10

The Basics of Linear Principal Components Analysis

Yaya Keho
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSEA), Abidjan
Côte d'Ivoire

1. Introduction

When you have obtained measures on a large number of variables, there may exist redundancy in those variables. Redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same "thing". Because of this redundancy, it should be possible to reduce the observed variables into a smaller number of variables. For example, if a group of variables are strongly correlated with one another, you do not need all of them in your analysis, but only one since you can predict the evolution of all the variables from that of one. This opens the central issue of how to select or build the representative variables of each group of correlated variables. The simplest way to do this is to keep one variable and discard all others, but this is not reasonable. Another alternative is to combine the variables in some way by taking perhaps a weighted average, as in the line of the well-known Human Development Indicator published by UNDP. However, such an approach calls the basic question of how to set the appropriate weights. If one has sufficient insight into the nature and magnitude of the interrelations among the variables, one might choose weights using one's individual judgment. Obviously, this introduces a certain amount of subjectivity into the analysis and may be questioned by practitioners. To overcome this shortcoming, another method is to let the data set uncover itself the relevant weights of variables. Principal Components Analysis (PCA) is a variable reduction method that can be used to achieve this goal. Technically this method delivers a relatively small set of synthetic variables called principal components that account for most of the variance in the original dataset.

Introduced by Pearson (1901) and Hotelling (1933), Principal Components Analysis has become a popular data-processing and dimension-reduction technique, with numerous applications in engineering, biology, economy and social science. Today, PCA can be implemented through statistical software by students and professionals but it is often poorly understood. The goal of this Chapter is to dispel the magic behind this statistical tool. The Chapter presents the basic intuitions for how and why principal component analysis works, and provides guidelines regarding the interpretation of the results. The mathematics aspects will be limited. At the end of this Chapter, readers of all levels will be able to gain a better understanding of PCA as well as the when, the why and the how of applying this technique. They will be able to determine the number of meaningful components to retain from PCA, create factor scores and interpret the components. More emphasis will be placed on examples explaining in detail the steps of implementation of PCA in practice.

We think that the well understanding of this Chapter will facilitate that of the following chapters and novel extensions of PCA proposed in this book (sparse PCA, Kernel PCA, Multilinear PCA, …).

2. The basic prerequisite – Variance and correlation

PCA is useful when you have data on a large number of quantitative variables and wish to collapse them into a smaller number of artificial variables that will account for most of the variance in the data. The method is mainly concerned with identifying variances and correlations in the data. Let us focus our attention to the meaning of these concepts. Consider the dataset given in Table 1. This dataset will serve to illustrate how PCA works in practice.

ID	X1	X2	X3	X4	X5
1	24	21.5	5	2	14
2	16.7	21.4	6	2.5	17
3	16.78	23	7	2.2	15
4	17.6	22	8.7	3	20
5	22	25.7	6.4	2	14.2
6	15.3	16	8.7	2.21	15.3
7	10.2	19	4.3	2.2	15.3
8	11.9	17.1	4.5	2	14
9	14.3	19.1	6	2.2	15
10	8.7	14.3	4.1	2.24	15.5
11	6.7	10	3.8	2.23	16
12	7.1	13	2.8	2.01	12
13	10.3	16	4	2	14.5
14	7.1	13	3.9	2.4	16.4
15	7.9	13.6	4	3.1	20.2
16	3	8	3.4	2.1	14.7
17	3	9	3.3	3	20.2
18	1	7.5	3	2	14
19	0.8	7	2.8	2	15.8
20	1	4	3.1	2.2	15.3

Table 1. **Example dataset, 5 variables obtained for 20 observations.**

The variance of a given variable x is defined as the average of the squared differences from the mean:

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{1}$$

The square root of the variance is the standard deviation and is symbolized by the small Greek sigma σ_x . It is a measure of how spread out numbers are.

The variance and the standard deviation are important in data analysis because of their relationships to correlation and the normal curve. Correlation between a pair of variables measures to what extent their values co-vary. The term covariance is undoubtedly associatively prompted immediately. There are numerous models for describing the behavioral nature of a simultaneous change in values, such as linear, exponential and more. The linear correlation is used in PCA. The linear correlation coefficient for two variables x and y is given by:

$$\rho(x,y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y} \quad (2)$$

where σ_x and σ_y denote the standard deviation of x and y, respectively. This definition is the most widely-used type of correlation coefficient in statistics and is also called Pearson correlation or product-moment correlation. Correlation coefficients lie between -1.00 and +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation. Correlation coefficients are used to assess the degree of collinearity or redundancy among variables. Notice that the value of correlation coefficient does not depend on the specific measurement units used.

When correlations among several variables are computed, they are typically summarized in the form of a correlation matrix. For the five variables in Table 1, we obtain the results reported in Table 2.

	X1	X2	X3	X4	X5
X1	1.00	0.94	0.77	-0.03	-0.08
X2		1.00	0.74	0.02	-0.04
X3			1.00	0.21	0.19
X4				1.00	0.95
X5					1.00

Table 2. **Correlations among variables**

In this Table a given row and column intersect shows the correlation between the two corresponding variables. For example, the correlation between variables X_1 and X_2 is 0.94.

As can be seen from the correlations, the five variables seem to hang together in two distinct groups. First, notice that variables X_1, X_2 and X_3 show relatively strong correlations with one another. This could be because they are measuring the same "thing". In the same way, variables X_4 and X_5 correlate strongly with each another, a possible indication that they measure the same "thing" as well. Notice that those two variables show very weak correlations with the rest of the variables.

Given that the 5 variables contain some "redundant" information, it is likely that they are not really measuring five different independent constructs, but two constructs or underlying factors. What are these factors? To what extent does each variable measure each of these factors? The purpose of PCA is to provide answers to these questions. Before presenting the mathematics of the method, let's see how PCA works with the data in Table 1.

In linear PCA each of the two artificial variables is computed as the linear combination of the original variables.

$$Z = \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_5 X_5 \tag{3}$$

where α_j is the weight for variable j in creating the component Z. The value of Z for a subject represents the subject's score on the principal component.

Using our dataset, we have:

$$Z_1 = 0.579 X_1 + 0.577 X_2 + 0.554 X_3 + 0.126 X_4 + 0.098 X_5 \tag{4}$$

$$Z_2 = -0.172 X_1 - 0.14 X_2 + 0.046 X_3 + 0.685 X_4 + 0.693 X_5 \tag{5}$$

Notice that different coefficients were assigned to the original variables in computing subject scores on the two components. X_1, X_2 and X_3 are assigned relatively large weights that range from 0.554 to 0.579, while variables X_4 and X_5 are assigned very small weights ranging from 0.098 to 0.126. As a result, component Z_1 should account for much of the variability in the first three variables. In creating subject scores on the second component, much weight is given to X_4 and X_5, while little weight is given to X_1, X_2 and X_3. Subject scores on each component are computed by adding together weighted scores on the observed variables. For example, the value of a subject along the first component Z_1 is 0.579 times the standardized value of X_1 plus 0.577 times the standardized value of X_2 plus 0.554 times the standardized value of X_3 plus 0.126 times the standardized value of X_4 plus 0.098 times the standardized value of X_5.

At this stage of our analysis, it is reasonable to wonder how the weights from the preceding equations are determined. Are they optimal in the sense that no other set of weights could produce components that best account for variance in the dataset? How principal components are computed?

3. Heterogeneity and standardization of data

3.1 Graphs and distances among points

Our dataset in **Table 1** can be represented into two graphs: one representing the subjects, and the other the variables. In the first, we consider each subject (individual) as a vector with coordinates given by the 5 observations of the variables. Clearly, the cloud of points belongs to a R^5 space. In the second one each variable is regarded as a vector belonging to a R^{20} space.

We can calculate the centroide of the cloud of points which coordinates are the 5 means of the variables, that is $g = (\bar{X}_1, ..., \bar{X}_5)$. Again, we can compute the overall variance of the points by summing the variance of each variable:

$$I = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - \bar{X}_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n} d^2(s_i, g) = \sum_{j=1}^{p}\sigma_j^2 \tag{6}$$

This quantity measures how spread out the points are around the centroid. We will need this quantity when determining principal components.

We define the distance between subjects s_i and $s_{i'}$ using the Euclidian distance as follows:

$$d^2(s_i, s_{i'}) = \|s_i - s_{i'}\|^2 = \sum_{j=1}^{p=5} (X_{ij} - X_{i'j})^2 \tag{7}$$

Two subjects are close one to another when they take similar values for all variables. We can use this distance to measure the overall dispersion of the data around the centroid or to cluster the points as in classification methods.

3.2 How work when data are in different units?

There are different problems when variables are measured in different units. The first problem is the meaning of the variance: how to sum quantities with different measurement units? The second problem is that the distance between points can be greatly influenced. To illustrate this point, let us consider the distances between subjects 7, 8 and 9. Applying Eq.(7), we obtain the following results:

$$d^2(s_7, s_8) = (10.2 - 11.9)^2 + (19 - 17.1)^2 + ... + (15.3 - 14)^2 = 8.27 \tag{8}$$

$$d^2(s_7, s_9) = (10.2 - 14.3)^2 + (19 - 19.1)^2 + ... + (15.3 - 15)^2 = 19.8 \tag{9}$$

Subject 7 is closer to subject 8 than to subject 9. Multiplying the values of variable X_5 by 10 yields:

$$d^2(s_7, s_8) = (10.2 - 11.9)^2 + (19 - 17.1)^2 + ... + (153 - 140)^2 = 175.58 \tag{10}$$

$$d^2(s_7, s_9) = (10.2 - 14.3)^2 + (19 - 19.1)^2 + ... + (153 - 150)^2 = 28.71 \tag{11}$$

Now we observe that subject 7 is closer to subject 9 than to subject 8. It is hard to accept how the measurement units of the variables can change greatly the comparison results among subjects. Indeed, we could by this way render a tall man as shorter as we want!

