MVA - Project 4

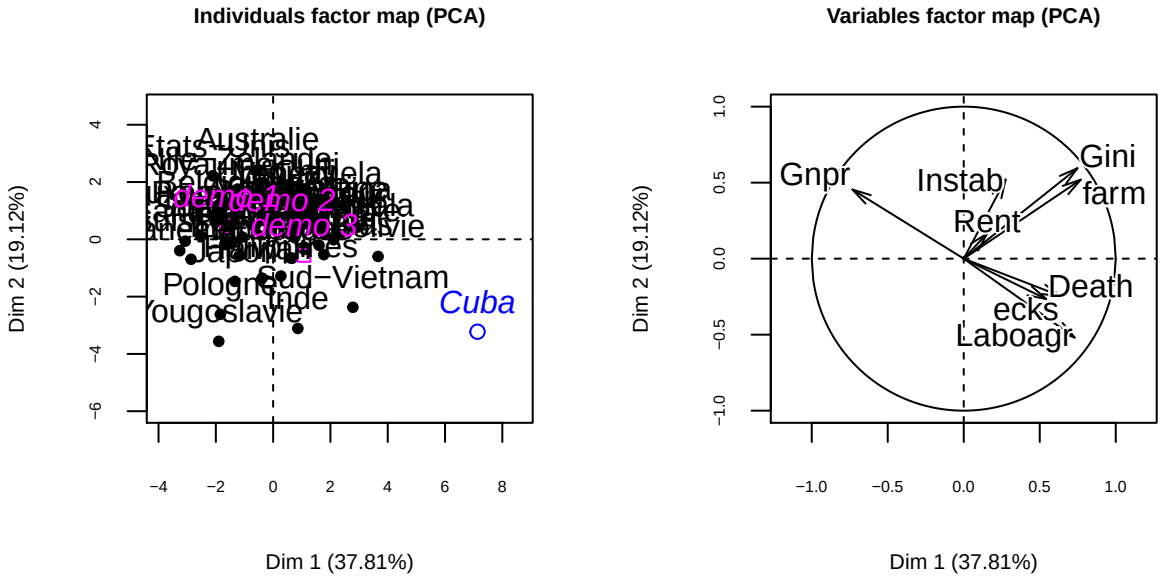
Carles Garriga Estrade i Balbina Virgili Rocosa

04/02/2018

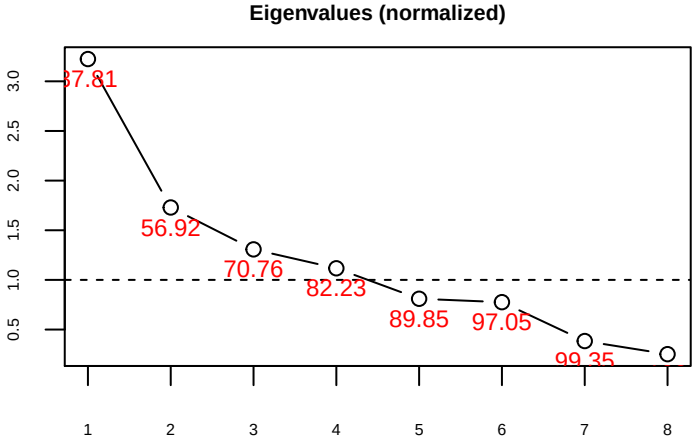
All the code created for us to develop this assignment can be found on LAB04.R file. This file is located in the same folder of this document.

Exercise 1, 2 and 3.

First of all, the Russet dataset has been read from the file and missing values have been filled with KNN method. Then, we have performed the PCA of the continuous variables using the FactoMineR library. To do it, we have defined the qualitative 'demo' variable as illustrative, by setting the parameter 'quali.sup' to the related index of the variable, and 'Cuba' individual as supplementary, by setting the attribute 'ind.sup' with the index of the outlier. The 'scale.unit' parameter has been set to 'TRUE' in order to execute the PCA using normalized euclidean distance. The results obtained are showed below and it can be seen that are the same ones that we obtained during the third homework.



