MVA - Project 2

Carles Garriga Estrade i Balbina Virgili Rocosa

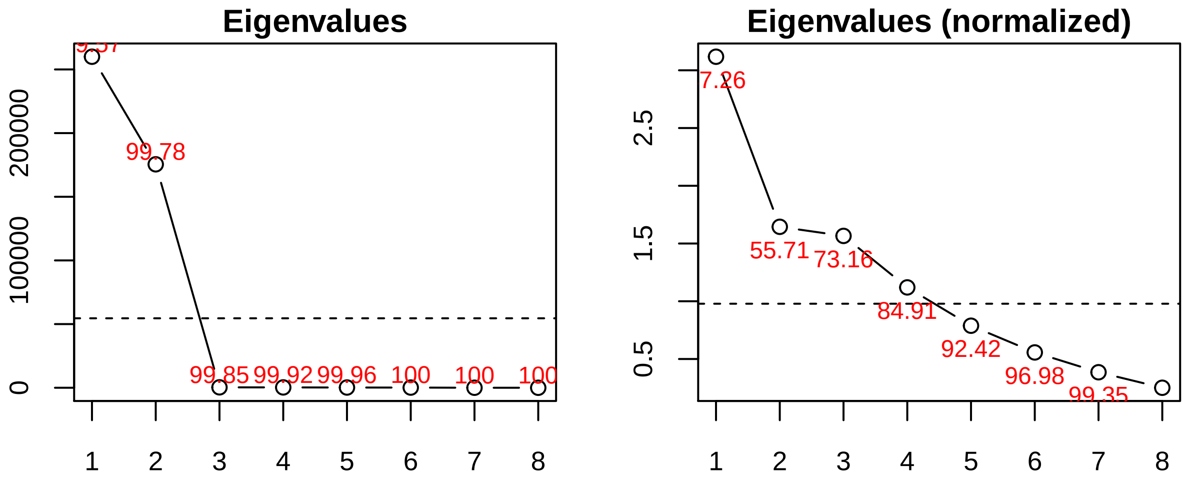
3/13/2018

All the code created for us to develop this assignment can be found on LAB02.R file. This file can be found in the same folder of this document.

Exercise 1. First of all, the Russet completed dataset has been read and the completed matrix has been defined with the continuous data. To do it, all data of the dataset except the one related with 'demo' variables has been included to the previous matrix.

Exercise 2. We have developed our own implementation of a PCA analysis function, which can be found on LAB02.R file. To be able to execute the function, the X matrix, the weights (as a vector) and the distance are required to be given. The distance can be 'normalized' or 'centered', depending on the metric that wants to be used ('normalized' for normalized euclidean distance or 'centered' for euclidean distance). The results are explained below.

The plots below represent the eigenvalues for each dimension as well as the accumulative percentage of inertia up to each dimension.

