Multivariate Analysis: Stuff Beyond Principal Component Analysis

Pol Serra i Lidón
Polytechnic University of Catalonia
FIB, MIRI, Multivariate Analysis
Barcelona, Spain
Email: pol_serra_lidon@outlook.es

***Abstract*—This work about Multivariate Analysis is focused on the concepts of PCA, in concrete in the Nonlinear Iterative Partial Least Squares (NIPLES) algorithm to realize a PCA, and the Varimax rotation. Also, the symmetrization, similarity, and dissimilarity concepts regarding matrices are presented. Besides, main R code to give answer to the questions in is attached.**

I. Introduction

This work regarding Multivariative Analysis puts into practice the concepts beyond PCA. The work is organized as follows:
Section II briefly presents the NIPALS algorithm and the Varimax rotation within a theoretical perspective.
Section III answers questions from 1 to 5 regarding the $russet.txt$ dataset, developing each one at some extent. Section IV then answers the questions 6 to 10, regarding the $quetaltecaen.txt$ dataset, introducing the concepts of similarity-dissimilarity matrices. Conclusions are drawn in section V.

II. NIPALS algorithm and Varimax rotation

A. NIPALS algorithm

The NIPALS algorithm is a techique used to determine the eigenvectors of a matrix to perform the PCA. SVD is another method presented during the course. The NIPALS algorithm provides a more numerically accurate results than when compared to the SVD of the covariance matrix, but is most costly within a comuptational perspective. The NIPALS algorithm step-by-step is presented below, but first let:

- *X* be the mean centered data matrix.
- *E0=X* be the E-matrix for the zero-th PC.
- *t* be vector set to a column in X, that wil be the scores for PCi.
- *p* be the loadings for PCi.
- *tol* be the low threshold to convergence check.

The, the algorithm iterates as the follows:

- 1) Project X onto t to find the corresponding loading p.
- 2) Normalize loaidng vector p to length 1.
- 3) Project X onto p to find the corresponding score vector t.
- 4) Check for convergence. If the difference between new and old eigenvectors is larger than tol, return to 1
- 5) Remove the estimated PC component from E_{i-1}

B. Varimax rotation

In statistics, a varimax rotation a popular scheme for orthogonal rotations, used to simplify the expression of a particular sub-space in terms of just a few major items. The actual coordinate system is unchanged, it is the orthogonal basis that is being rotated to align with those coordinates. The sub-space found with principal component analysis or factor analysis is expressed as a dense basis with many non-zero weights which makes it hard to interpret. Varimax is so called because it maximizes the sum of the variances of the squared loadings (squared correlations between variables and factors).

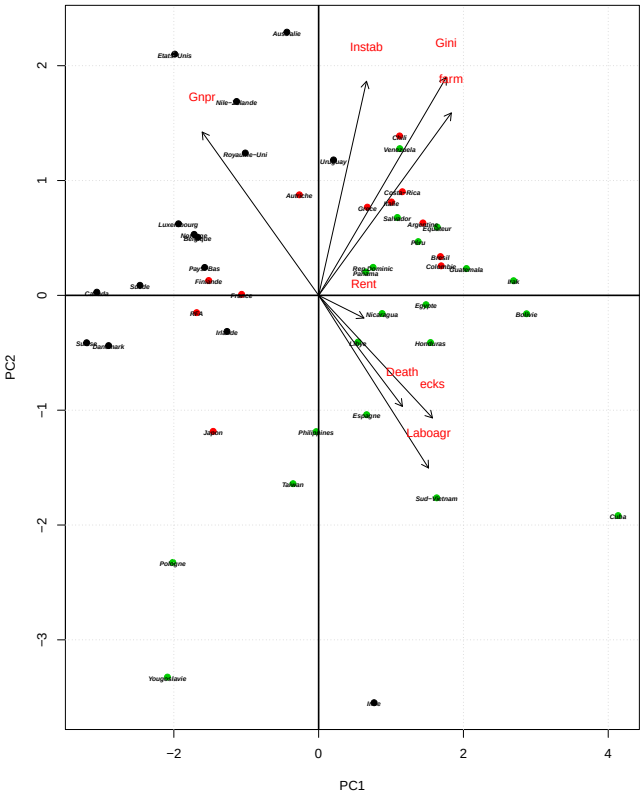

Fig. 1: Biplot of the $russet.txt$ dataset PCA obtained with the NIPALS algorithm

