Descriptive Statistics vs. Factor Analysis

Descriptive statistics will inform on the prevalence of a phenomenon, among a given population, captured by specific indicators.

This provides a valuable insight but it is not sufficient to capture existing underlying phenomena nor to unfold the problem in a comprehensive manner.

Factor Analysis

Many statistical methods are used to study the relation between independent and dependent variables.

There is a domain of statistics that deals with this type of data analysis which is called **factor** or **multivariate analysis**.

Factor Analysis

The purpose of **factor analysis** is to discover simple patterns in the network of relationships among variables. In particular, it seeks to discover if the observed variables can be explained largely or entirely in terms of a much smaller number of variables called factors.

Factor Analysis

The multivariate statistical analysis provides several techniques to analyze continuous and categorical variables, to capture the essence of their relationship and build new indicators (i.e. factors or principal components) conveying this relationship.

Principal Component Analysis

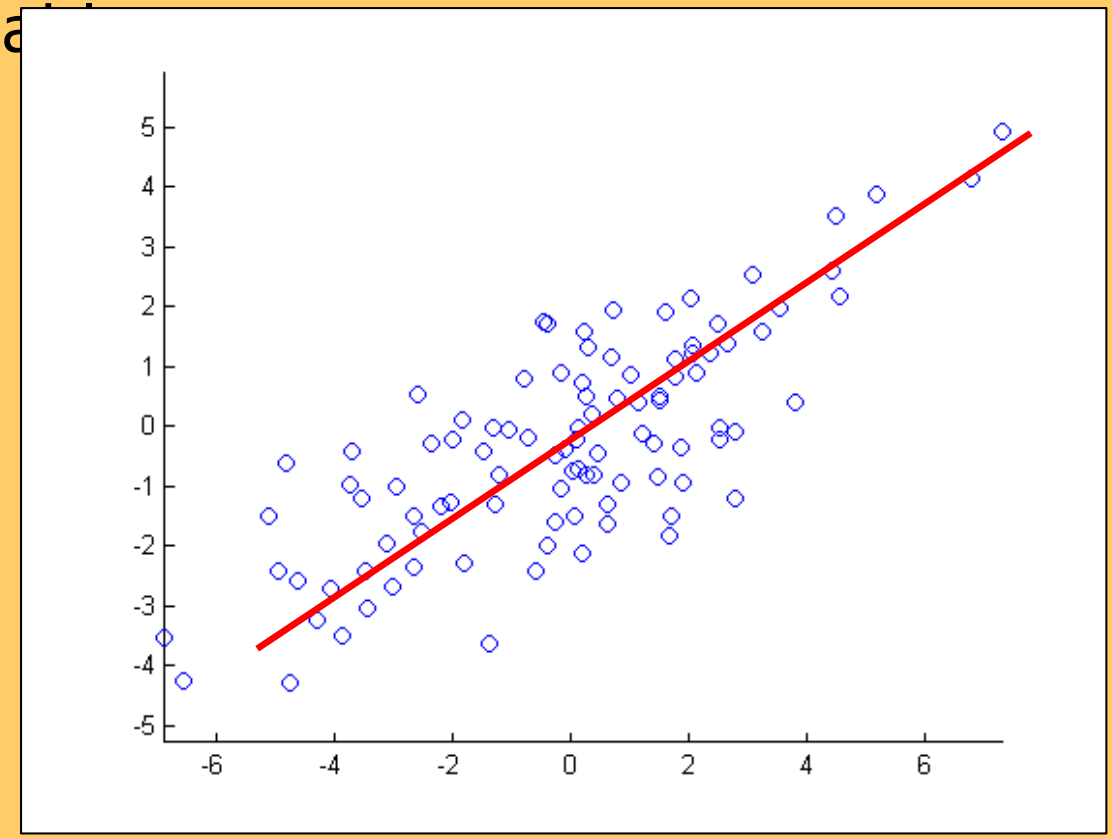
The objectives of a PCA are:

- To discover or reduce the dimensionality of the data set;
- To identify new meaningful underlying variables.

Principal Component Analysis

Assume to have two correlated variables on a scatter-plot.