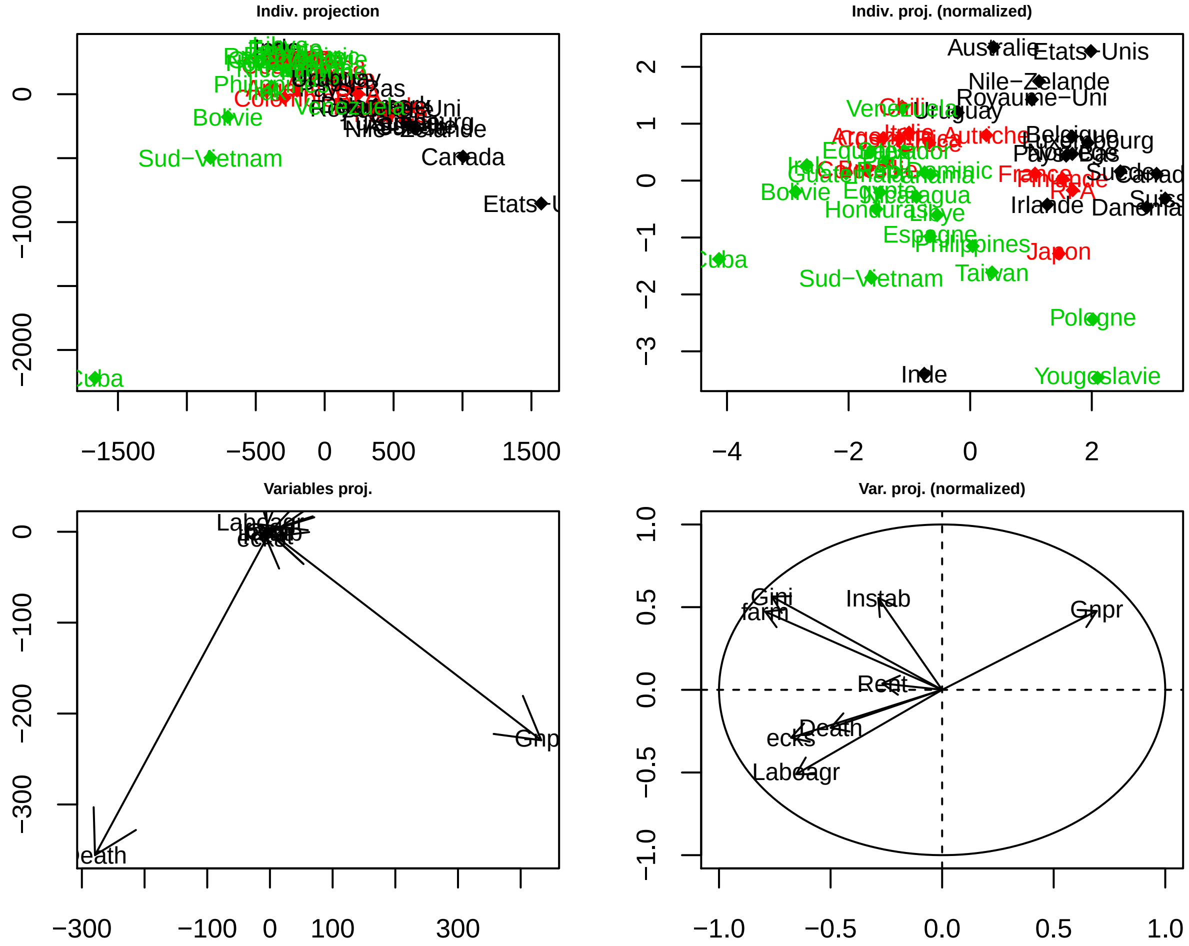


In order to define the number of significant dimensions, several methods can be used:

- Kaiser rule: Performing the mean average of the eigenvalues and taking as much dimensions by comparing their eigen values with the average.
    - Using the centered euclidean distance, the two first dimensions would be the significant ones.
    - Using the normalized one, the average of the eigenvalues is always 1. Under this assumption, The first four dimensions should be extracted.
- Last elbow rule:
    - Using the centered euclidean distance, the two first dimensions would be the significant ones.
    - Using the normalized one, there's no "relevant' plateau on the last dimensions to consider a last elbow. Therefore, we should take all the dimensions up to a fixed percentage of inertia (using the method below).
- Taking all factorial coordinates up to a fixed percentage of inertia (~90%):
    - Using the centered euclidean distance, the two first components would be the significant ones.
    - Using the normalized one, the five (92.42%) first dimensions should be necessary.
- Performing an statistical test of significance of the eigenvalues (not used).

Having introduced all the methods, since it was not easy to idenitify a clear last elbow for the computation of the significant dimensions, we've decided to compute those taking into account the fixed percentage of inertia. Therefore, the significant dimensions are the five first ones for the normalized euclidean distance and the two first ones for the centered.

Once significant dimensions are decided, the projections of individuals and variables are calculated. For simplification, we have just plot them on the first factorial plane.