As seen, PCA is sensitive to scale. If you multiply one variable by a scalar you get different results. In particular, the principal components are dependent on the units used to measure the original variables as well as on the range of values they assume (variance). This makes comparison very difficult. It is for these reasons we should *often* standardize the variables prior to using PCA. A common standardization method is to subtract the mean and divide by the standard deviation. This yields the following:

$$X_i^* = \frac{X_i - \bar{X}}{\sigma_x} \tag{12}$$

where $\bar{X}$ and σ_x are the mean and standard deviation of X, respectively.

Thus, the new variables all have zero mean and unit standard deviation. Therefore the total variance of the data set is the number of observed variables being analyzed.

Throughout, we assume that the data have been centered and standardized. Graphically, this implies that the centroid or center of gravity of the whole dataset is at the origin. In this case, the PCA is called normalized principal component analysis, and will be based on the correlation matrix (and not on variance-covariance matrix). The variables will lie on the unit sphere; their projection on the subspace spanned by the principal components is the "correlation circle". Standardization allows the use of variables which are not measured in the same units (e.g. temperature, weight, distance, size, etc.). Also, as we will see later, working with standardized data makes interpretation easier.

4. The mathematics of PCA: An eigenvalue problem

Now we have understood the intuitions of PCA, we present the mathematics behind the method by considering a general case. More details on technical aspects can be found in Cooley & Lohnes (1971), Stevens (1986), Lebart, Morineau & Piron (1995), Cadima & Jolliffe (1995), Hyvarinen, Karhunen & Oja (2001), and Jolliffe (2002).

Consider a dataset consisting of p variables observed on n subjects. Variables are denoted by $(x_1, x_2, ..., x_p)$. In general, data are in a table with the rows representing the subjects (individuals) and the columns the variables. The dataset can also be viewed as a $n \times p$ rectangular matrix X. Note that variables are such that their means make sense. The variables are also standardized.

We can represent these data in two graphs: on the one hand, in a subject graph where we try to find similarities or differences between subjects, on the other, in a variable graph where we try to find correlations between variables. Subjects graph belongs to an p-dimensional space, i.e. to R^p, while variables graph belongs to an n-dimensional space, i.e. to R^n. We have two clouds of points in high-dimensional spaces, too large for us to plot and see something in them. We cannot see beyond a three-dimensional space! The PCA will give us a subspace of reasonable dimension so that the projection onto this subspace retains "as much as possible" of the information present in the dataset, i.e., so that the projected clouds of points be as "dispersed" as possible. In other words, the goal of PCA is to compute another basis that best re-express the dataset. The hope is that this new basis will filter out the noise and reveal hidden structure.

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ . \\ . \\ . \\ x_{pi} \end{pmatrix} \rightarrow \text{reduce dimensionality} \rightarrow z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ . \\ . \\ . \\ z_{qi} \end{pmatrix} \text{ with } q < p \qquad (13)$$

Dimensionality reduction implies information loss. How to represent the data in a lower-dimensional form without losing too much information? Preserve as much information as possible is the objective of the mathematics behind the PCA procedure.

We first of all assume that we want to project the data points on a 1-dimensional space. The principal component corresponding to this axis is a linear combination of the original variables and can be expressed as follows:

$$z_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + ... + \alpha_{1p}x_p = Xu_1 \tag{14}$$

where $u_1 = (\alpha_{11}, \alpha_{12}, ...\alpha_{1p})'$ is a column vector of weights. The principal component z_1 is determined such that the overall variance of the resulting points is as large as possible. Of course, one could make the variance of z_1 as large as possible by choosing large values for the weights $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$. To prevent this, weights are calculated with the constraint that their sum of squares is one, that is u_1 is a unit vector subject to the constraint:

$$\alpha_{11}^2 + \alpha_{12}^2 + ... + \alpha_{1p}^2 = \|u_1\|^2 = 1 \tag{15}$$

Eq.(14) is also the projections of the *n* subjects on the first component. PCA finds u_1 so that

$$Var(z_1) = \frac{1}{n}\sum_{i=1}^{n} z_{1i}^2 = \frac{1}{n}\|z_1\|^2 = \frac{1}{n}u'_1 X'Xu_1 \text{ is maximal} \tag{16}$$

The matrix $C = \frac{1}{n}X'X$ is the correlation matrix of the variables. The optimization problem is:

$$\underset{\substack{u_1 \\ \|u_1\|^2=1}}{Max}\, u'_1 Cu_1 \tag{17}$$

This program means that we search for a unit vector u_1 so as to maximize the variance of the projection on the first component. The technique for solving such optimization problems (linearly constrained) involves a construction of a Lagrangian function.

$$\Im_1 = u'_1 Cu_1 - \lambda_1(u'_1 u_1 - 1) \tag{18}$$

Taking the partial derivative $\partial\Im_1 / \partial u_1 = Cu_1 - \lambda_1 u_1$ and solving the equation $\partial\Im_1 / \partial u_1 = 0$ yields:

$$Cu_1 = \lambda_1 u_1 \tag{19}$$

By premultiplying each side of this condition by u'_1 and using the condition $u'_1 u_1 = 1$ we get:

$$u'_1 Cu_1 = \lambda_1 u'_1 u_1 = \lambda_1 \tag{20}$$

It is known from matrix algebra that the parameters u_1 and λ_1 that satisfy conditions (19) and (20) are the maximum eigenvalue and the corresponding eigenvector of the correlation matrix C. Thus the optimum coefficients of the original variables generating the first principal component z_1 are the elements of the eigenvector corresponding to the largest eigenvalue of the correlation matrix. These elements are also known as loadings.

The second principal component is calculated in the same way, with the condition that it is uncorrelated (orthogonal) with the first principal component and that it accounts for the largest part of the remaining variance.

$$z_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + ... + \alpha_{2p}x_p = Xu_2 \tag{21}$$

where $u_2 = (\alpha_{21}, \alpha_{22}, ...\alpha_{2p})'$ is the direction of the component. This axis is constrained to be orthogonal to the first one. Thus, the second component is subject to the constraints:

$$\alpha_{21}^2 + \alpha_{22}^2 + ... + \alpha_{2p}^2 = \|u_2\|^2 = 1, \quad u'_1 u_2 = 0 \tag{22}$$

The optimization problem is therefore:

$$\underset{\substack{u_2 \\ \|u_2\|^2=1 \\ u'_1 u_2=0}}{Max}\ u'_2 C u_2 \tag{23}$$

Using the technique of Lagrangian function the following conditions:

$$Cu_2 = \lambda_2 u_2 \tag{24}$$

$$u'_2 C u_2 = \lambda_2 \tag{25}$$

are obtained again. So once more the second vector comes to be the eignevector corresponding to the second highest eigenvalue of the correlation matrix.

Using induction, it can be proven that PCA is a procedure of eigenvalue decomposition of the correlation matrix. The coefficients generating the linear combinations that transform the original variables into uncorrelated variables are the eigenvectors of the correlation matrix. This is a good new, because finding eigenvectors is something which can be done rapidly using many statistical packages (SAS, Stata, R, SPSS, SPAD…), and because eigenvectors have many nice mathematical properties. Note that rather than maximizing variance, it might sound more plausible to look for the projection with the smallest average (mean-squared) distance between the original points and their projections on the principal components. This turns out to be equivalent to maximizing the variance (Pythagorean Theorem).

An interesting property of the principal components is that they are all uncorrelated (orthogonal) to one another. This is because matrix C is a real symmetric matrix and then linear algebra tells us that it is diagonalizable and the eigenvectors are orthogonal to one another. Again because C is a covariance matrix, it is a positive matrix in the sense that $u'Cu \geq 0$ for any vector u. This tells us that the eigenvalues of C are all non-negative.

$$\text{var}(z) = \begin{bmatrix} \lambda_1 & 0 & . & 0 \\ 0 & \lambda_2 & & \\ . & & . & \\ 0 & & & \lambda_p \end{bmatrix} \tag{26}$$

The eigenvectors are the "preferential directions" of the data set. The principal components are derived in decreasing order of importance; and have a variance equal to their corresponding eigenvalue. The first principal component is the direction along which the data have the most variance. The second principal component is the direction orthogonal to

the first component with the most variance. It is clear that all components explain together 100% of the variability in the data. This is why we say that PCA works like a change of basis. Analyzing the original data in the canonical space yields the same results than examining it in the components space. However, PCA allows us to obtain a linear projection of our data, originally in R^p, onto R^q, where $q < p$. The variance of the projections on to the first q principal components is the sum of the eigenvalues corresponding to these components. If the data fall near a q-dimensional subspace, then p-q of the eigenvalues will be nearly zero.

Summarizing the computational steps of PCA

Suppose $x_1, x_2, ..., x_p$ are $p \times 1$ vectors collected from n subjects. The computational steps that need to be accomplished in order to obtain the results of PCA are the following:

Step 1. Compute mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Step 2. Standardize the data: $\Phi_i = \frac{x_i - \bar{x}}{\sigma_x}$

Step 3. Form the matrix $A = \left[\Phi_1, \Phi_2, ..., \Phi_p\right]$ ($p \times n$ matrix), then compute:

$$C = \frac{1}{n}\sum_{i=1}^{n} \Phi'_i \Phi_i$$

Step 4. Compute the eigenvalues of C: $\lambda_1 > \lambda_2 > ... > \lambda_p$

Step 5. Compute the eigenvectors of C: $u_1, u_2, ..., u_p$

Step 6. Proceed to the linear tranformation R^p ->R^q that performs the dimensionality reduction.