A regression line can be fitted to represent the linear relationship between the two vari



Principal Component Analysis

If we could define a variable that would approximate the regression line on the plot, then the new variable would capture most of the essence of the two variables.

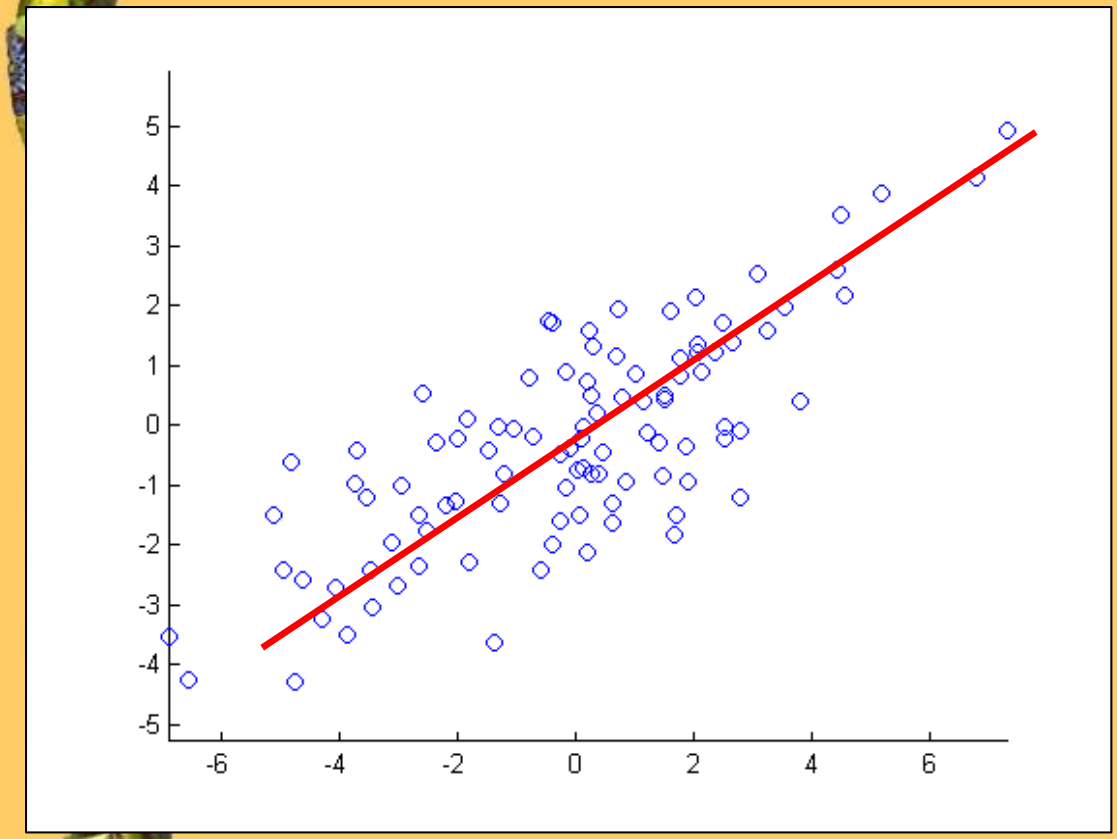
Individual single scores on the new factor can be used to represent the essence of the two variables.

The new factor is thus a linear combination of the two variables.

We have reduced the two variables to one factor.