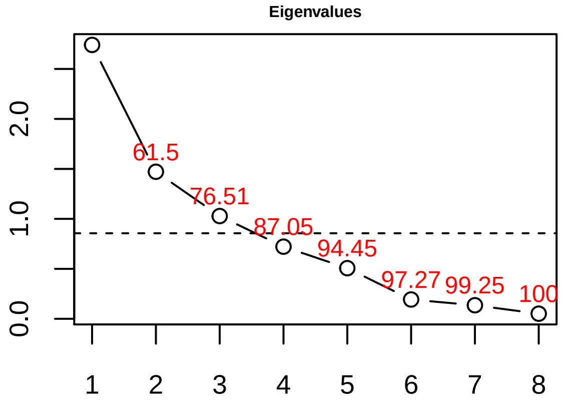


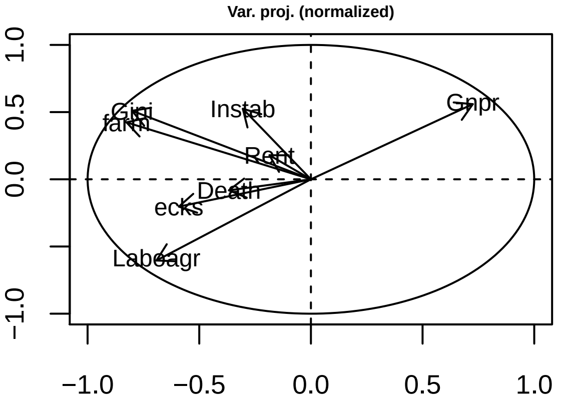
Taking a look at the correlation of the variables with the significant principal components, we can identify *Gini*, *farm*, *Gnpr*, *Laboagr* and *ecks* as the most correlated ones in the first factorial plane. In the third principal component, the most correlated variables are *Rent* as well as *Death*. For the fourth and fifth component, both *instab* and *rent* are the most correlated ones.

On the one hand, executing the PCA using the euclidean distance instead of the normalized one, the resulting plot of variables shows a big difference between *Death* and *Gnpr* to the rest of them, due to the value of its covariance, making impossible to distinguish any other possible relationship between the rest of the variables in the first factorial plane. Also, the plot of individuals shows "Cuba" being clearly farther away from the rest of individuals (contribution the most the PCA).

On the other hand, executing the PCA using the normalized euclidean, the relationship between all the individuals and variables of the dataset end up in a way more "clear" results. Therefore, the metric M that should be used in the previous pca must be the inverse of the variance $M = S^{-2}$.

Exercise 3. We have executed our developed PCA function for the exercise 2 but changing the weight of Cuba individual to 0. To be able to do it, the vector of weights needs to be recalculated by setting the row of weight 0 and to accomplish that the sum of weights equals 1. The results obtained are showed below but first, we have observed that the *centroid* of each variable has not changed but the relationship between variables and individuals has been affected.



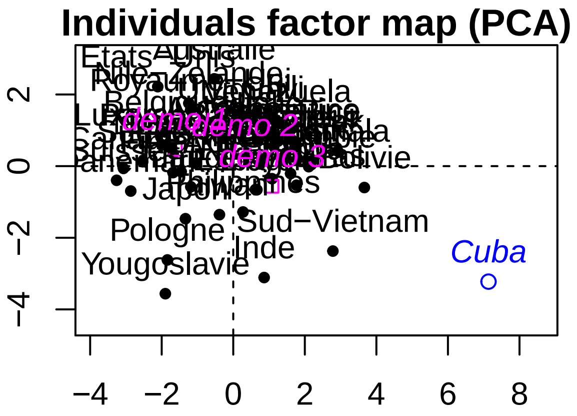


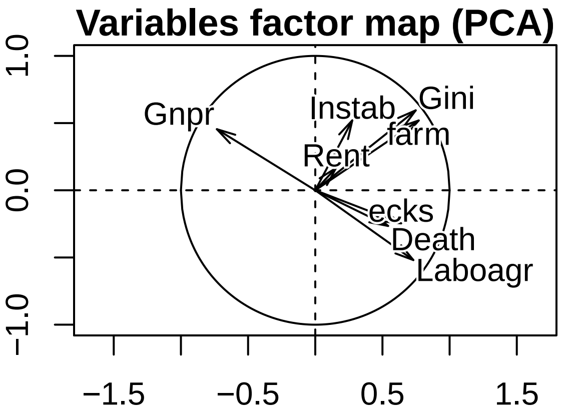
With the above retrieved results, we can see that the accumulative percentage of variablitity has changed, but the number of significant components according to the fixed percentage of inertia has not changed. So, the most significant number of components are still the first five. Furthermore, we can see that the relationship between the variables has been effected because the correlation of *Rent* variable has changed considerably. That is why we can say that Cuba is an outlier that influences the other data.