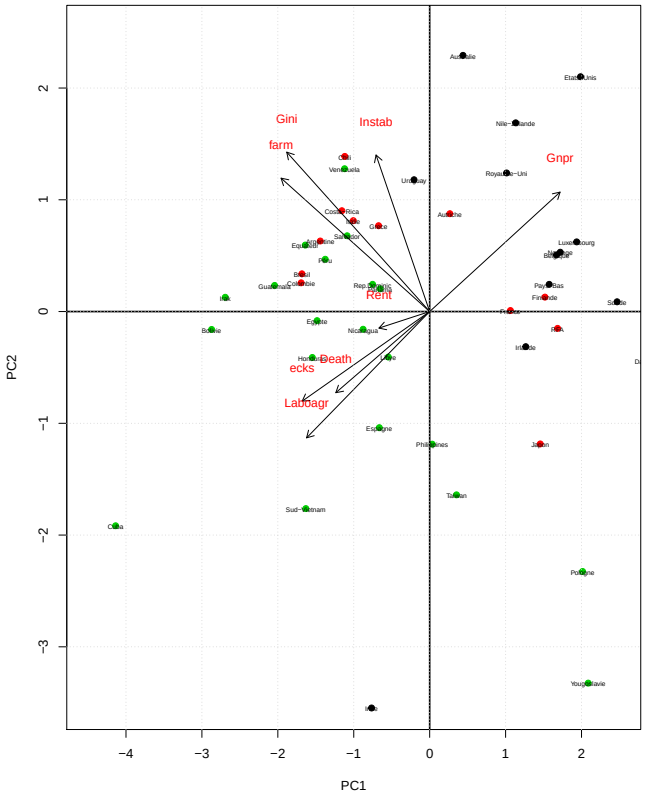

Fig. 2: Biplot of the $russet.txt$ dataset obtained with the $PCA()$ function

III. Questions 1 to 5

1) Read again the Russet data set and impute the missing values. Define as X the matrix formed by the standardized continuous variables.

The Russet data set, that is used as an example, tries to find the relation between political instability of countries and the economical and agricultural inequality, through the collected data of 47 countries on the period after the Second World War (1945-1962). Missing data is imputed to the data set using the kNN algorithm. X matrix is standardized with the R function $scale()$.

2) Obtain the Principals Components till the significant dimension you stated, using the NIPALS algorithm.

All in all, untill 3 Principal Components (PCs) were taken as significant by the Kaiser Rule. Although, during the realization of this report, only 2 PCs are taken into account, as only the representation of the scores of the countries in the first factorial plane are presented. The $nipals$ package is now used. The procedure can be seen in the commented attached R code.

3) With the results of the NIPALS, obtain the biplot of Rp. Interpret the results.

The biplot in Rp can be observed in Figure 1. As it can be seen compared to Figure 2, that represents the biplot obtained with the $PCA()$ function, only an axis rotation and arrow length modifications are noticed. Then, the interpretation of the results is quite similar than the one performed my previous work regarding the exploratory analysis of the Russet dataset. The PC1 represents the agro-economic situation of a country, while PC2 represents the unequality in each country.

Countries from Unstable countries and Dicatatorships tend to stay separated in PC1 from the ones with consolidated Democracies, as the democracies in tend to be wealthier than others (at least in the 50's). As we can observe in Figure 1 and 2, the countries that are ruled in a Democracy tend to be all together in the top of PC2 as they tend to be more equal, Unstable countries also tend to be equal

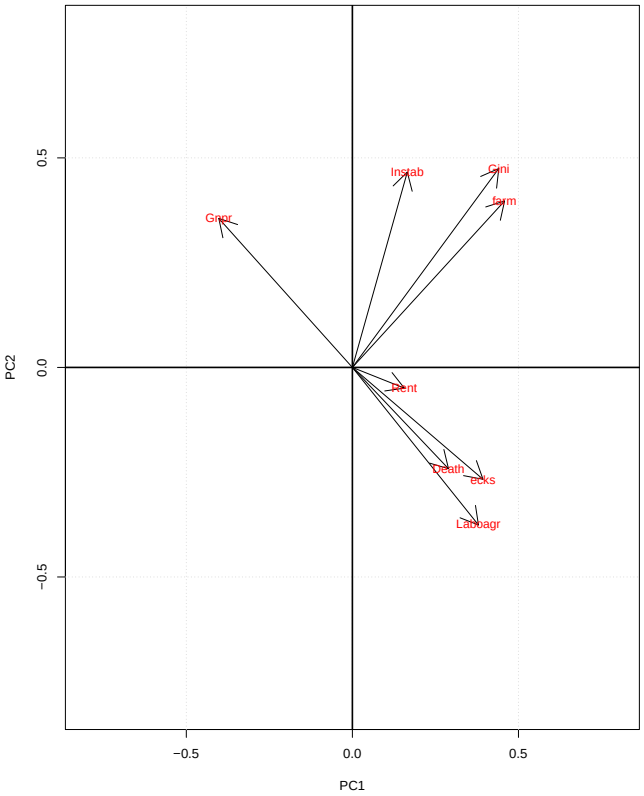

Fig. 3: Representation of the $russet.txt$ dataset loadings in the first factorial plane.

