Homework 2: Practice of PCA

1. **Read again the Russet completed data set. Define as X matrix the one defined by the continuous variables. (Now, just using matrix operation)**
2. Write a function for PCA analysis, having as input, the raw data matrix X, and a vector containing the weights per individual, and allowing either the Euclidean metric or the normalized Euclidean metric (M metric of R^p). The function should provide the screeplot of the eigenvalues, then you can choose the number of significant ones, the projection of individuals and variables upon the significant dimensions, the plot of individuals on the first factorial plane and the plot of variables on the first factorial plane.
    a. Define the matrix N of weights of individuals (with uniform weights).
    b. Compute the centroid G of individuals.
    c. Compute the covariance or correlation matrix of X (be aware of dividing by `sum(weights_i))`.
    d. Compute the centered X matrix and standardized X matrix.
    e. Diagonalize X centered.
    f. Do the screeplot of the eigenvalues and define the number of significant dimensions. How much is the retained information?
    g. Compute the projections of individuals in the significant dimensions.
    h. Compute the projection of variables in the significant dimensions
    i. Plot the individuals in the first factorial plane of R^p. Color the individuals according the "demo" variable.
    j. Plot the variables (as arrows) in the first factorial plane of R^n.
    k. According to the Russet complete data, justify which metric M is appropriate for this problem.
    l. Compute the correlation of the variables with the significant principal components and interpret them.
3. Redo 2, but taking the weight of Cuba equal to 0.
4. Now, study the sensibility of the performed PCA respect to considering Cuba as an outlier. Compute the correlations of the obtained *significant* principal components (Cuba 0 weight) with the previous obtained ones (all cases equal weights).

5. **Do again the PCA, but now using the library "FactoMineR" (be aware of using the completed data file with the "demo" factor as illustrative and the selected Metric).**
6. What is the country best represented in the first factorial plane?. And what is the worse?.
7. What are the three countries most influencing the formation of the first principal component?, and what are the three countries most influencing the formation of the second principal component?
8. What is the variable best represented in the first factorial plane?. And what is the worse?.

9. What are the three variables most influencing the formation of the first principal component?, and what are the three variables most influencing the formation of the second principal component?
10. Which modalities of the variable "demo" are significant in the first two principal components.
11. Use the NIPALS algorithm to obtain Principal Components in standardized PCA (as determined in previous questions) and with the results of the NIPALS, obtain the *biplot* of R^p. Interpret the results. Use unweighted data only.
12. Perform the *Varimax* rotation and plot the *rotated* variables. Interpret the new rotated components. Use unweighted data only.
13. Compute the scores of individuals in the rotated components `Psi.rot`. Interpret them (`xxxx$ind$coord[,1:nd] = Psi.rot; dimdesc(xxxx,axes=1:nd)`. Use unweighted data only.

~~**14. Read the PCA_quetaltecaen data.**~~

~~15. Symmetrize the data matrix, expressing the joint feeling between CCAA.~~

~~16. Transform the similarity matrix into a dissimilarity (notice that max. similarity = 10).~~

~~17. Perform the PCA upon the formed dissimilarity matrix.~~

~~18. Plot the first two components.~~

Lecturer Comments:

Your assignment main report is limited to 10 pages (an annex can be additionally included in the same document) and should be posted as a .pdf (Task Homework 2 – PDF). PDF file name should be Name1FamilyName1-Name2FamilyName2-HW2.pdf.

RScript/R Markdown has to be posted in ATENEA in the Task Homework 2 – Files/Scripts. Homework will be graded taken into account the following topics:

- Imputation and Weight selection. Normalized/Non-Normalized PCA selection.
- Unweighted PCA: own procedure
- Weighted PCA: own procedure
- Unweighted PCA with FactoMineR and Comparison issues with your own procedure
- Weighted PCA with FactoMineR and Comparison issues with your own procedure
- NIPALS and Biplot
- VARIMAX Rotation and Latent Factor Interpretation
- Sensibility PCA to outliers
- Final Conclusions