A Principal Component Analysis: Dictatorship and agriculture sector in the 50's

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

***Abstract*—This work within the scope of multivariative analysis, is focused on the concept of Principal Component Analysis (PCA), applied to a data set regarding the agriculture sector of some countries and its development. The Russet data set, that is under study, 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).**

I. Introduction

This work about the concept of PCA is developed trying to answer some questions regarding the analysis of the data. The work is organized as follows:
Section II presents the PCA approach very bierfly. There is so much material for the interested reader in Internet about PCA.
Section III presents the results of the function $PCAPolSerra.R$ when applied in the Russet data set, and presents answers from 1 to 4.
Section IV tries to answer the questions 5 to 10, using the $PCA()$ function of the $FactoMineR$ package. Finally, conclusions are drawn in Section V.

II. PCA Concept

Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.

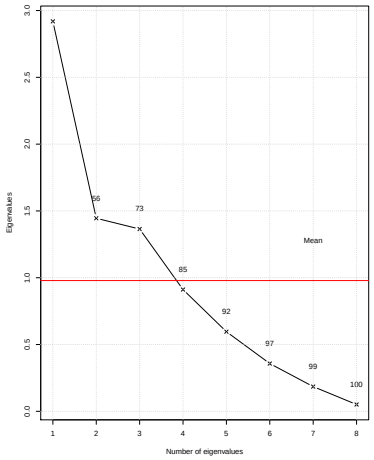

Fig. 1: Screeplot of the eigenvalues

III. Points 1 to 4

This sections tries to complete the points 1 to 4:
1) Read again the Russet comlete data set. Define as X matrix the one defined by the continuous variables.
To start the work properly, the Russet data set missign values are imputed with a kNN algorithm, and some treatments are applied.
2.b) Execute this function with the Russet complete data. Justify which metric M is appropriate for this problem. Compute the correlation of the variables with the significant principal components and interpret them.
The adequate M metric is the standardized one, as the variables are represented in different units, i.e. number of deaths in demonstrations and GNP.
The results of executing the function with the complete Russet data set can be found in Figure 1 and Figure 2, that represent the screeplot of eigenvalues and their contributions, the variables and observations in the first factorial plane formed by PC1 and PC2.
Regarding the correlation of the variables with the

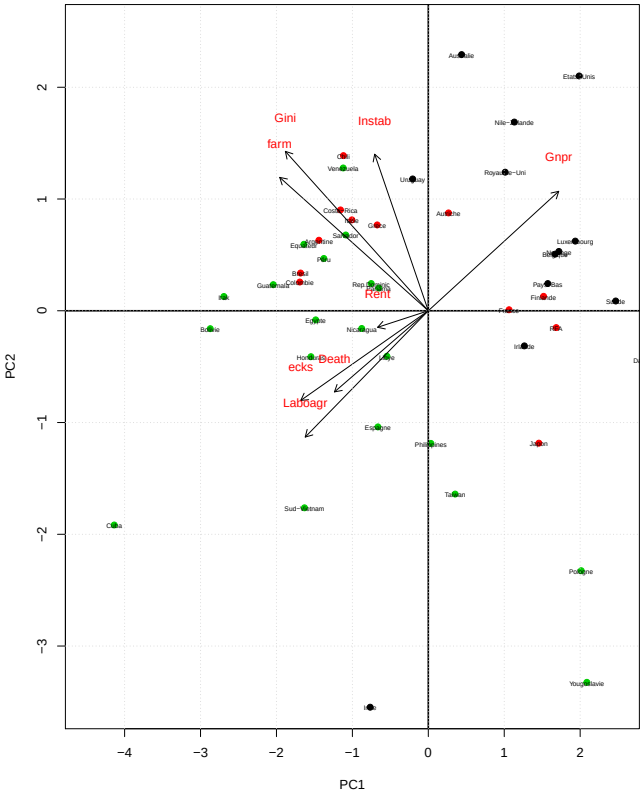

Fig. 2: Complete Russet data set representation of individuals and variables in the first factorial plane, formed by PC1 and PC2. Colours indicate if each country is Stable, Instable, or Dictatorship