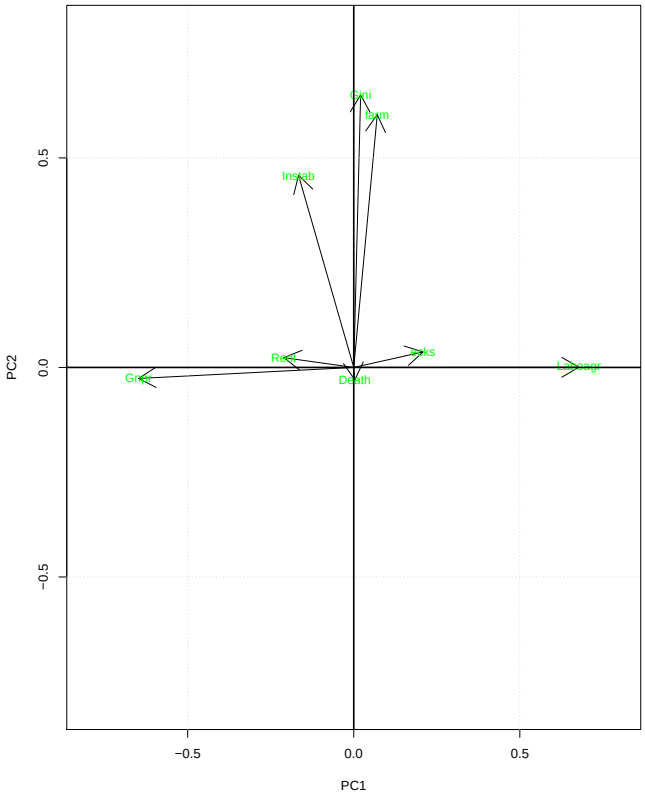

Fig. 4: Representation of the $russet.txt$ dataset loadings in the first factorial plane after the Varimax rotation.

(not as much as consolidated democracies), while Dictatorships tend to be down PC2, as they tend to be unequal.

4) Perform the Varimax rotation and plot the rotated variables. Interpret the new rotated components.

The $varimax()$ function is applied to the variables in the first factorial plane. After and later results can be observed in Figure 3 and Figure 4. As it can be observed, varimax performed an orthogonal rotation of the PCA applied on the factor loadings b analyzing the co-variance matrix of the factor loadings. The rotated components present information in a more explanatory way than the ones not rotated, offering a better reproducibility of the results.

5) Compute the scores of individuals in the rotated components Psi.rot.

Interpret them (xxxxindcoord[,1:nd] = Psi.rot; dimdesc(xxxx,axes=1:nd).

The representation of the scores of the individuals in the rotation components can be observed in Figure 5. As it is observed, Figure 5 makes the results much more visual than Figures 1 and 2.

IV. Questions 6 to 10

6) Read the PCAquetaltecaen data.

The $quetaltecaen.txt$ data represents what opinion the inhabitants in certain regions of Spain have about each other. In concrete, the question answered is how friendly are the other spaniard from certain regions to you. 8 regions that act as observations and variables are observed. No imputation or other pre-processing procedures are required.

7) Symmetrize the data matrix, expressing the joint feeling between CCAA.

To symmetrize the data, one possible procedure is to sum the matrix X to the transpose of X and divide by 2 ($(X + t(X))/2$), producing a symmetric matrix that express the joint feeling between regions.