Notice that, in this analysis, we gave the same weight to each subject. We could have give more weight to some subjects, to reflect their representativity in the population.

5. Criteria for determining the number of meaningful components to retain

In principal component analysis the number of components extracted is equal to the number of variables being analyzed (under the general condition $n > p$). This means that an analysis of our 5 variables would actually result in 5 components, not two. However, since PCA aims at reducing dimensionality, only the first few components will be important enough to be retained for interpretation and used to present the data. It is therefore reasonable to wonder how many independent components are necessary to best describe the data.

Eigenvalues are thought of as quantitative assessment of how much a component represents the data. The higher the eigenvalues of a component, the more representative it is of the data. Eigenvalues are therefore used to determine the meaningfulness of components. Table 3 provides the eigenvalues from the PCA applied to our dataset. In the column headed "Eignenvalue", the eigenvalue for each component is presented. Each raw in the table presents information about one of the 5 components: the raw "1" provides information about the first component (PCA1) extracted, the raw "2" provides information about the second component (PCA2) extracted, and so forth. Eigenvalues are ranked from the highest to the lowest.

It can be seen that the eigenvalue for component 1 is 2.653, while the eigenvalue for component 2 is 1.98. This means that the first component accounts for 2.653 units of total variance while the second component accounts for 1.98 units. The third component accounts for about 0.27 unit of variance. Note that the sum of the eigenvalues is 5, which is also the number of variables. How do we determine how many components are worth interpreting?

Component	Eigenvalue	% of variance	Cumulative %
1	2.653	53.057	53.057
2	1.980	39.597	92.653
3	0.269	5.375	98.028
4	0.055	1.095	99.123
5	0.044	0.877	100.000

Table 3. **Eigenvalues from PCA**

Several criteria have been proposed for determining how many meaningful components should be retained for interpretation. This section will describe three criteria: the Kaiser eigenvalue-one criterion, the Cattell Scree test, and the cumulative percent of variance accounted for.

5.1 Kaiser method

The Kaiser (1960) method provides a handy rule of thumb that can be used to retain meaningful components. This rule suggests keeping only components with eigenvalues greater than 1. This method is also known as the eigenvalue-one criterion. The rationale for this criterion is straightforward. Each observed variable contributes one unit of variance to the total variance in the data set. Any component that displays an eigenvalue greater than 1 is accounts for a greater amount of variance than does any single variable. Such a component is therefore accounting for a meaningful amount of variance, and is worthy of being retained. On the other hand, a component with an eigenvalue of less than 1 accounts for less variance than does one variable. The purpose of principal component analysis is to reduce variables into a relatively smaller number of components; this cannot be effectively achieved if we retain components that account for less variance than do individual variables. For this reason, components with eigenvalues less than 1 are of little use and are not retained. When a covariance matrix is used, this criterion retains components whose eigenvalue is greater than the average variance of the data (Kaiser-Guttman criterion).

However, this method can lead to retaining the wrong number of components under circumstances that are often encountered in research. The thoughtless application of this rule can lead to errors of interpretation when differences in the eigenvalues of successive components are trivial. For example, if component 2 displays an eigenvalue of 1.01 and component 3 displays an eigenvalue of 0.99, then component 2 will be retained but component 3 will not; this may mislead us into believing that the third component is meaningless when, in fact, it accounts for almost exactly the same amount of variance as the second component. It is possible to use statistical tests to test for difference between successive eigenvalues. In fact, the Kaiser criterion ignores error associated with each

eigenvalue due to sampling. Lambert, Wildt and Durand (1990) proposed a bootstrapped version of the Kaiser approach to determine the interpretability of eigenvalues.

Table 3 shows that the first component has an eigenvalue substantially greater than 1. It therefore explains more variance than a single variable, in fact 2.653 times as much. The second component displays an eigenvalue of 1.98, which is substantially greater than 1, and the third component displays an eigenvalue of 0.269, which is clearly lower than 1. The application of the Kaiser criterion leads us to retain unambiguously the first two principal components.

5.2 Cattell scree test

The scree test is another device for determining the appropriate number of components to retain. First, it graphs the eigenvalues against the component number. As eigenvalues are constrained to decrease monotonically from the first principal component to the last, the scree plot shows the decreasing rate at which variance is explained by additional principal components. To choose the number of meaningful components, we next look at the scree plot and stop at the point it begins to level off (Cattell, 1966; Horn, 1965). The components that appear *before* the "break" are assumed to be meaningful and are retained for interpretation; those appearing *after* the break are assumed to be unimportant and are not retained. Between the components before and after the break lies a scree.

The scree plot of eigenvalues derived from Table 3 is displayed in Figure 1. The component numbers are listed on the horizontal axis, while eigenvalues are listed on the vertical axis. The Figure shows a relatively large break appearing between components 2 and 3, meaning the each successive component is accounting for smaller and smaller amounts of the total variance. This agrees with the preceding conclusion that two principal components provide a reasonable summary of the data, accounting for about 93% of the total variance.

Sometimes a scree plot will display a pattern such that it is difficult to determine exactly where a break exists. When encountered, the use of the scree plot must be supplemented with additional criteria, such as the Kaiser method or the cumulative percent of variance accounted for criterion.

5.3 Cumulative percent of total variance accounted for

When determining the number of meaningful components, remember that the subspace of components retained must account for a reasonable amount of variance in the data. It is usually typical to express the eigenvalues as a percentage of the total. The fraction of an eigenvalue out of the sum of all eigenvalues represents the amount of variance accounted by the corresponding principal component. The cumulative percent of variance explained by the first q components is calculated with the formula:

$$r_q = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100 \tag{27}$$

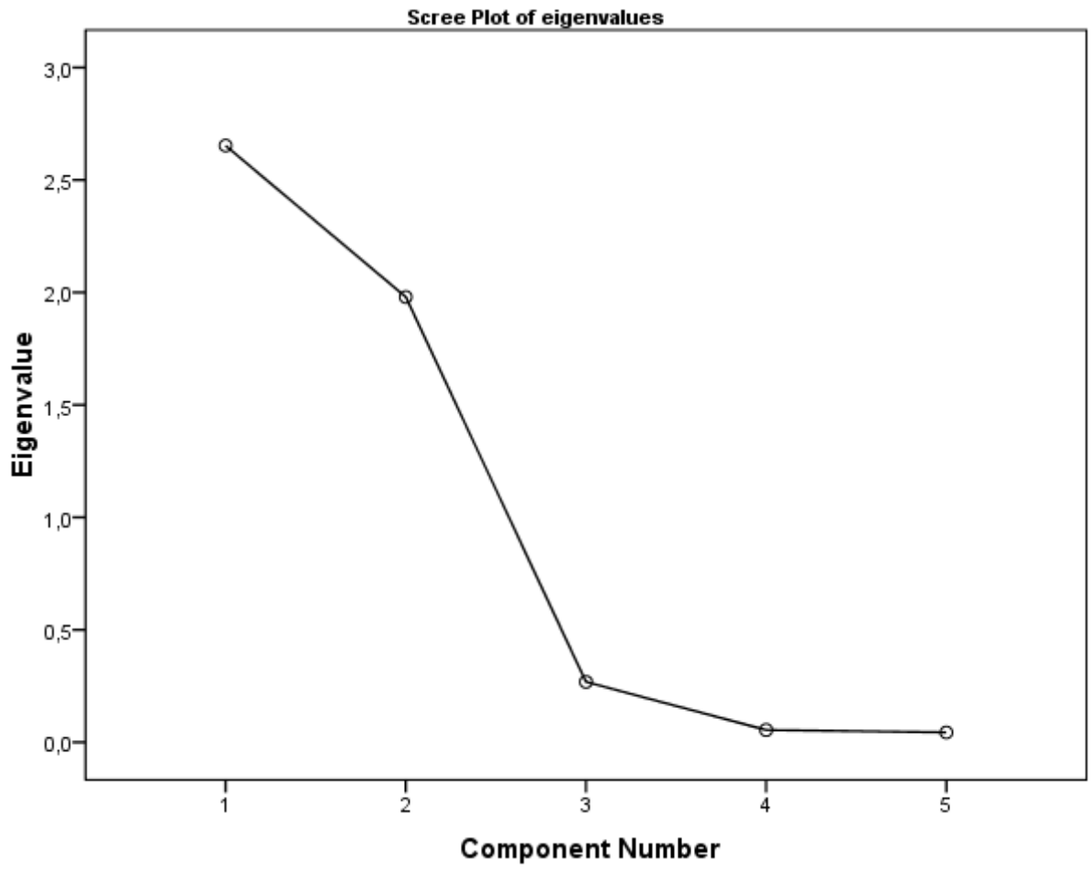

Fig. 1. **Scree plot of eigenvalues**

How many principal components we should use depends on how big an r_q we need. This criterion involves retaining all components up to a total percent variance (Lebart, Morineau & Piron, 1995; Jolliffe, 2002). It is recommended that the components retained account for at least 60% of the variance. The principal components that offer little increase in the total variance explained are ignored; those components are considered to be noise. When PCA works well, the first two eigenvalues usually account for more than 60% of the total variation in the data.

In our current example, the percentage of variance accounted for by each component and the cumulative percent variance appear in Table 3. From this Table we can see that the first component alone accounts for 53.057% of the total variance and the second component alone accounts for 39.597% of the total variance. Adding these percentages together results in a sum of 92.65%. This means that the cumulative percent of variance accounted for by the first two components is about 93%. This provides a reasonable summary of the data. Thus we can keep the first two components and "throw away" the other components.