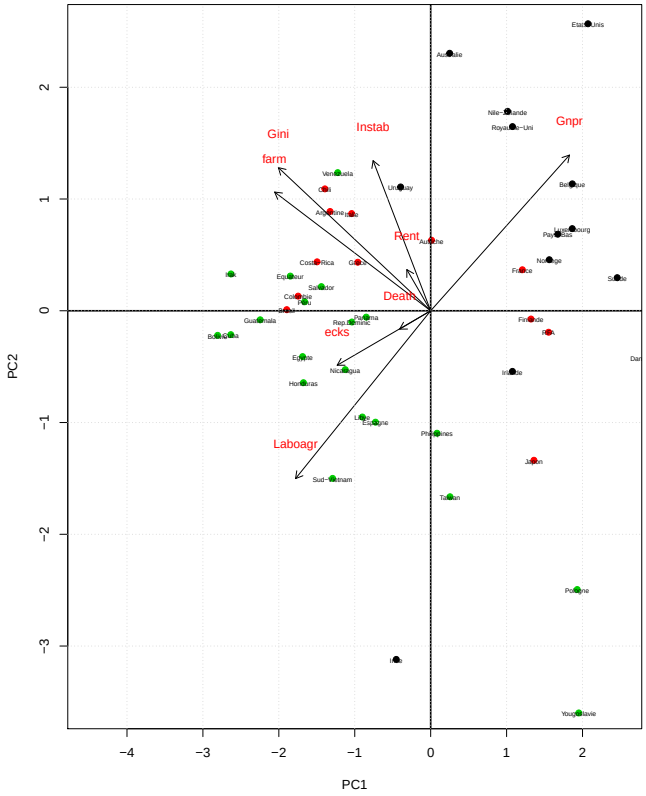

Fig. 3: Russet data set with Cuba taken as an outlier representation of individuals and variables in the first factorial plane, formed by PC1 and PC2. Colours indicate if each country is Stable, Instable, or Dictatorship

significant principal components, we can observe that in the first PC (PC1), all the variables have negative correlation unless Gnrp. This indicates that the countries with a high GNP per capita use to have low levels in the other variables. It seems coherent, as countries with higher GNP tend to have less armed conflicts, demonstrations, population working in the agriculture, present less deaths in demonstrations, or even have a smaller numer of prime ministers (this last no too much correlated even though).

In PC2, variables Laboagr, ecks, Death, and Rent, present a positive correlation, while Instab, Farm, and Gnpr present a negative one. It seems that the countries with more demonstrations, deaths and % of farmers not cultivating their own land, tend to have low Gnpr, and low % of small farmers with 50% of land.

3) Redo 2, but taking the weight of Cuba equal to 0. The projections of individuals and variables in the first factorial planes taking the weight of Cuba equal to 0 can be observed in Figure 3.

4) Now, study the sensibility of the performed PCA respect to considering Cuba as an outlier. Compute the correlations of the obtained significant components with the previous obtained ones.

Taking Cuba as an outlier, the variables Death and ecks now are less dominant than in the previous case, as Cuba was considered an outlier because his incredibly high number of armed conflicts and deaths during demonstrations during the decade of the 50s.

IV. Points 5 to 10

5) Do again the PCA, but now using the library "FactoMineR". (be aware of using the completed data file with the demo variable as illustrative).

In this Section, the PCA is performed through the "$FactoMineR$" package $PCA()$ function, and several questions are answered. Figure 4 and Figure 5 represent the correlation between variables, and the contribution of individuals in the first factorial plane of Rp. The call to the $PCA()$ function and the posterior analysis can