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

To obtain the dissimilarity matrix, a matrix C is

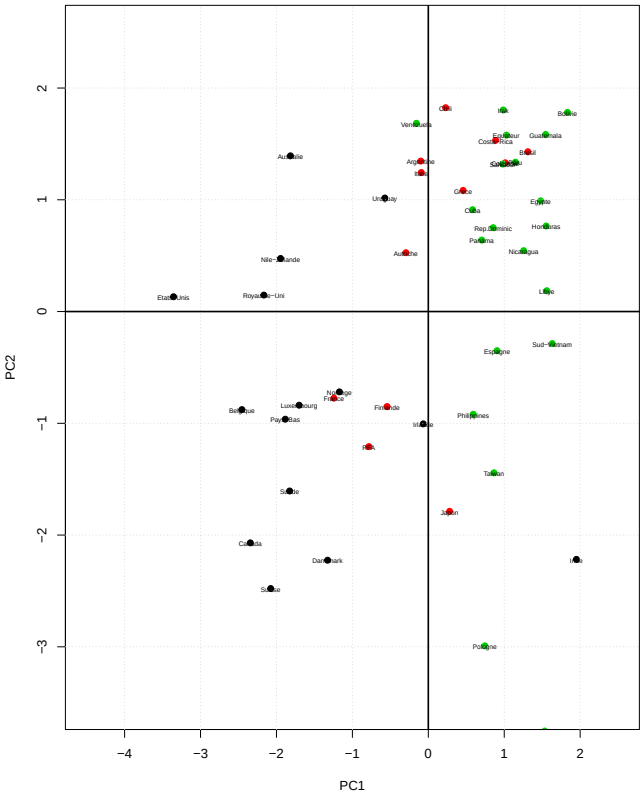

Fig. 5: Representation of the $russet.txt$ dataset scores in the first factorial plane after applying the Varimax rotation.

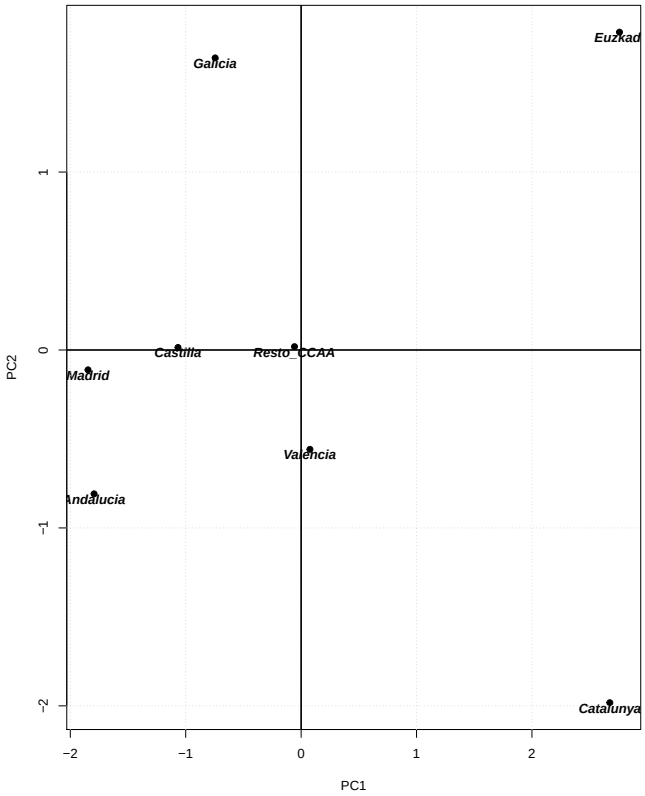

Fig. 6: Representation of the $quetaltecaen.txt$ dataset scores in the first factorial plane.

composed by the number 10 repeated among all rows and collums (8x8). The dissimilarity matrix is the result of substracting from this C matrix full of tens ($diss(X) = 10 - sim(X)$).

9) Perform the PCA upon the formed dissimilarity matrix.

The PCA is performed upon the dissimilarity matrix using the NIPALS algorithm that was previously used in Question 2. Therefore, PC1 and PC2 are obtained and ready to be used. The whole process can be observed in the attached code.

10) Plot the first two components.

The representation of the observations in the first factorial plane formed by PC1 and PC2 is plotted in Figure 6. As it can be observed, Catalonia and Euzkadi differ from the others, as they are the inhabitants living there are not as well perceived by the other habitants of Spain, perhaps speaking another language makes the mixing and perception between communities more difficult. Valencia and Galicia have a little different behaviour, while the others seem to stay close.

V. Conclusions

During this work, the NIPALS algorithm was defined, plus the Varimax rotation. This procedures were applied to the $russet.txt$ dataset, and results similar than the ones obtained in my previous analysis of the Russet dataset were obtained. Furthermore, the $quetaltecaen.txt$ dataset was read, symmetrized, transformed to a dissimilarity matrix and applied the PCA. Finally, the individual scores in the first factorial plane was plotted.