A number of other criteria have been proposed to select the number of components in PCA and factorial analysis. Users can read Lawley (1956), Horn (1965), Humphreys and Montanelli (1975), Horn and Engstrom (1979), Zwick and Velicer (1986), Hubbard and Allen (1987) and Jackson (1993), among others.

6. Interpretation of principal components

Running a PCA has become easy with statistical software. However, interpreting the results can be a difficult task. Here are a few guidelines that should help practitioners through the analysis.

6.1 The visual approach of correlation

Once the analysis is complete, we wish to assign a name to each retained component that describes its content. To do this, we need to know what variables explain the components. Correlations of the variables with the principal components are useful tools that can help interpreting the meaning of components. The correlations between each variable and each principal component are given in Table 4.

Variables	PCA 1	PCA 2
X1	0.943	-0.241
X2	0.939	-0.196
X3	0.902	0.064
X4	0.206	0.963
X5	0.159	0.975

Notes : PCA1 and PCA2 denote the first and second principal component, respectively.

Table 4. **Correlation variable-component**

Those correlations are also known as component loadings. A coefficient greater than 0.4 in absolute value is considered as significant (see, Stevens (1986) for a discussion). We can interpret PCA1 as being highly positively correlated with variables X_1, X_2 and X_3, and weakly positively correlated to variables X_4 and X_5. So X_1, X_2 and X_3 are the most important variables in the first principal component. PCA2, on the other hand, is highly positively correlated with X_4 and X_5, and weakly negatively related to X_1 and X_2. So X_4 and X_5 are most important in explaining the second principal component. Therefore, the name of the first component comes from variables X_1, X_2 and X_3 while that of the second component comes from X_4 and X_5.

It can be shown that the coordinate of a variable on a component is the correlation coefficient between that variable and the principal component. This allows us to plot the reduced dimension representation of variables in the plane constructed from the first two components. Variables highly correlated with a component show a small angle. Figure 2 represents this graph for our dataset. For each variable we have plotted on the horizontal dimension its loading on component 1, on the vertical dimension its loading on component 2.

The graph also presents a visual aspect of correlation patterns among variables. The cosine of the angle θ between two vectors x and y is computed as:

$$< x, y >= \|x\|\|y\|\cos(x, y) \tag{28}$$

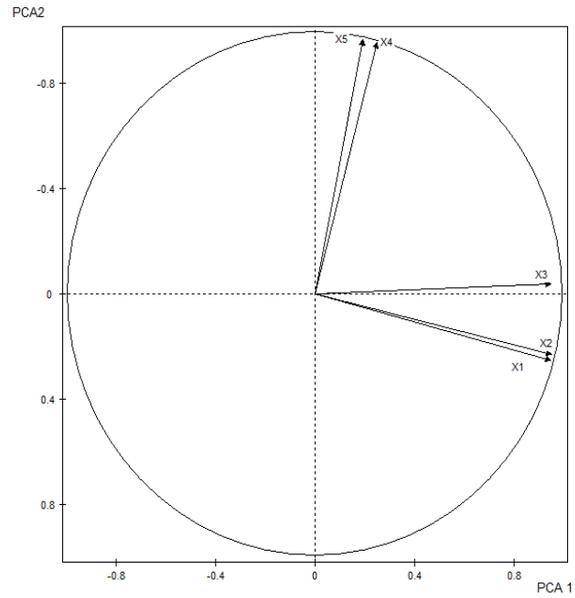

Fig. 2. **Circle of Correlation**

Replacing x and y with our transformed vectors yields:

$$\cos(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}} = \rho(x,y) \tag{29}$$

Eq.(29) shows the connection between the cosine measurement and the numerical measurement of correlation: the cosine of the angle between two variables is interpreted in terms of correlation. Variables highly positively correlated with each another show a small angle, while those are negatively correlated are directed in opposite sense, i.e. they form a flat angle. From Figure 2 we can see that the five variables hang together in two distinct groups. Variables X_1, X_2 and X_3 are positively correlated with each other, and form the first group. Variables X_4 and X_5 also correlate strongly with each other, and form the second group. Those two groups are weakly correlated. In fact, Figure 2 gives a reduced dimension representation of the correlation matrix given in Table 2.

It is extremely important, however, to notice that the angle between variables is interpreted in terms of correlation only when variables are well-represented, that is they are close to the border of the circle of correlation. Remember that the goal of PCA is to explain multiple variables by a lesser number of components, and keep in mind that graphs obtained from that reduction method are projections that optimize global criterion (i.e. the total variance). As such some relationships between variables may be greatly altered. Correlations between variables and components supply insights about variables that are not well-represented. In a subspace of components, the quality of representation of a variable is assessed by the sum-of-squared component loadings across components. This is called the communality of the

variable. It measures the proportion of the variance of a variable accounted for by the components. For example, in our example, the communality of the variable X_1 is $0.943^2+0.241^2=0.948$. This means that the first two components explain about 95% of the variance of the variable X_1. This is quite substantial to enable us fully interpreting the variability in this variable as well as its relationship with the other variables. Communality can be used as a measure of goodness-of-fit of the projection. The communalities of the 5 variables of our data are displayed in Table 5. As shown by this Table, the first two components explain more than 80% of variance in each variable. This is enough to reveal the structure of correlation among the variables. Do not interpret as correlation the angle between two variables when at least one of them has a low communality. Using communality prevent potential biases that may arise by directly interpreting numerical and graphical results yielded by the PCA.

Variables	Value
X1	0.948
X2	0.920
X3	0.817
X4	0.970
X5	0.976

Table 5. **Communalities of variables**

All these interesting results show that outcomes from normalized PCA can be easily interpreted without additional complicated calculations. From a visual inspection of the graph, we can see the groups of variables that are correlated, interpret the principal components and name them.

6.2 Factor scores and their use in multivariate models

A useful by product of PCA is factor scores. Factor scores are coordinates of subjects (individuals) on each component. They indicate where a subject stands on the retained component. Factor scores are computed as weighted values on the observed variables. Results for our dataset are reported in Table 6.

Factor scores can be used to plot a reduced representation of subjects. This is displayed by Figure 3.

How do we interpret the position of points on this diagram? Recall that this graph is a projection. As such some distances could be spurious. To distinguish wrong projections from real ones and better interpret the plot, we need to use that is called "the quality of representation" of subjects. This is computed as the squared of the cosine of the angle between a subject s_i and a component z, following the formula:

$$\cos^2(s_i, z_j) = \frac{z_i^2}{\|s_i\|^2} = \frac{z_i^2}{\sum_{k=1}^{p} x_{ki}^2} \tag{30}$$

ID	PCA1	PCA2	$Cos^2 1$	$Cos^2 2$	QL12= $Cos^2 1$+ $Cos^2 2$	CTR1	CTR2
1	1.701	-1.610	0.436	0.390	0.826	5.458	6.547
2	1.701	0.575	0.869	0.099	0.969	5.455	0.837
3	1.972	-0.686	0.862	0.104	0.966	7.333	1.191
4	3.000	2.581	0.563	0.417	0.981	16.974	16.832
5	2.382	-1.556	0.687	0.293	0.980	10.700	6.116
6	1.717	-0.323	0.522	0.018	0.541	5.558	0.264
7	0.193	-0.397	0.062	0.263	0.325	0.070	0.400
8	0.084	-1.213	0.004	0.972	0.977	0.013	3.718
9	1.071	-0.558	0.765	0.208	0.974	2.162	0.787
10	-0.427	-0.110	0.822	0.054	0.877	0.344	0.030
11	-1.088	0.176	0.093	0.024	0.933	2.232	0.078
12	-1.341	-1.673	0.344	0.536	0.881	3.393	7.075
13	-0.291	-0.996	0.071	0.835	0.906	0.160	2.507
14	-0.652	0.567	0.543	0.411	0.955	0.801	0.812
15	-0.062	3.166	0.000	0.957	0.957	0.007	25.325
16	-1.830	-0.375	0.929	0.039	0.968	6.318	0.356
17	-1.181	3.182	0.119	0.868	0.988	2.630	25.572
18	-2.244	-0.751	0.877	0.098	0.976	9.493	1.424
19	-2.288	-0.150	0.933	0.004	0.937	9.874	0.057
20	-2.417	0.155	0.938	0.003	0.942	11.019	0.060

Notes: Columns PCA1 and PCA2 display the factor scores on the first and second components, respectively. $Cos^2 1$ and $Cos^2 2$ indicate the quality of representation of subjects on the first and second components, respectively. QL12= $Cos^2 1$+ $Cos^2 2$ measures the quality of representation of subjects on the plane formed by the first two components. CTR1and CTR2 are the contribution of subjects on component 1 and component 2, respectively.

