Contemporary Engineering Sciences, Vol. 11, 2018, no. 16, 763 - 769
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ces.2018.8120

Analysis of Economic and Social Indicators through the Principal Components Analysis

Fernando Mesa

Department of Mathematics and GEDNOL
Universidad Tecnológica de Pereira
Pereira, Colombia

German Correa Velez

Department of Mathematics
Universidad Tecnológica de Pereira
Pereira, Colombia

Pedro Pablo Cárdenas Alzate

Department of Mathematics and GEDNOL
Universidad Tecnológica de Pereira
Pereira, Colombia

Abstract

Principal component analysis (PCA) is a method used to identify patterns in the data and express them in such a way as to highlight their similarities and differences. Since data patterns can be difficult to find in arrays with high dimensions, where it is very complicated to make a graph, the PCA becomes a powerful tool for data analysis. Among the advantages that the PCA has is that once these patterns are found in the data, these are compressed and reduce the dimensions of the matrix and helping the graphic interpretation of this [1].

Keywords: PCA, Analysis of economic and social indicators

1 Introduction

The problem of dimension reduction underlies most of the Multivariate Analysis methods. Generically, it can be considered as follows: In engineering, you can find a wide variety of linear algebraic problems; and when trying to solve differential equations numerically [2], equations with an excessive number of variables can arise with a system of equations of the same size:

Is it possible to describe the information contained in some data by a number of variables lower than the number of variables observed?. The analysis of main components starts from a data matrix (centered) of n rows and p columns, which can be considered as a sample of size n of a random vector of dimension p,

$$X = (X_1, \ldots, X_p)'$$

It is considered a linear (univariate) combination of X

$$y = X't,$$

where t is a vector of weights of dimension p. The first principal component appears as a solution to the problem of finding the vector t that maximizes the variance of Y with the normalization condition $t't = 1$. In other words, the expression $var(Y)$ as a function of the vector of weights t gives rise to to a variational problem that has as its solution the first main component [3]. This problem is equivalent to finding the eigenvalues and auto-vectors of the covariance matrix of X. So that the successive principal components are obtained from the diagonalization of the covariance matrix of X,

$$S = TAT',$$

where T is an orthogonal matrix $p \times p$ whose columns are the coefficients of the main components.

1.1 Analysis of a correlation matrix

An analysis of major components makes sense if there are high correlations between the variables, since this is indicative of redundant information and, therefore, few factors explain much of the total variability.

The choice of factors is made in such a way that the first one picks up as much as possible of the original variability; the second factor must collect the maximum possible variability not collected by the first and so on, from the total of factors will be collected those, which collect the percentage of variability

that is considered sufficient. These will be called principal components.

Once the main components are selected, they will be represented as a matrix. Each element of this represents the factorial coefficients of the variables (the correlations between the variables and the main components). The matrix will have as many columns as main components and as many rows as there are variables [4].

For a factor to be easily interpretable it must have the following characteristics, which are difficult to achieve:

- The factor coefficients should be close to 1.
- A variable must have high coefficients with only one factor.
- There should not be factors with similar coefficients.

They are the scores that have the main components for each case, that allow us their graphical representation. It is calculated by the expression:

$$X_{ij} = a_{i1} \cdot Z_{1i} + \cdots + a_{ik} \cdot Z_{kj} = \sum_{s=1}^{k} a_{is} \cdot Z_{sk}$$

The a are the coefficients and the Z are the standardized values ??that have the variables in each of the subjects of the sample.

2 Process

To carry out a practical application [5] of the analysis of principal components, a table of real data with 11 economic and social indicators from 96 countries is drawn up. The observed variables are:

- $X1$: Annual rate of population growth,
- $X2$: Infant mortality rate per 1000 live births.
- $X3$: Percentage of women in the active population.
- $X4$: GNP in 1995 (in millions of dollars).
- $X5$: Production of electricity (in million Kw/h).
- $X6$: Telephone lines per 1000 inhabitants
- $X7$: Water consumption per-capita.

- $X8$: Proportion of the area of the country covered by forests.
- $X9$: Annual deforestation rate.
- $X10$: Energy consumption per-capita,
- $X11$: Emission of CO_2 per-capita.

An analysis of main components was carried out and if a conclusion from which matrix, S and R, is more appropriate, the first two components were also interpreted.

Let us first observe that the units of measure of the variables X_i are very different (percentages, dollars, kWh, ...). In addition, the high variances of $X4$ and $X5$ make it foresee that a principal component analysis made from the covariance matrix S will result in a first and second principal components that will basically coincide with these two observed variables. Therefore, the principal component analysis should be carried out from the correlation matrix R.

