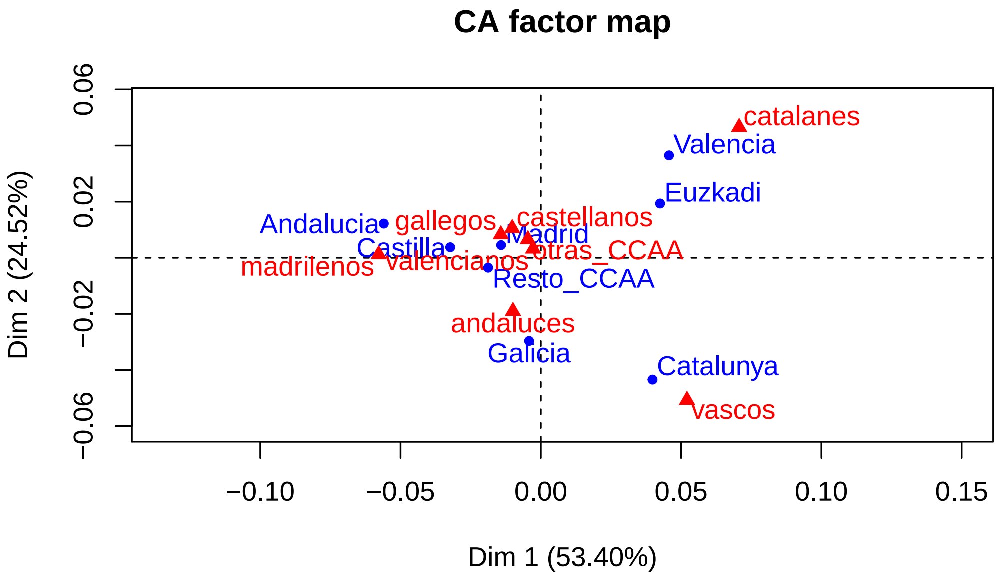
In order to compute the percentage of inertia due to the diagonal cells, we have computed the percentage based on the previous contribution to the inertia for each cell and we have calculated the sum of the diagonal. As we can observe with the result obtained, we realize that the values of the diagonal have a huge contribution and nearly only the ~25% of the inertia is given by the cells that are not in the diagonal. Therefore, we know that values provided on the diagonal are based on marks that each communities give to itself and we can realize that they are considerably higher than the rest.

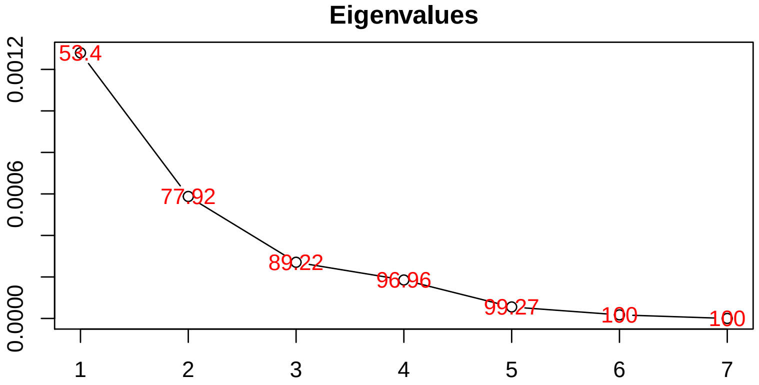
Exercise 4 and 5

So, we need to recompute new values for the diagonal in order to obtain values that do not influence the results obtained. To do it, we try to reduce this influence by imputing the diagonal values by the independence hypothesis values of the product of marginal probabilities, using the formula $(n * fi. * f.j)$.

```
## [1] 7.893078e-23
```

With the new values obtained for the diagonal, the percentage of inertia of the values of the diagonal is almost 0. So, they won't influence on the results obtained any more.





After computing again the CA with the new diagonal values calculated, we can see that the significant dimensions have not changed and they are still the first four ones and the accumulative variance of the results obtained on the first factorial plane have increased up to the 77%.

Regarding the first factorial plane results, we can see that the results retrieved have changed. In other words, now gallegos, castellanos, valencianos and otras_CCAA seem to receive similar punctuation from the other communities, but madrilenos and andaluces have been slightly separated from this group of communities. Catalanes and vascos are still the ones that receive more extreme punctuation. Moreover, we can see that Madrid, Castilla and Resto_CCAA give similar punctuation to other communities, while Catalunya is the most different one. We can also realize that Euzkadi and Valencia give more similar punctuation now than before.

To conclude, we can say that it is important to detect non-truthful or non-important information of a dataset and try to minimize the impact of it to the final results of the CA, otherwise, the results obtained on the analysis may not be the correct and real ones.