Table 6. **Factor Scores of Subjects, Contributions and Quality of Representation**

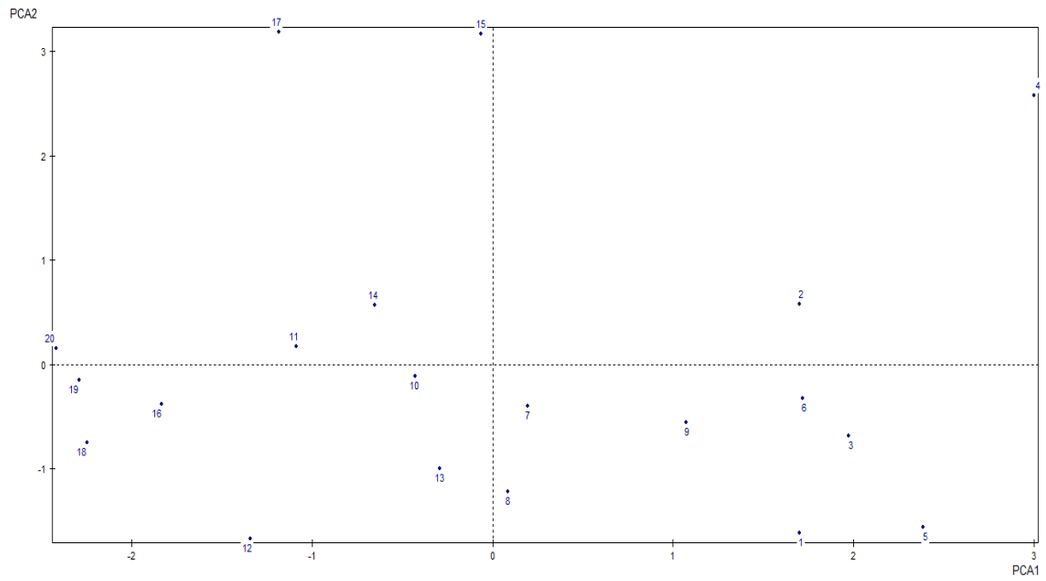

Fig. 3. **Scatterplot of subjects in the first two factors**

Cos^2 is interpreted as a measure of goodness-of-fit of the projection of a subject on a given component. Notice that in Eq. (30), $\|s_i\|^2$ is the distance of subject s_i from the origin. It measures how far the subject is from the center. So if cos^2=1 the component extracted is reproducing a great amount of the original behavior of the subject. Since the components are orthogonal, the quality of representation of a subject in a given subspace of components is the sum of the associated cos^2. This notion is similar to the concept of communality previously defined for variables.

In Table 6 we also reported these statistics. As can be seen, the two components retained explain more than 80% of the behavior of subjects, except for subjects 6 and 7. Now we are confident that almost all the subjects are well-represented, we can interpret the graph. Thus, we can tell that subjects located in the right side and having larger coordinates on the first component, i.e.1, 9, 6, 3 and 5, have values of X_1, X_2 and X_3 greater than the average. Those located in the left side and having smaller coordinates on the first axis, i.e. 20, 19, 18, 16, 12, 11 and 10, record lesser values for these variables. On the other hand, subjects 15 and 17 are characterized by highest values for variables X_4 and X_5, while subjects 8 and 13 record lowest values for these variables.

Very often a small number of subjects can determine the direction of principal components. This is because PCA uses the notions of mean, variance and correlation; and it is well known that these statistics are influenced by outliers or atypical observations in the data. To detect what are these atypical subjects we define the notion of "contribution" that measures how much a subject contributes to the variance of a component. Contributions (CTR) are computed following:

$$CTR(s_i, z_j) = \frac{z_i^2}{n\lambda_i} \times 100 \tag{31}$$

Contributions are reported in the last two columns of Table 6. Subject 4 contributes greatly to the first component with a contribution of 16.97%. This indicates that subject 4 explains alone 16.97% of the variance of the first component. Therefore, this subject takes higher values for X1, X2 and X3. This can be easily verified from the original Table 1. Regarding the second component, over 25% of the variance of the data accounted for by this component is explained by subjects 15 and 17. These subjects exhibit high values for variables X_4 and X_5.

The principal components obtained from PCA could be used in subsequent analyses (regressions, poverty analysis, classification…). For example, in linear regression models, the presence of correlated variables poses the econometric well-known problem of multicolinearity that makes instable regression coefficients. This problem is avoided when using the principal components that are orthogonal with one another. At the end of the analysis you can re-express the model with the original variables using the equations defining principal components. If there are variables that are not correlated with the other variables, you can delete them prior to the PCA, and reintroduce them in your model once the model is estimated.

7. A Case study with illustration using SPSS

We collected data on 10 socio-demographic variables for a sample of 132 countries. We use these data to illustrate how performing PCA using the SPSS software package. By following the indications provided here, user can try to reproduce himself the results obtained.

To perform a principal components analysis with SPSS, follow these steps:

1. Select **Analyze/Data Reduction/ Factor**
2. Highlight all of the quantitative variables and Click on the **Variables** button. The character variable Country is an identifier variable and should not be included in the Variables list.
3. Click on the **Descriptives** button to select **Univariate Descriptives, Initial Solution, KMO and Bartlett's test of Sphericity**.
4. Click on the **Extraction** button, and select **Method=Principal Components, Display Unrotated factor solution, Scree Plot**. Select **Extract Eigenvalue over 1** (by default).
5. Click on the **Rotation** button, and select **Display Loading Plot(s).**
6. Click on **Scores** and select **Save as variables**, **Method=Regression**. Select the case below.
7. Click on **Options**, and select **Exclude Cases Listwise** (option by default).

In what follows, we review and comment on the main outputs.

- **Correlation Matrix**

To discover the pattern of intercorrelations among variables, we examine the correlation matrix. That is given in Table 7:

	Life_ exp	Mortality	Urban	Iliteracy	Water	Telephone	Vehicles	Fertility	Hosp_ beds	Physici-ans
Life_exp	1.000	-0.956	0.732	-0.756	0.780	0.718	0.621	-0.870	0.514	0.702
Mortality		1.000	-0.736	0.809	-0.792	-0.706	-0.596	0.895	-0.559	-0.733
Urban			1.000	-0.648	0.692	0.697	0.599	-0.642	0.449	0.651
Iliteracy				1.000	-0.667	-0.628	-0.536	0.818	-0.603	-0.695
Water					1.000	0.702	0.633	-0.746	0.472	0.679
Telephone						1.000	0.886	-0.699	0.622	0.672
Vehicles							1.000	-0.602	0.567	0.614
Fertility								1.000	-0.636	-0.763
Hosp_beds									1.000	0.701
Physicians										1.000

Note : Figures reported in this table are correlation coefficients.

Table 7. **Correlation Matrix**

The variables can be grouped into two groups of correlated variables. We will see this later.

- **Testing for the Factorability of the Data**

Before applying PCA to the data, we need to test whether they are suitable for reduction. SPSS provides two tests to assist users:

Kaiser-Meyer-Olkin Measure of Sampling Adequacy (Kaiser, 1974): This measure varies between 0 and 1, and values closer to 1 are better. A value of 0.6 is a suggested minimum for good PCA.

Bartlett's Test of Sphericity (Bartlett, 1950): This tests the null hypothesis that the correlation matrix is an identity matrix in which all of the diagonal elements are 1 and all off diagonal elements are 0. We reject the null hypothesis when the level of significance exceeds 0.05.

The results reported in Table 8 suggest that the data may be grouped into smaller set of underlying factors.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.913
Bartlett's Test of Sphericity	Approx. Chi-Square	1407.151
	df	45
	Sig.	.000

Table 8. **Results of KMO and Bartlett's Test**

- **Eigenvalues and number of meaningful components**

Table 9 displays the eigenvalues, percent of variance and cumulative percent of variance from the observed data. Earlier it was stated that the number of components computed is equal to the number of variables being analyzed, necessitating that we decide how many components are truly meaningful and worthy of being retained for interpretation.

Here only component 1 demonstrates an eigenvalue greater than 1.00. So the Kaiser eigenvalue-one criterion would lead us to retain and interpret only this component. The first component provides a reasonable summary of the data, accounting for about 72% of the total variance of the 10 variables. Subsequent components each contribute less than 8%.

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.194	71.940	71.940	7.194	71.940	71.940
2	.780	7.801	79.741	.780	7.801	79.741
3	.667	6.675	86.416			
4	.365	3.654	90.070			
5	.302	3.022	93.092			
6	.236	2.361	95.453			
7	.216	2.162	97.615			
8	.106	1.065	98.680			
9	.095	.946	99.626			
10	.037	.374	100.000			

Table 9. **Eigenvalues**

The scree plot is displayed in Figure 4. From the second component on, we observe that the line is almost flat with a relatively large break following component 1. So the scree test would lead us to retain only the first component. The components appearing after the break (2-10) would be regarded as trivial (less than 10%).

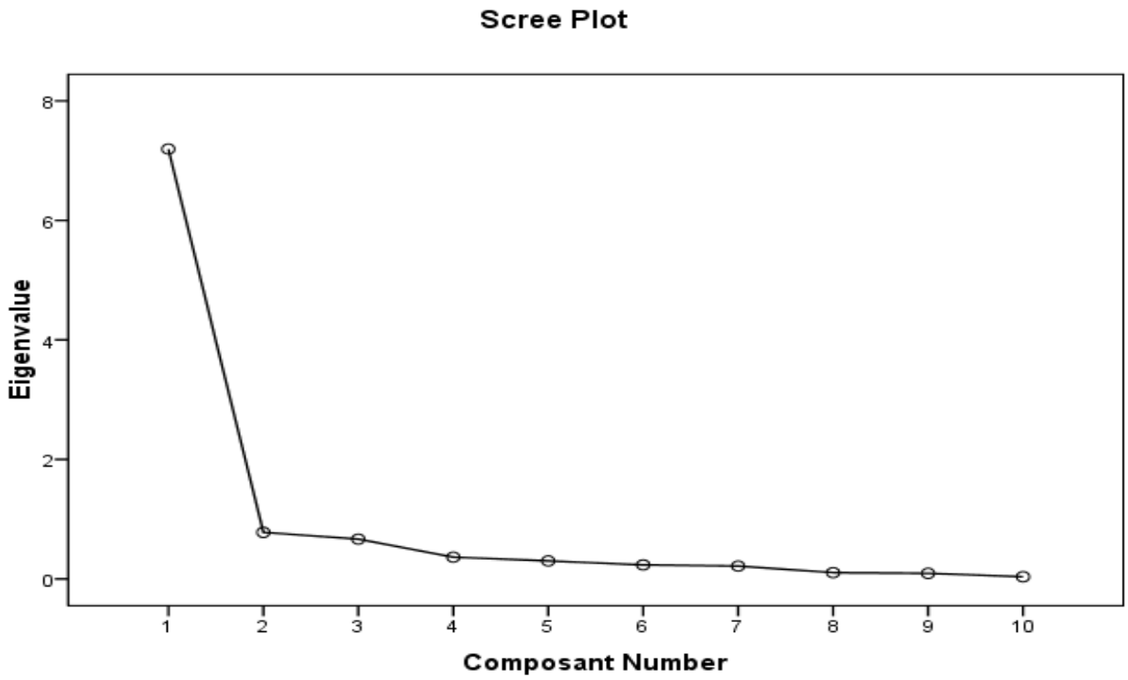

Fig. 4. **Scree Plot**

In conclusion, the dimensionality of the data could be reduced to 1. Nevertheless, we shall add the second component for representation purpose. Plot in a plane is easier to interpret than a three or 10-dimensional plot. Note that by default SPSS uses the Kaiser criterion to extract components. It belongs to the user to specify the number of components to be extracted if the Kaiser-criterion under-estimate the appropriate number. Here we specified 2 as the number of components to be extracted.