Now, in order to be able to select the signigicant dimensions, we have plotted below together the eigenvalues for each dimension as well as the accumulative percentage of inertia up to each dimension.



As we discussed during the second homework, there are several methods that can be used to define the number of significant dimensions. We have decided, once again, to compute those taking into account the fixed percentage of inertia (>90). Therefore, the significant dimensions are the five first ones for the results obtained.

Exercise 4

Exercise 5

After the clusters are defined, we need to interpret and represent them in the first factorial display using the function 'catdes' of the FactoMineR library. To do it, we need to pass to the function the defined clusters with the dataset used to define them. The number of response variables is set to 1. The most significant results obtained for each cluster are showed below.

Cluster 1	farm	Gini	Laboagr	Instab	Gnpr
v.test	4.496867	4.210871	3.311864	3.156425	-3.624185
p.test	6.896198e-06	2.543884e-05	9.267655e-04	1.597162e-03	2.898740e-04

Cluster 2	Gnpr	Rent	Laboagr
v.test	4.146770	2.487275	-3.959736
p.test	3.371991e-05	1.287260e-02	7.503270e-05

Cluster 4	Gnpr	ecks	Rent	Gini	farm
v.test	2.070405	-2.477637	-3.291283	-4.543160	-5.048875
p.test	3.841447e-02	1.322556e-02	9.973150e-04	5.541724e-06	4.444186e-07

Cluster 3	Death	ecks	**Cluster 5**	Rent	Instab
v.test	5.838112	2.126104	v.test	2.627079	-4.592467
p.test	5.279572e-09	3.349457e-02	p.test	8.612130e-03	4.380374e-06

With catdes function we want to realize which are the most significant continuous variables on each clustered defined. This way, we are able to find out the profiling, which is the significant characteristics which make the group of individuals different than the other whole set of individuals. For it, we first have a look on p.test calculated value. The p.test is less than 5% when one category is significantly linked to another categories, so we realize that all variables that catdes return, are already below this stablized threshold. For this reason, we need to also have a look to v.test value, which is the one that gives us more information.

After ordering them by v.test value, we then state that if the v.test is positive, it means that the category is over-expressed for the category and if the v-test is negative it means that the category is under-expressed for the category. So, with the results retrieved we can see that the first cluster is best expressed by farm and Gini, but it is also expressed by Laboagr and Instab but it is under-expressed by Gnpr. On the other hand, the second cluster is best expressed by Gnpr, also expressed by Rent, and under-expressed by Laboagr. The third one is well-expressed by Death and ecks. The forth one is over-expressed by Gnpr but under-expressed by ecks, Rent, Gini and farm. Finally, the fifth cluster is over-expressed by Rent but under-expressed by Instab. To conclude, we can see that all clusters are expressed with different variables so we are able to say that we have defined clusters that have significant different characteristics to the other ones.

Exercise 6

Finally, with the results developed during the previous exercises, we would like to assign Cuba to one of the

defined clusters. To do it, we are going to calculate the distance between each centroid of the defined clusters and the new individual Cuba that we want to assign. That is why the euclidean distance of the five principal dimensions defined from Cuba to each cluster is calculated. As we want Cuba to be on the same plane as the clusters' centroids, otherwise we are not able to correctly compare them, we extract the Cuba's coordinates, which are in the first fatorial plane, from the PCA response. The results obtained from the calculations developed are showed below.

```
##   result   result   result   result   result
## 15.19420 15.88672 11.17291 16.26204 15.44225
```

There are two different methods to define which cluster is more accurate for Cuba, crisp or soft. As we have been able to calculate the euclidean distance from each cluster centroid to the individuals, we are going to use crisp for this exercise. As we can see with the results retrieved, the smallest distance calculeted is obtained from the third cluster. So, we should assign cuba to the third cluster.