Exercise 4. Computing the correlation of the principal components of the last exercise (setting "Cuba" as a supplementary individual) with the ones of the second exercise, we find the results showed on the table below. Taking a look at the diagonal of the correlation matrix, we can observe that setting "Cuba" as a supplementary individual doesn't get changed neither first nor the second p. components. However, starting with the third p. component the difference (regarding both executions) has increased.

P.Comp 1	P.Comp 2	P.Comp 3	P.Comp 4	P.Comp 5
0.9914313	0.3933545	-0.0506858	-0.0363321	0.0905388
0.2242108	0.9754859	-0.2730099	0.4367374	0.3553962
0.4931758	0.0785477	**0.6690508**	-0.6105427	-0.5168897
-0.1481258	0.0564144	0.7359841	**0.4748138**	0.2363561
-0.0457279	0.0297233	-0.0429381	0.4136966	**-0.6738181**

Exercise 5. Now, we have performed the PCA using the FactoMineR library. As we decided on exercise 2, the best way to execute the PCA with the given dataset is using normalized euclidean distance. That is why, the 'scale.unit' parameter has been set to 'TRUE'. Also, to be able to use the 'demo' variable as illustrative, we have set the parameter 'quali.sup' to the related index of the variable. At this point, we already know that Cuba is an outlier of the dataset which influences the other data, so we have used this individual as supplementary. To do so, we have set the attribute 'ind.sup' with the index of the outlier. The results obtained are showed below and it can be seen that are really similiar as the obtained on the exercise 3 but with the oposite sign.





Exercise 6. We have identified the best and the worst represented countries by taking the individuals with the maximum and minimum importance of *cos2* in the first factorial plane (which is composed by the first and the second dimensions). With the results obtained, we can confirm that the best represented country is Suisse and the worst represented is Espagne.

Suisse	Dim.1	Dim.2	**Espagne**	Dim.1	Dim.2
	0.94223934	0.01390865		0.03275566	0.03638849

Exercise 7. We have identified the three countries that influence the most the formation of the first principal component by sorting the contribution (*contrib* variable) of each individual on the first dimension and we got the three with higher values. We repeated the same procedure for the second dimension, checking the values obtained for the second measure. With the results obtained, we can confirm that the three countries that most influence the formation of the first principal component are Bolivie, Suisse and Canada. And the countries for the second component are Yougoslavie, Inde and Pologne.

1st Component	**Bolivie**	**Suisse**	**Canada**	2nd Component	**Yougoslavie**	**Inde**	**Pologne**
	9.648733	7.638387	6.766401		18.028899	13.755210	9.751908

Exercise 8. We have identified the best and the worst represented variables by taking the variables with the maximum and minimum importance of *cos2* in the first factorial plane (which is composed by the first and the second dimensions). With the results obtained, we can confirm that the best represented variable is Gini and the worst represented is Rent.

Gini	Dim.1	Dim.2	**Rent**	Dim.1	Dim.2
	0.5592258	0.3537580		0.02357010	0.02727845

Exercise 9. We have identified the three variables that influence the most the formation of the first principal component by sorting the contribution (*contrib* variable) of each variable on the first dimension and we got the first three with higher values. We repeated the same procedure for the second dimension, checking the values obtained for the second measure. With the results obtained, we can confirm that the three variables that most influence the formation of the first principal component are *farm*, *Gini* and *Gnpr*. And the countries for the second component are *Gini*, *Instab* and *farm*.

1st Component	**farm**	**Gini**	**Gnpr**	2nd Component	**Gini**	**Instab**	**farm**
	19.61556	18.49016	17.73829		23.13257	17.70380	17.65349

Exercise 10. By taking a look to the resulting tests vectors of supplementary qualitative data, we can confirm that the first and third modalities of the demo variable are more significant among the second one.

	Dim.1	Dim.2
demo 1	-4.3524010	1.7742851
demo 2	0.7231947	0.9741501
demo 3	3.4988292	-2.5579943