- **Component loadings**

Table 10 displays the loading matrix. The entries in this matrix are correlations between the variables and the components. As can be seen, all the variables load heavily on the first component. It is now necessary to turn to the content of the variables being analyzed in order to decide how this component should be named. What common construct do variables seem to be measuring?

In Figure 5 we observe two opposite groups of variables. The right-side variables are positively correlated one with another, and deal with social status of the countries. The left-side variables are also positively correlated one with another, and talk about another aspect of social life. It is therefore appropriate to name the first component the "social development" component.

	Component	
	1	2
Life_exp	.911	-.268
Mortality	-.926	.287
Urbanisation	.809	-.093
Iliteracy	-.848	.200
Water	.850	-.139
Telephone	.862	.355
Vehicles	.780	.483
Fertility	-.911	.183
Hosp_beds	.713	.396
Physicians	.850	.087

Table 10. **Component Matrix**

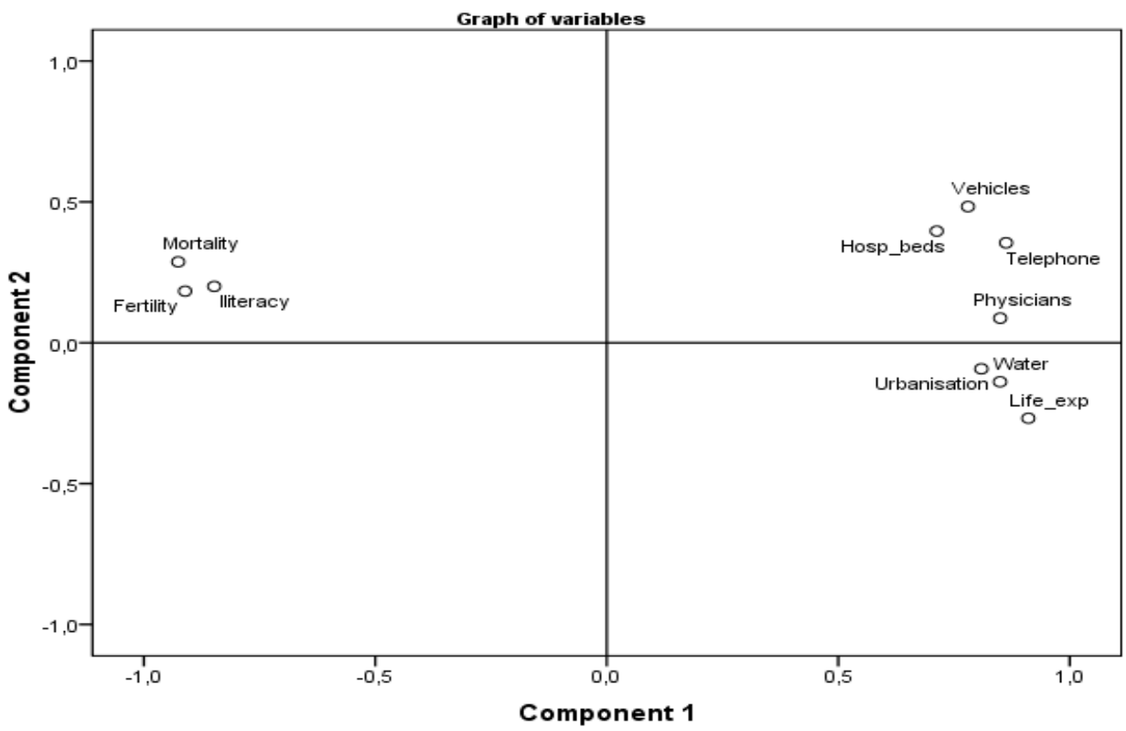

Fig. 5. **Scatterplot of variables**

- **Factor scores and Scatterplot of the Countries**

Since we have named the component, it is desirable to assign scores to each country to indicate where that country stands on the component. Here scores are indicating the level of social development of the countries. The values of the scores are to be interpreted paying attention to the signs of component loadings. From Figure 5 we say that countries with high

positive scores on the first component demonstrate higher level of social development relatively to countries with negative scores. In Figure 6 we can see that countries such as Burkina Faso, Niger, Sierra Leone, Tchad, Burundi, Centrafrique and Angola belong to the under-developed group.

SPSS does not provide directly the scatterplot for subjects. Since factor scores have been created and saved as variables, we can use the Graph menu to request a scatterplot. This is an easy task on SPSS. The character variable Country is used as an identifier variable. Notice that in SPSS factor scores are standardized with a mean zero and a standard deviation of 1.

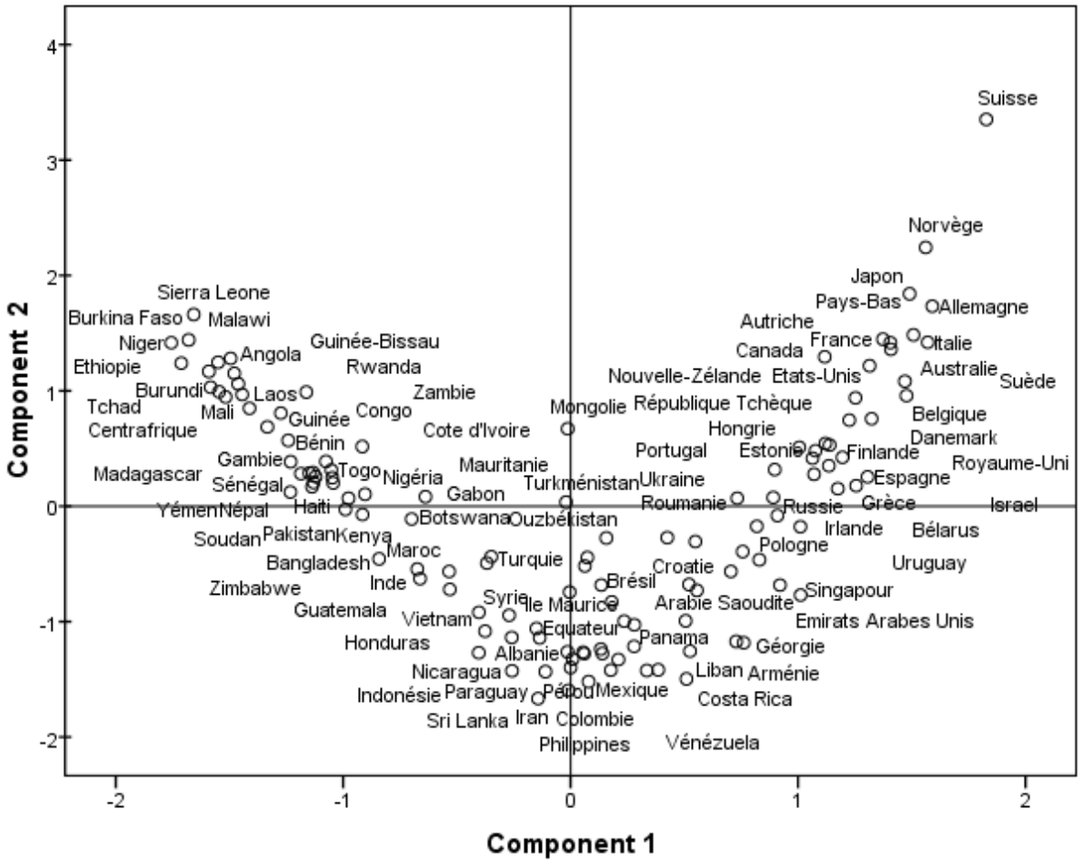

Fig. 6. **Scatterplot of the Countries**

A social development index is most useful to identify the groups of countries in connection with their level of development. The construction of this index assigns a social development-ranking score to each country. We rescale factor scores as follows:

$$SI_i = \frac{F_i - F_{\min}}{F_{\max} - F_{\min}} \times 100 \tag{32}$$

where F_{min} and F_{max} are the minimum and maximum values of the factor scores *F*. Using the rescaled-scores, countries are sorted in ascending. Lower scores identify socially under-developed countries, whereas higher scores identify socially developed countries.

8. Conclusion

Principal components analysis (PCA) is widely used in statistical multivariate data analysis. It is extremely useful when we expect variables to be correlated to each other and want to reduce them to a lesser number of factors. However, we encounter situations where variables are non linearly related to each other. In such cases, PCA would fail to reduce the dimension of the variables. On the other hand, PCA suffers from the fact each principal component is a linear combination of all the original variables and the loadings are typically nonzero. This makes it often difficult to interpret the derived components. Rotation techniques are commonly used to help practitioners to interpret principal components, but we do not recommend them.