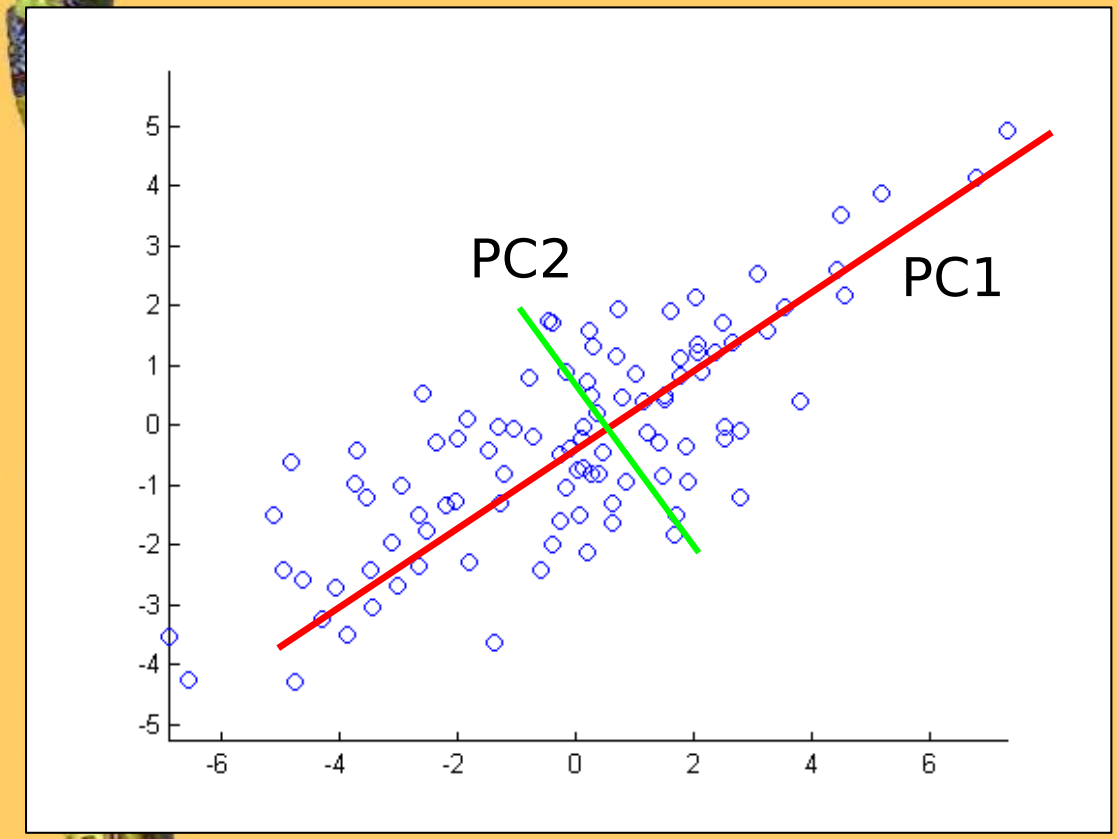
The Principal Components are calculated from the correlation matrix.

Principal Component Analysis



Graphically, the first principal component lies along the line of greatest variation and it is as close to all of the data as possible (red line).

Principal Component Analysis



The second PCA axis also must be completely uncorrelated i.e. at right angles, or "orthogonal" to PCA axis 1 (green line).

Principal Component Analysis

In a typical PCA however, there are more than two variables i.e more than two dimensions.

If the second principal component will be both perpendicular to the first, and along the line of second next greatest variation. The third principal component will be along the line of the following greatest variation and perpendicular to the first two principal components.

The same applies to the N dimensions under analysis. With several variables the computation is more complicated but the basic principle to express two or more variables by a single factor remains the same.

Principal Component Analysis

New factors, e.g. principal components are created by rotating the data plotted on orthogonal axes.

So doing, PCA helps to determine whether there is/are hidden factors/components along which the data vary.

It computes a compact and optimal description of the data set.

Principal Component Analysis

Basically a PCA transforms a set of more or less correlated variables into a set of uncorrelated variables which are ordered by reducing variability.

The uncorrelated variables are linear combinations of the original variables and the **last of these variables can be removed with minimum loss of real data.**

Principal Component Analysis

If we have more than three initial variables, how do we determine how many axes we worth interpreting.

This is left to the analyst, however the eigenvalues calculated by the PCA give us a big hint.

Principal Component Analysis

Every axis has an eigenvalue associated with it, that is the variance extracted by the factor. They are ranked from the highest to the lowest and their level is related to the amount of variation explained by the axis.

The sum of the eigenvalues is the number of variables.

Usually, the eigenvalues are expressed as a percentage of the total.

Principal Component Analysis

Example:

PCA Axis 1: eigenvalue 63%

PCA Axis 2: eigenvalue 33%

PCA Axis 3: eigenvalue 4%

In this example the first PCA axis explains 63% (about 2/3) of the variation of the entire data set and the second axis almost all the remaining variation. Axis 3 explain a trivial amount and can be dropped.