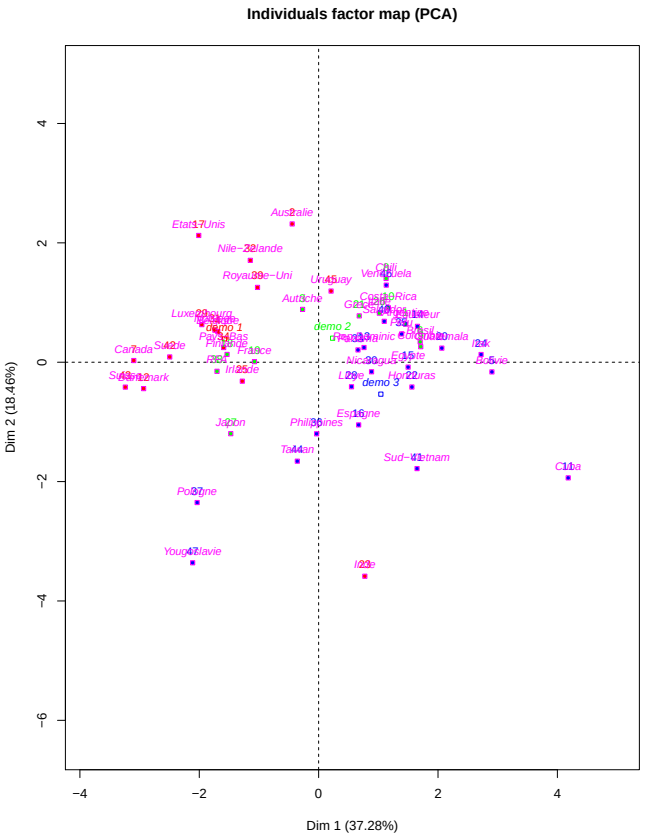

Fig. 4: Complete Russet data set representation of individuals in the first factorial plane, formed by PC1 and PC2.

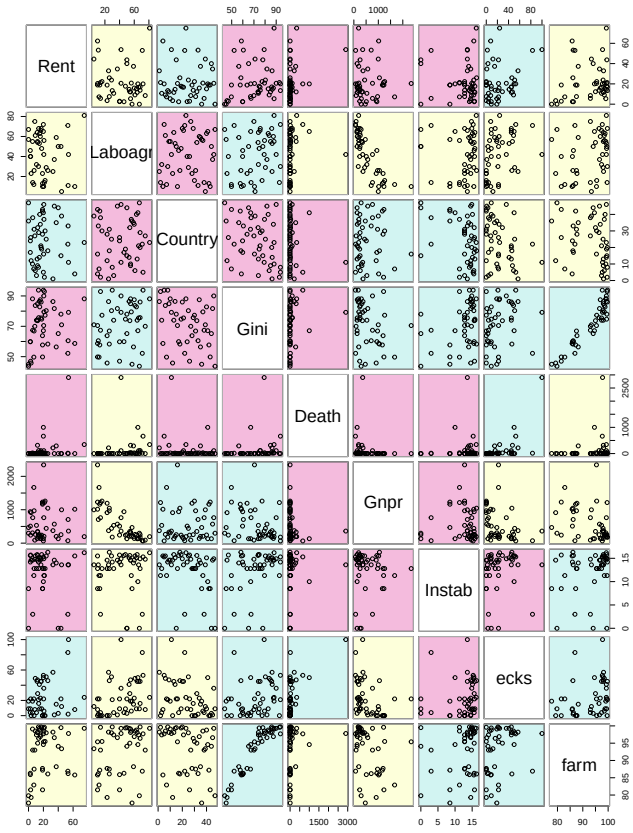

Fig. 5: Graphic showing the correlation between variables

be performed with following codelines:

```
X<-data[,1:10];
res<-PCA(X,quali.sup=c(1,10));
summary(res);
plot(res,habillage=10,cex=0.8,
col.hab=c("red","green","blue"));
```

Its result can be observed in Figure 4.

6) What is the country best represented in the first factorial plane? And what is the worse?

The contry best represented in the first factorial plane is Cuba, while the worst represented is Suisse.

As Cuba presents a low Gnpr, and presents high numbers in the other variables, highlighting and a very high number of deaths during demonstrations and violent conflicts, is situated in the farthest place. Contrary, Suisse presents the opposite conditions, as it is supposed to be a very wealthy country with low levels of social conflictivity.

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?

The three countries most infulencing the formation of the first principal component are Cuba, Canada, and Suisse.

They are the ones that have more impact as they present very high Gnpr numbers and low numbers in terms of (deaths, ecks, farm, gini, etc) in the case of Canada and Suisse, or viceversa, in the case of Cuba.

The three countries most infulencing the formation of the second principal component are Inde, Yougoslavie, and Pologne.

Those countries had a low number of prime ministers (Instab), Gini and Grp, while have had high number of deaths in demonstrations and armed conflicts. The relations between those variables are explained in the following questions.

8) What is the variable best represented in the first factorial plane?. And what is the worse?

The best represented variable in the first factorial plane is Farm, while the worse represented is Rent.

It could be interpreted that the percentage of small farmers with 50% of land has much more impact in the

other variables than the variable Rent, that represents the % of farmers not cultivating their own land, that seems less correlated to the others.

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?

The three variables most influencing the formation of the first principal component are Farm, Gini, and Gnpr. Farm represents the % of small farmers owners of at least half of their land, Gini represents the concentration index (Lorenz curve), and Gnpr is the GNP per capita in 1995.

This first PC shows the difference between the Gnpr (countries that are rich) and the other variables, that all have negative connotations. Besides, the three variables most influencing the formation of the second principal component are Gini, Instab, and Farm: representing Gini the concentration index (Lorenz curve), Instab the number of prime ministers during the whole period, and Farm the percentage of small farmers owners of at least half of their land.

It can be observed the second factorial plane presents the contraposition between two groups of variables:

- Gnpr, Instab, Gini and farm, considered in this case as positive.
- Death, ecks and Laboagr.

It can be observed that the second PC shows the difference between the countries that are very unequal in terms of distribution of the agrucultural resources and the countries that are very poor (lots of deaths, conflicts, and low Gnpr). All in all, one country can be very rich but very unequal, and that is what the second PC shows.

10)Which modalities of the variable demo are significant in the first two principal components?

As we can observe in Figure 3, countries with dictatorships regimes tend to be further from the centroid in the first factorial plane. Meanwhile, the countries in a unstable situation are disperse in the first PC but closer to the centroid, and the countries in a stable situation are distributed all together far from the centroid. Contributions of the variables in the first factorial plane obtained through the $PCA()$ function can be observed in Figure 6. The conclusion that we can extract from the answer of this question, is that countries within this data set ruled by dictators present contrary characteristics in terms of development and agriculture distribution than the democratic ones. Also, its worth mentioning that observing the second PC, it is noticialbe that countries

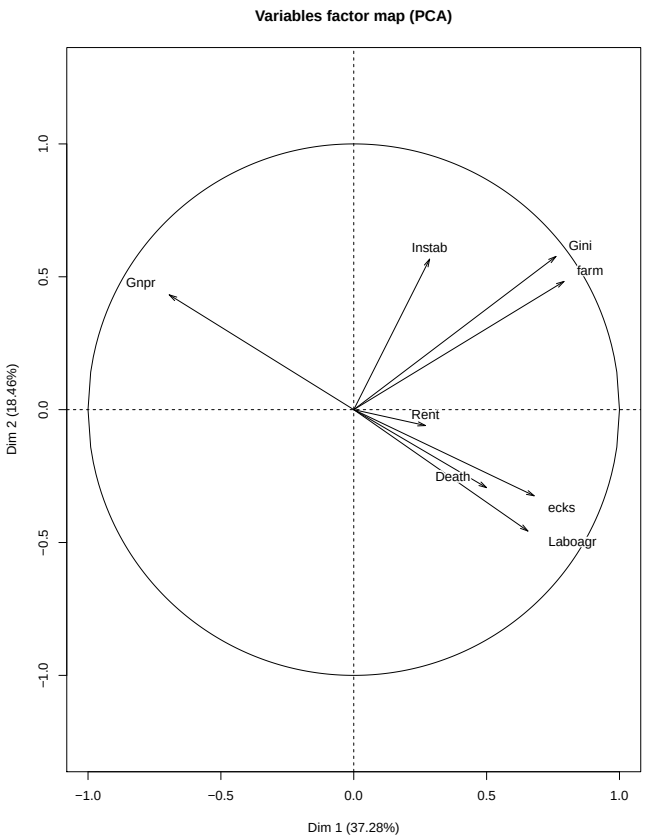

Fig. 6: Plot of the contributions of the variables in the first factorial plane.

with dictatorships tend to distribute the agriculture in a less fair way (as it is expected).

V. Conclusions

PCA permits to reduce the dimensionality of a data set, making it easier to deduce the hidden relations between the variables and individuals, usually difficult to notice in the first sight. It is not trivial to choose what PC should be taken as relevant, and several decisions can be applied following different rules (Kaiser, mean, elbow...). Computing and plotting the projections of individuals and variables in the first factorial plane is the main ilustrative source in PCA. R offers some standardized functions that allow the programmer to apply a PCA automatically, and observe the data in a more simple way instantly.