Recently, other new methods of data analysis have been developed to generalize linear PCA. These include Sparse Principal Components Analysis (Tibshirani, 1996; Zou, Hastie & Tibshirani, 2006), Independent Component Analysis (Vasilescu & Terzopoulos, 2007), Kernel Principal Components Analysis (Schölkopf, Smola & Müller, 1997, 1998), and Multilinear Principal Components Analysis (Haiping, Plataniotis & Venetsanopoulos, 2008).

9. Appendix

9.1 Data for the case study

Pays	Life_exp	Mortality	Urban	Iliteracy	Water	Telephone	Vehicles	Fertility	Hosp_beds	Physicians
Albanie	72.00	25.00	40.00	16.50	76.00	31.00	27.00	2.5	3.2	1.4
Algérie	71.00	35.00	59.00	35.00	90.00	53.00	25.00	3.5	2.1	0.8
Angola	47.00	124.00	33.00	41.00	32.00	6.00	18.00	6.7	1.3	0
Argentine	73.00	19.00	89.00	3.00	65.00	203.00	137.00	2.6	3.3	2.7
Arménie	74.00	15.00	69.00	2.00	99.00	157.00	0	1.3	7.6	3
Australie	79.00	5.00	85.00	3.00	99.00	512.00	488.00	1.8	8.5	2.5
Autriche	78.00	5.00	65.00	2.00	100.00	491.00	481.00	1.3	9.2	2.8
Azerbeidjan	71.00	17.00	57.00	3.00	97.00	89.00	36.00	2	9.7	3.8
Bangladesh	59.00	73.00	23.00	60.00	84.00	3.00	1.00	3.1	0.3	0.2
Bélarus	68.00	11.00	71.00	0.5	100.00	241.00	2.00	1.3	12.2	4.3
Belgique	78.00	6.00	97.00	2.00	100.00	500.00	435.00	1.6	7.2	3.4
Bénin	53.00	87.00	41.00	61.50	50.00	7.00	7.00	5.7	0.2	0.1
Bolivie	62.00	60.00	61.00	15.50	55.00	69.00	32.00	4.1	1.7	1.3
Botswana	46.00	62.00	49.00	24.50	70.00	65.00	15.00	4.2	1.6	0.2
Brésil	67.00	33.00	80.00	16.00	72.00	121.00	88.00	2.3	3.1	1.3
Bulgarie	71.00	14.00	69.00	1.50	99.00	329.00	220.00	1.1	10.6	3.5
Burkina Faso	44.00	104.00	17.00	77.50	42.00	4.00	4.00	6.7	1.4	0
Burundi	42.00	118.00	8.00	54.00	52.00	3.00	2.00	6.2	0.7	0.1
Cambodge	54.00	102.00	15.00	61.50	13.00	2.00	5.00	4.5	2.1	0.1
Cameroun	54.00	77.00	47.00	26.50	41.00	5.00	7.00	5	2.6	0.1
Canada	79.00	5.00	77.00	35.00	99.00	634.00	455.00	1.6	4.2	2.1
Centrafrique	44.00	98.00	40.00	55.50	19.00	3.00	0	4.8	0.9	0.1
Tchad	48.00	99.00	23.00	60.00	24.00	1.00	3.00	6.4	0.7	0
Chili	75.00	10.00	85.00	4.50	85.00	205.00	71.00	2.2	2.7	1.1
Chine	70.00	31.00	31.00	17.00	90.00	70.00	3.00	1.9	2.9	2
Hong Kong	79.00	3.00	100.00	7.50	100.00	558.00	56.00	1.1		1.3
Colombie	70.00	23.00	73.00	9.00	78.00	173.00	21.00	2.7	1.5	1.1
Congo Démocratique	51.00	90.00	30.00	41.00	27.00	-	9.00	6.3	1.4	0.1
Congo	48.00	90.00	61.00	21.50	47.00	8.00	14.00	6	3.4	0.3
Costa Rica	77.00	13.00	47.00	5.00	92.00	172.00	85.00	2.6	1.9	1.4
Cote d'Ivoire	46.00	88.00	45.00	55.50	72.00	12.00	18.00	5	0.8	0.1
Croatie	73.00	8.00	57.00	2.00	63.00	348.00	17.00	1.5	5.9	2
République Tchèque	75.00	5.00	75.00	3.00	97.00	364.00	358.00	1.2	9.2	2.9

Pays	Life_exp	Mortality	Urban	Iliteracy	Water	Telephone	Vehicles	Fertility	Hosp_beds	Physicians
Danemark	76.00	5.00	85.00	1.00	100.00	660.00	355.00	1.8	4.7	2.9
Equateur	70.00	32.00	63.00	9.50	70.00	78.00	41.00	2.9	1.6	1.7
Egypte	67.00	49.00	45.00	46.50	64.00	60.00	23.00	3.2	2	2.1
El Salvador	69.00	31.00	46.00	22.00	55.00	80.00	30.00	3.3	1.6	1
Erythrée	51.00	61.00	18.00	48.00	7.00	7.00	1.00	5.7		0
Estonie	70.00	9.00	69.00	2.00	100.00	343.00	312.00	1.2	7.4	3.1
Ethiopie	43.00	107.00	17.00	64.00	27.00	3.00	1.00	6.4	0.2	0
Finlande	77.00	4.00	66.00	1.00	98.00	554.00	145.00	1.8	9.2	2.8
France	78.00	5.00	75.00	1.00	100.00	570.00	442.00	1.8	8.7	2.9
Gabon	53.00	86.00	79.00	29.00	67.00	33.00	14.00	5.1	3.2	0.2
Gambie	53.00	76.00	31.00	65.50	76.00	21.00	8.00	5.6	0.6	0
Géorgie	73.00	15.00	60.00	4.00	100.00	115.00	80.00	1.3	4.8	3.8
Allemagne	77.00	5.00	87.00	1.00	100.00	567.00	506.00	1.4	9.6	3.4
Ghana	60.00	65.00	37.00	31.00	56.00	8.00	5.00	4.8	1.5	
Grèce	78.00	6.00	60.00	3.50	100.00	522.00	238.00	1.3	5	3.9
Guatemala	64.00	42.00	39.00	32.50	67.00	41.00	12.00	4.4	1	0.9
Guinée	47.00	118.00	31.00	41.00	62.00	5.00	2.00	5.4	0.6	0.2
Guinée-Bissau	44.00	128.00	23.00	63.00	53.00	7.00	6.00	5.6	1.5	0.2
Haiti	54.00	71.00	34.00	52.00	28.00	8.00	4.00	4.3	0.7	0.2
Honduras	69.00	36.00	51.00	27.00	65.00	38.00	7.00	4.2	1.1	0.8
Hongrie	71.00	10.00	64.00	1.00	99.00	336.00	233.00	1.3	9.1	3.4
Inde	63.00	70.00	28.00	45.00	81.00	22.00	5.00	3.2	0.8	0.4
Indonésie	65.00	43.00	39.00	14.50	62.00	27.00	12.00	2.7	0.7	0.2
Iran	71.00	26.00	61.00	25.50	83.00	112.00	26.00	2.7	1.6	0.9
Irlande	76.00	6.00	59.00	1.00	100.00	435.00	279.00	1.9	3.7	2.1
Israel	78.00	6.00	91.00	4.00	99.00	471.00	215.00	2.7	6	4.6
Italie	78.00	5.00	67.00	1.50	100.00	451.00	539.00	1.2	6.5	5.5
Jamaique	75.00	21.00	55.00	14.00	70.00	166.00	40.00	2.6	2.1	1.3
Japon	81.00	4.00	79.00	2.00	96.00	503.00	394.00	1.4	16.2	1.8
Jordanie	71.00	27.00	73.00	11.50	89.00	86.00	48.00	4.1	1.8	1.7
Kazakhstan	65.00	22.00	56.00	7.00	93.00	104.00	62.00	2	8.5	3.5
Kenya	51.00	76.00	31.00	19.50	53.00	9.00	11.00	4.6	1.6	0
Corée du Sud	73.00	9.00	80.00	2.50	83.00	433.00	163.00	1.6	4.6	1.1
Koweit	77.00	12.00	97.00	19.50	100.00	236.00	359.00	2.8	2.8	1.9
Laos	54.00	96.00	22.00	54.00	39.00	6.00	3.00	5.5	2.6	0.2
Liban	70.00	27.00	89.00	15.00	100.00	194.00	21.00	2.4	2.7	2.8
Lesotho	55.00	93.00	26.00	18.00	52.00	10.00	6.00	4.6		0.1
Lituanie	72.00	9.00	68.00	0.5	97.00	300.00	265.00	1.4	9.6	3.9
Madagascar	58.00	92.00	28.00	35.00	29.00	3.00	4.00	5.7	0.9	0.3
Malawi	42.00	134.00	22.00	41.50	45.00	3.00	2.00	6.4	1.3	0
Malaisie	72.00	8.00	56.00	13.50	100.00	198.00	145.00	3.1	2	0.5
Mali	50.00	117.00	29.00	61.50	37.00	3.00	3.00	6.5	0.2	0.1
Mauritanie	54.00	90.00	55.00	58.50	64.00	6.00	8.00	5.4	0.7	0.1
Ile Maurice	71.00	19.00	41.00	16.50	98.00	214.00	71.00	2	3.1	0.9
Mexique	72.00	30.00	74.00	9.00	83.00	104.00	97.00	2.8	1.2	1.2
Mongolie	66.00	50.00	62.00	38.50	45.00	37.00	16.00	2.5	11.5	2.6
Maroc	67.00	49.00	55.00	53.00	52.00	54.00	38.00	3	1	0.5
Mozambique	45.00	134.00	38.00	57.50	32.00	4.00	0	5.2	0.9	
Namibie	54.00	67.00	30.00	19.00	57.00	69.00	46.00	4.8		0.2
Népal	58.00	77.00	11.00	60.50	44.00	8.00	0	4.4	0.2	0
Pays-Bas	78.00	5.00	89.00	1.00	100.00	593.00	391.00	1.6	11.3	2.6
Nouvelle-Zélande	77.00	5.00	86.00	2.00	97.00	479.00	470.00	1.9	6.1	2.1
Nicaragua	68.00	36.00	55.00	30.50	81.00	31.00	18.00	3.7	1.5	0.8
Niger	46.00	118.00	20.00	85.50	53.00	2.00	4.00	7.3	0.1	0
Nigéria	53.00	76.00	42.00	39.00	39.00	4.00	9.00	5.3	1.7	0.2
Norvège	78.00	4.00	75.00	1.00	100.00	660.00	402.00	1.8	15	2.5
Oman	73.00	18.00	81.00	32.50	68.00	92.00	103.00	4.6	2.2	1.3
Pakistan	62.00	91.00	36.00	56.50	60.00	19.00	5.00	4.9	0.7	0.6
Panama	74.00	21.00	56.00	8.50	84.00	151.00	79.00	2.6	2.2	1.7