This is equivalent to standardizing each of the X_i to mean zero and unit variance and considering the covariance matrix of the standardized variables. The following Matlab function performs the analysis of principal components, first starting with S and, secondly, starting with R.

3 Results

Figure 4.1 contains the representation in major components of these countries and the percentage of variability explained by the first two components. We will interpret only the components calculated from R, since they are the most appropriate in this case. The coefficients of these two components are the first two columns of the matrix $T2$. The cumulative variability percentages are found in the accum vector.

$T2(:, 1:2)$	0.3141	0.3924
	-0.3484	0.0414
	0.0735	0.1776
	0.4403	0.1340
	0.3297	-0.0834
	-0.1839	-0.0866
	0.1629	0.6398
	-0.0948	-0.3231
	-0.5218	0.2903
	0.3467	-0.3896
	-0.1006	0.1749
acum2(1:2)	36.6353	54.1806

Table 1: Cumulative variability percentages

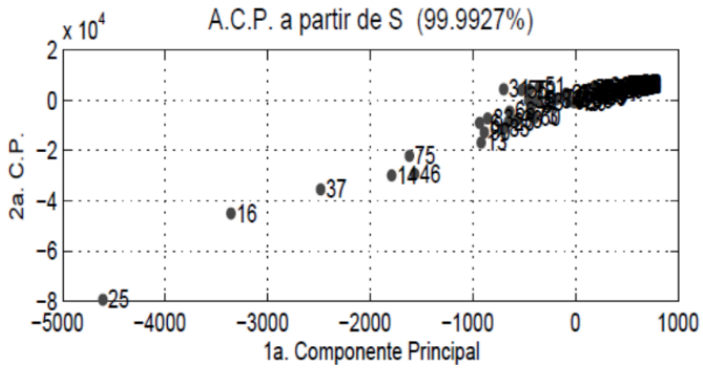

Figure 1: Main component.

4 Conclusion

From the graph obtained after reducing the dimension of the data to $R2$, the variables $X1, X2, X4, X5, X9$ and $X10$ are the ones that contribute the most in the first main component, which can be interpreted as an index of wealth. While $X1, X7, X8$ and $X10$ are the most contributing in the second component, which could be interpreted as an index of rurality. Thus, for example, the group of countries formed by Canada (16), China (25), France (37) and the United Kingdom (75) would be the richest according to this index that we have built, while Bangladesh (8) and Haiti (42) would be the poorest. On the other hand, Iran (48) and Pakistan (69) are the countries with the highest rurality index, while Finland (36) and Sweden (83) are on the opposite side.

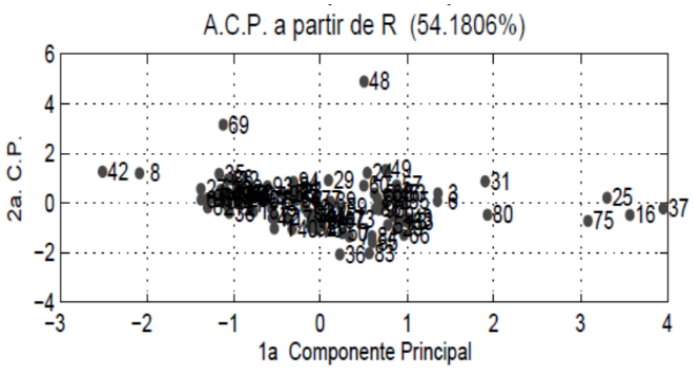

Figure 2: Main component.

- The abscissa collects the most variably combined data from all the initial data.
- The ordinate collects all the data that we still need to explain since the final graphs we can make a conclusion, it is also a linear combination of data.
- The information obtained at the end is sufficient to interpret the behavior of our data.
- The figure presents a clear classification of the data, which were organized from the transposed matrices of the adjusted data and the eigenvalues.

Acknowledgements. We would like to thank the referee for his valuable suggestions that improved the presentation of this paper and our gratitude to the Department of Mathematics of the Universidad Tecnológica de Pereira (Colombia) and the group GEDNOL.

References

[1] J. Baró and R. Alemany, Estadística II, Ed. Fundació per a la Universitat Oberta de Catalunya, 2000.

[2] D. Peña, Estadística, Modelos y Métodos, Alianza Editorial, Madrid, Vol. 84, 1987.

[3] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, **23** (1958), 187-200. https://doi.org/10.1007/bf02289233

[4] J. Kim and C. Mueller, *An Introduction to Factor Analysis*, What it is and how to do it, Sage Publications, Inc., 1978. https://doi.org/10.4135/9781412984652

[5] J. Múciga and R. Maya, El comportamiento del consumidor, Análisis del proceso de compra, (1997).

Received: February 12, 2018; Published: April 9, 2018