Pays	Life_exp	Mortality	Urban	Iliteracy	Water	Telephone	Vehicles	Fertility	Hosp_beds	Physicians
Paraguay	70.00	24.00	55.00	7.50	70.00	55.00	14.00	3.9	1.3	1.1
Pérou	69.00	40.00	72.00	11.00	80.00	67.00	26.00	3.1	1.5	0.9
Philippines	69.00	32.00	57.00	5.00	83.00	37.00	10.00	3.6	1.1	0.1
Pologne	73.00	10.00	65.00	0	98.00	228.00	230.00	1.4	5.4	2.3
Portugal	75.00	8.00	61.00	8.50	82.00	413.00	309.00	1.5	4.1	3
Roumanie	69.00	21.00	56.00	2.00	62.00	162.00	116.00	1.3	7.6	1.8
Russie	67.00	17.00	77.00	0.5	95.00	197.00	120.00	1.2	12.1	4.6
Rwanda	41.00	123.00	6.00	31.00	56.00	2.00	1.00	6.1	1.7	0
Arabie Saoudite	72.00	20.00	85.00	26.50	93.00	143.00	98.00	5.7	2.3	1.7
Sénégal	52.00	69.00	46.00	64.50	50.00	16.00	10.00	5.5	0.4	0.1
Sierra Leone	37.00	169.00	35.00	39.00	34.00	4.00	5.00	6	1.2	0.1
Singapour	77.00	4.00	100.00	8.00	100.00	562.00	108.00	1.5	3.6	1.4
Slovaquie	73.00	9.00	57.00	3.00	92.00	286.00	222.00	1.4	7.5	3
Slovénie	75.00	5.00	50.00	0	98.00	375.00	403.00	1.2	5.7	2.1
Afrique du Sud	63.00	51.00	53.00	15.50	70.00	115.00	85.00	2.8		0.6
Espagne	78.00	5.00	77.00	3.00	100.00	414.00	385.00	1.2	3.9	4.2
Sri Lanka	73.00	16.00	23.00	9.00	46.00	28.00	15.00	2.1	2.7	0.2
Soudan	55.00	69.00	34.00	44.50	50.00	6.00	9.00	4.6	1.1	0.1
Suède	79.00	4.00	83.00	1.00	100.00	674.00	428.00	1.5	5.6	3.1
Suisse	79.00	4.00	68.00	1.00	100.00	675.00	477.00	1.5	20.8	3.2
Syrie	69.00	28.00	54.00	32.50	85.00	95.00	9.00	3.9	1.5	1.4
Tadjikistan	69.00	23.00	28.00	1.00	69.00	37.00	0	3.4	8.8	2.1
Tanzanie	47.00	85.00	31.00	26.50	49.00	4.00	1.00	5.4	0.9	0
Thailande	72.00	29.00	21.00	5.00	94.20	84.00	27.00	1.9	2	0.4
Togo	49.00	78.00	32.00	45.00	63.00	7.00	19.00	5.1	1.5	0.1
Tunisie	72.00	28.00	64.00	31.50	99.00	81.00	30.00	2.2	1.7	0.7
Turquie	69.00	38.00	73.00	16.00	49.00	254.00	64.00	2.4	2.5	1.1
Turkménistan	66.00	33.00	45.00	8.00	60.00	82.00	1.00	2.9	11.5	0.2
Ouganda	42.00	101.00	14.00	35.00	34.00	3.00	2.00	6.5	0.9	0
Ukraine	67.00	14.00	68.00	0.5	55.00	191.00	0	1.3	11.8	4.5
Emirats Arabes Unis	75.00	8.00	85.00	25.00	98.00	389.00	11.00	3.4	2.6	1.8
Royaume-Uni	77.00	6.00	89.00	1.00	100.00	557.00	375.00	1.7	4.5	1.6
Etats-Unis	77.00	7.00	77.00	2.00	100.00	661.00	483.00	2	4	2.6
Uruguay	74.00	16.00	91.00	2.50	99.00	250.00	154.00	2.4	4.4	3.7
Ouzbékistan	69.00	22.00	38.00	12.00	57.00	65.00	0	2.8	8.3	3.3
Vénézuela	73.00	21.00	86.00	8.00	79.00	117.00	69.00	2.9	1.5	2.4
Vietnam	68.00	34.00	20.00	7.00	36.00	26.00	1.00	2.3	3.8	0.4
Yémen	56.00	82.00	24.00	56.00	74.00	13.00	14.00	6.3	0.7	0.2
Zambie	43.00	114.00	39.00	23.50	43.00	9.00	15.00	5.5	3.5	0.1
Zimbabwe	51.00	73.00	34.00	12.50	77.00	17.00	28.00	3.7	0.5	0.1

10. References

Bartlett M.S. (1950). Tests of Significance in Factor Analysis. *The British Journal of Psychology*, 3: 77-85.

Cadima J. & Jolliffe I. T. (1995). Loadings and Correlations in the Interpretation of Principal Components. *Journal of Applied Statistics*, 22(2):203–214.

Cattell R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research, 1,*245-276.

Cooley W.W. & Lohnes P.R. (1971). *Multivariate Data Analysis*. John Wiley & Sons, Inc., New York.

Haiping L., Plataniotis K.N. & Venetsanopoulos A.N. (2008). MPCA: Multilinear Principal Component Analysis of Tensor Objects, *Neural Networks, IEEE Transactions on*, 19(1): 18 – 39.

Hotelling H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6):417–441.

Horn J. L. (1965). A Rationale and Test for the Number Factors in Factor Analysis. *Psychometrika*, 30, 179-185.

Horn J. L. & Engstrom R. (1979). Cattell's Scree Test in Relation to Bartlett's Chi-Square Test and other Observations on the Number of Factors Problem. *Multivariate Behavioral Research*, 14, 283-300.

Hubbard R. & Allen S. J. (1987). An Empirical Comparison of Alternative Methods for Principal Component Extraction. *Journal of Business Research, 15,* 173-190.

Humphreys L. G. & Montanelli R. G. (1975). An Investigation of the Parallel Analysis Criterion for Determining the Number of Common Factors. *Multivariate Behavioral Research*, *10*, 193-206.

Hyvarinen A., Karhunen J. & Oja E. (2001). *Independent Component Analysis*. New York: Wiley.

Jackson D.A. (1993). Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology* 74(8): 2204-2214.

Jolliffe I. T. (2002). *Principal Component Analysis*. Second ed. New York: Springer-Verlag

Jolliffe I.T., Trendafilov N.T. & Uddin M. (2003). A Modified Principal Component Technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

Kaiser H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement,* 20:141-151.

Kaiser H. F. (1974). An Index of Factorial Simplicity. *Psychometrika*, 39:31-36.

Lambert Z. V., Wildt A.R. & Durand R.M. (1990). Assessing Sampling Variation Relative to number-of-factor Criteria. *Educational and Psychological Measurement,* 50:33-49.

Lawley D.N. (1956). Tests of Significance for the Latent Roots of Covariance and Correlation Matrices. *Biometrika*, 43: 128-136.

Lebart L., Morineau A. & Piron M. (1995). *Statistique Exploratoire Multidimensionnelle*. Paris : Dunod.

Pearson K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, Series 6, 2(11), 559–572.

Stevens J. (1986). *Applied Multivariate Statistics for the Social Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, series B 58(267-288).

Schölkopf B., Smola A.J. & Müller K.-R. (1997). Kernel Principal Component Analysis. *Lecture Notes in Computer Science*, Vol.1327, Artificial Neural Networks — ICANN'97, pp.583-588.

Schölkopf B., Smola A.J. & Müller K.-R. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319.

Vasilescu M. A. O. & Terzopoulos D. (2007). Multilinear (Tensor) ICA and Dimensionality Reduction. *Lecture Notes in Computer Science*, Vol. 4666, Independent Component Analysis and Signal Separation, pp. 818-826.

Zou H., Hastie T. & Tibshirani R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, 15(2): 265-286.

Zwick W.R. & Velicer W.F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin,* 99, 432-442.

Principal Component Analysis

Edited by Dr. Parinya Sanguansat

ISBN 978-953-51-0195-6
Hard cover, 300 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

This book is aimed at raising awareness of researchers, scientists and engineers on the benefits of Principal Component Analysis (PCA) in data analysis. In this book, the reader will find the applications of PCA in fields such as image processing, biometric, face recognition and speech processing. It also includes the core concepts and the state-of-the-art methods in data analysis and feature extraction.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen