

# Running a Principal Component Analysis (PCA) with XLSTAT

[demoPCA.xls](#)

## Dataset for running a Principal Component Analysis

An Excel sheet containing both the data and the results for use in this tutorial can be downloaded by clicking [here](#).

The data are from the US Census Bureau and describe the changes in the population of 51 states between 2000 and 2001. The initial dataset has been transformed to rates per 1000 inhabitants, with the data for 2001 serving as the focus for the analysis.

## Goal of this Principal Component Analysis

Our goal is to analyze the correlations between the variables and to find out if the changes in population in some states are very different from the ones in other states. This example is also used in our Hierarchical Clustering tutorial.

## Principal Component Analysis

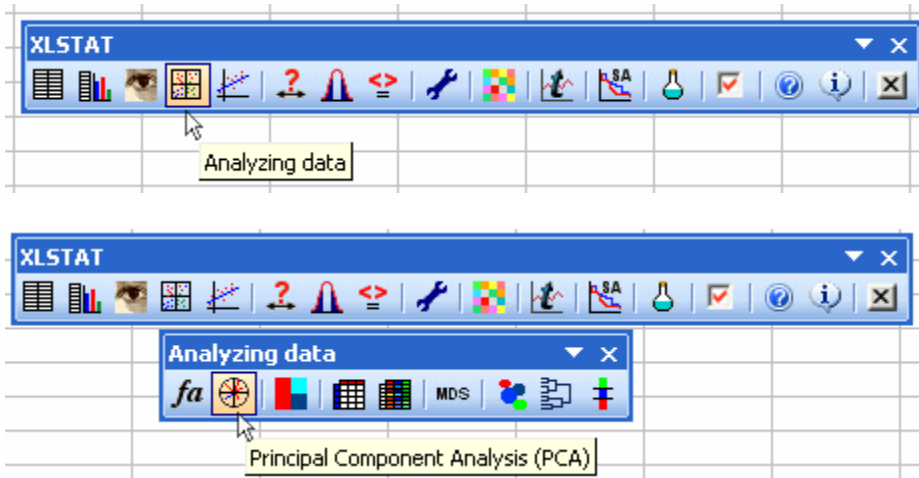
Principal Component Analysis is a very useful method to analyze numerical data structured in a M observations / N variables table. It allows to:

- Quickly visualize and analyze correlations between the N variables,
- Visualize and analyze the M observations (initially described by the N variables) on a low dimensional map, the optimal view for a variability criterion,
- Build a set of P uncorrelated factors ( $P \leq N$ ) that can be reused as input for other statistical methods (such as regression).

The limits of Principal Component Analysis stem from the fact that it is a projection method, and sometimes the visualization can lead to false interpretations. There are however some tricks to avoid these pitfalls.

## Setting up a Principal Component Analysis

Once XLSTAT-Pro is activated, select the **XLSTAT / Analyzing data / Principal components analysis** command, or click on the corresponding button of the **Analyzing Data** toolbar (see below).



The Principal Component Analysis dialog box will appear.

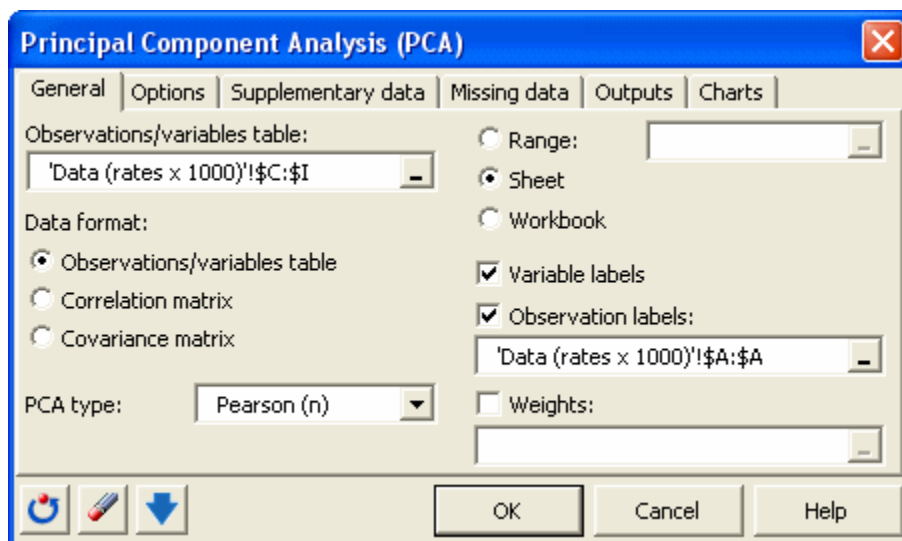
Select the data on the Excel sheet.

*Note: There are several ways of selecting data with XLSTAT - for further information, please check the section on selecting data.*

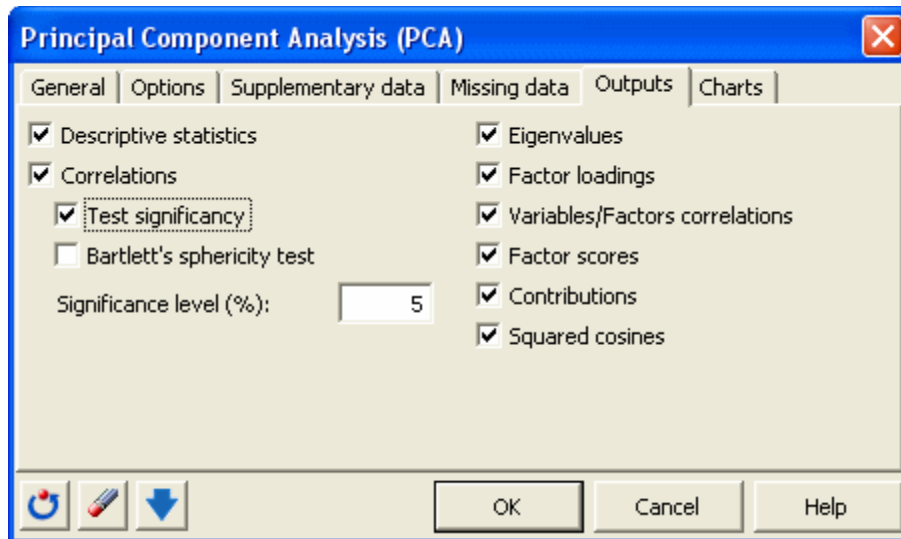
In this example, the data start from the first row, so it is quicker and easier to use columns selection. This explains why the letters corresponding to the columns are displayed in the selection boxes.

The **Data format** chosen is **Observations/variables** because of the format of the input data.

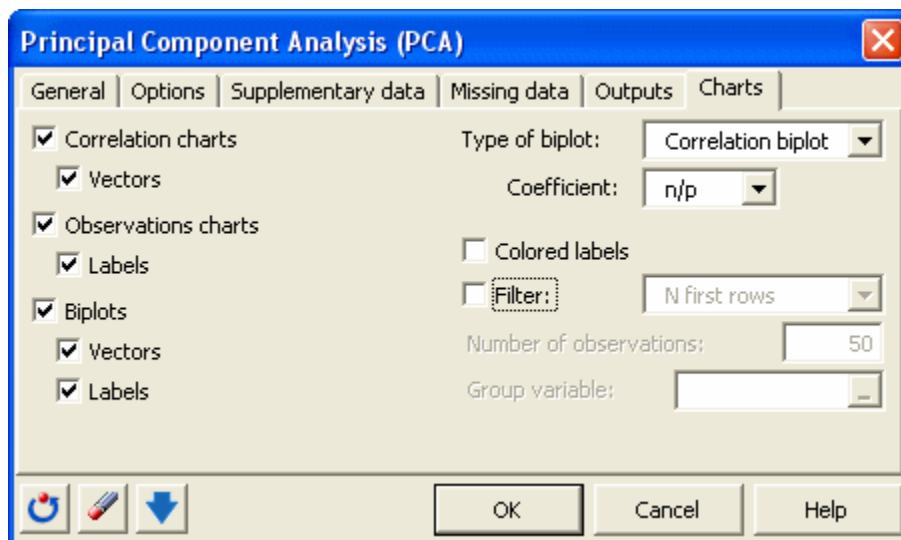
The **PCA type** that will be used during the computations is the Pearson's correlation matrix, which corresponds to the classical correlation coefficient.



In the **Outputs** tab, we choose to activate the option to display significant correlations in bold characters ("Test significance").



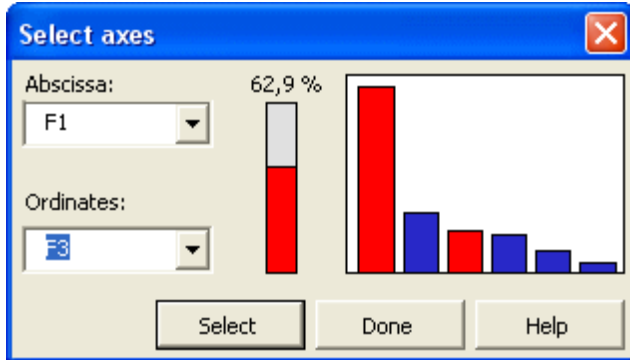
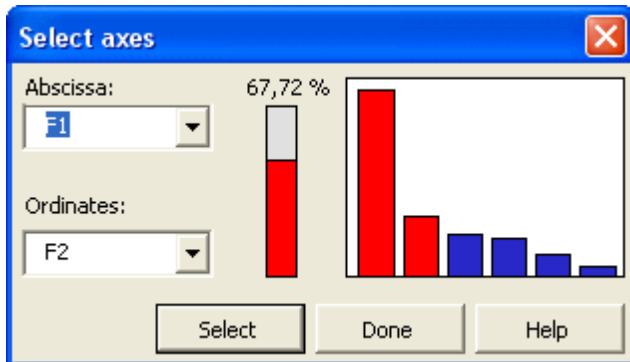
In the **Charts** tab, in order to display the labels on all charts, and to display all the observations (observations charts and biplots), the filtering option is unchecked. If there is a lot of data, displaying the labels might slow down the global display of the results. Displaying all the observations might make the of the results unreadable. In these cases, filtering the observations to display is recommended.



The computations begin once you have clicked on **OK**. You are asked to confirm the number of rows and columns.

*Note: This message can be bypassed by un-selecting the "Ask for selections confirmation" in the XLSTAT options panel.*

Then you should confirm the axes for which you want to display plots. In this example, the percentage of variability represented by the first two factors is not very high (67.72%); to avoid a misinterpretation of the results, we have decided to complement the results with a second chart on axes 1 and 3.



## Interpreting the results of a Principal Component Analysis

The first result to look at is the correlation matrix. We can see right away that the rates of people below and above 65 are negatively correlated ( $r = -1$ ). Either of the two variables could have been removed without effect on the quality of the results. We can also see that the Net Domestic Migration has low correlation with the other variables, including the Net International migration. This means that U.S. nationals and non-nationals may be moving to a state for different sets of reasons.

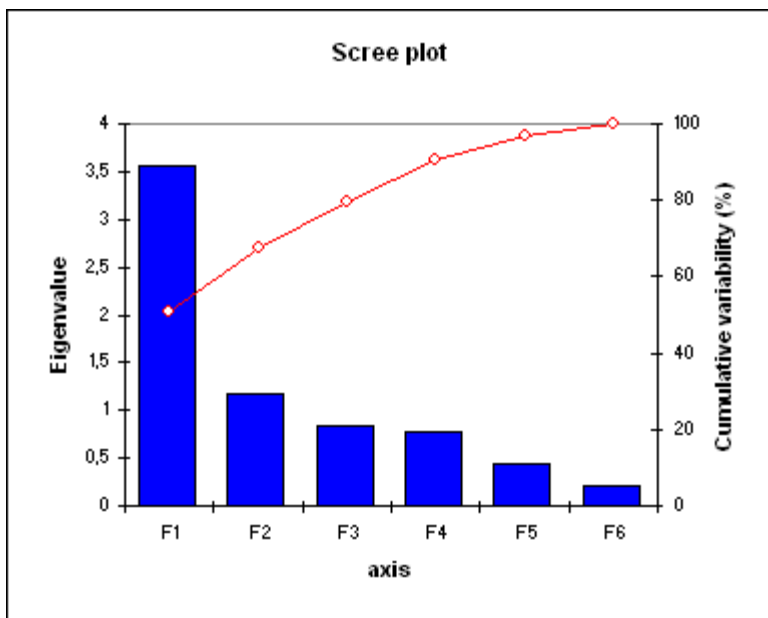
Correlation matrix:							
	Net Domestic	Net Int. Migrat	Period Birth	Period Death	<65 Pop. Es	>65 Pop. Est.	
Net Domestic	1	0.020	0.206	-0.060	-0.232	0.095	-0.095
Federal/Civ	0.020	1	-0.133	<b>-0.308</b>	<b>0.422</b>	<b>-0.377</b>	<b>0.377</b>
Net Int. Mi	0.206	-0.133	1	<b>0.295</b>	<b>-0.412</b>	0.204	-0.204
Period Birt	-0.060	<b>-0.308</b>	<b>0.295</b>	1	<b>-0.506</b>	<b>0.640</b>	<b>-0.640</b>
Period Dea	-0.232	<b>0.422</b>	<b>-0.412</b>	<b>-0.506</b>	1	<b>-0.779</b>	<b>0.779</b>
< 65 Pop.	0.095	<b>-0.377</b>	0.204	<b>0.640</b>	<b>-0.779</b>	1	<b>-1.000</b>
> 65 Pop.	-0.095	<b>0.377</b>	-0.204	<b>-0.640</b>	<b>0.779</b>	<b>-1.000</b>	1

*In bold, significant values (except diagonal) at the level of significance alpha=0.050 (Two-tailed test)*

The next table and the corresponding chart are related to a mathematical object, the eigenvalues, which reflect the quality of the projection from the N-dimensional initial table (N=7 in this example) to a lower number of dimensions. In this example, we can see that the first eigenvalue equals 3.567 and represents 51% of the total variability. This means that if we represent the data on only one axis, we will still be able to see % of the total variability of the data.

Each eigenvalue corresponds to a factor, and each factor to a one dimension. A factor is a linear combination of the initial variables, and all the factors are un-correlated ( $r=0$ ). The eigenvalues and the corresponding factors are sorted by descending order of how much of the initial variability they represent (converted to %).

Eigenvalues:						
	F1	F2	F3	F4	F5	F6
Eigenvalue	3.567	1.173	0.835	0.776	0.444	0.204
variance %	50.964	16.756	11.932	11.091	6.342	2.914
cumulated	50.964	67.720	79.652	90.744	97.086	100.000



Ideally, the first two or three eigenvalues will correspond to a high % of the variance, ensuring us that the maps based on the first two or three factors are a good quality projection of the initial multi-dimensional table. In this example, the first two factors allow us to represent 67.72% of the initial variability of the data. This is a good result, but we'll have to be careful when we interpret the maps as some information might be hidden in the next factors. We can see here that although we initially had 7 variables, the number of factors is 6. This is due to the two age variables, which are negatively correlated (-1). The number of "useful" dimensions has been automatically detected.

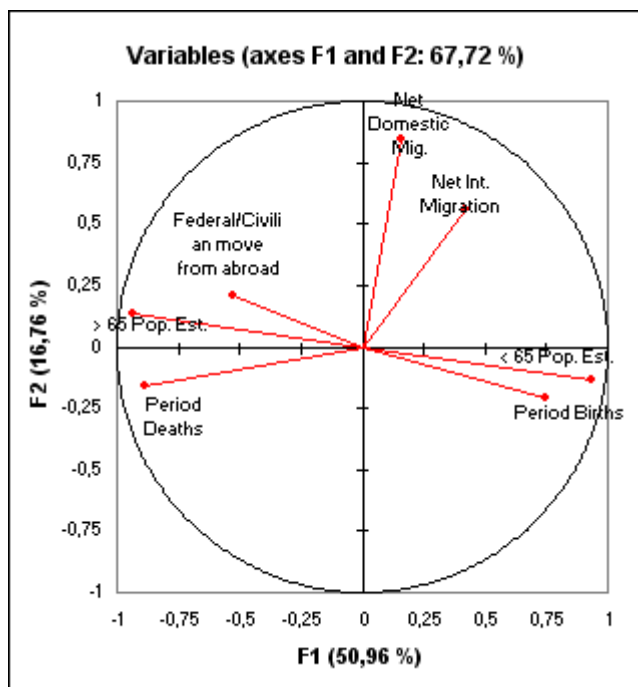
The first map is called the correlation circle (below on axes F1 and F2). It shows a projection of the initial variables in the factors space. When two variables are far from the center, then, if they are:

Close to each other, they are significantly positively correlated ( $r$  close to 1);

If they are orthogonal, they are not correlated ( $r$  close to 0);

If they are on the opposite side of the center, then they are significantly negatively correlated ( $r$  close to -1).

When the variables are close to the center, it means that some information is carried on other axes, and that any interpretation might be hazardous. For example, we might be tempted to interpret a correlation between the variables Net Domestic migration and Net International Migration although, in fact, there is none. This can be confirmed either by looking at the correlation matrix or by looking at the correlation circle on axes F1 and F3.

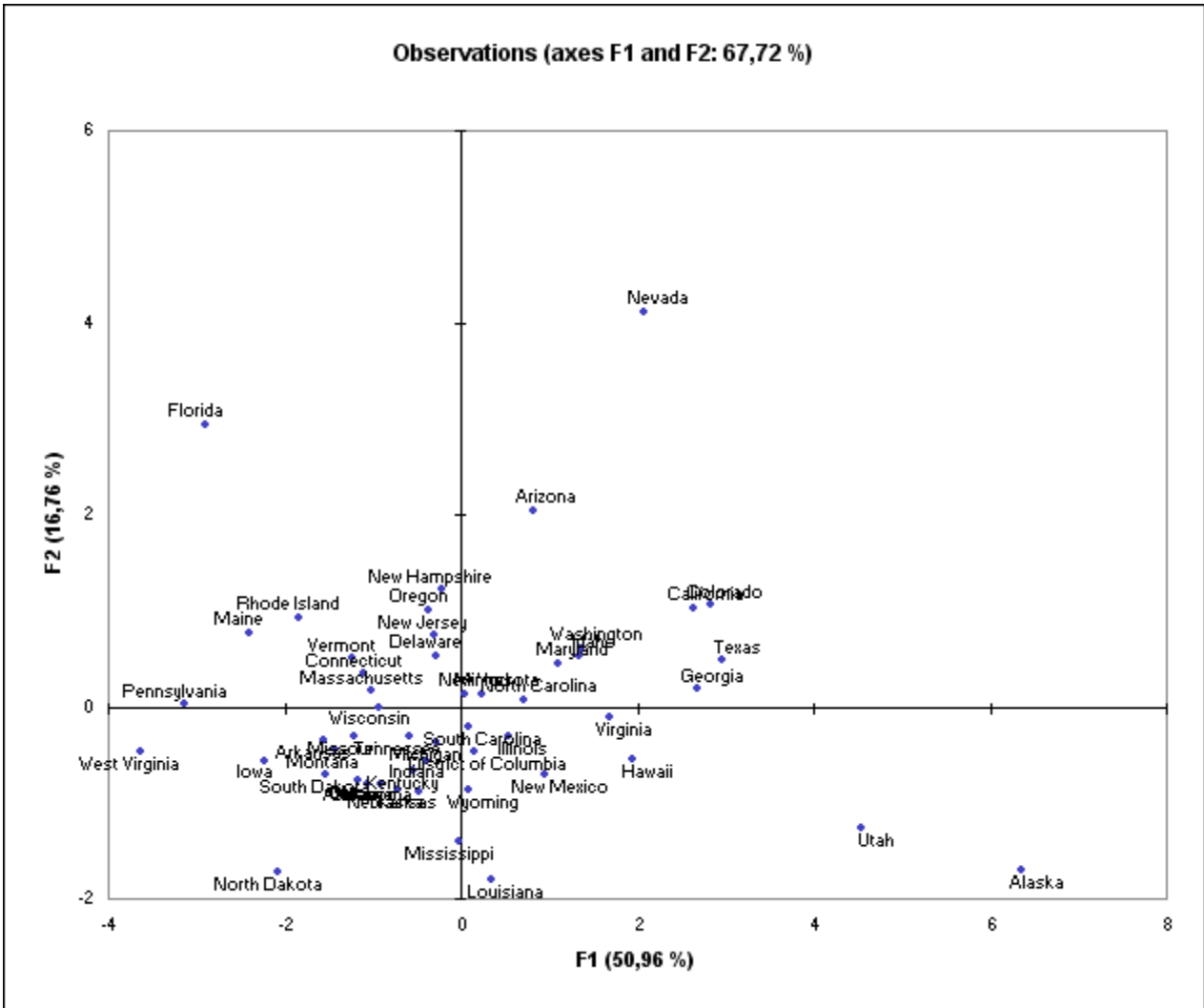


The correlation circle is useful in interpreting the meaning of the axes. In this example, the horizontal axis is linked with age and population renewal, and the vertical axis with domestic migration. These trends will be helpful in interpreting the next map. To confirm that a variable is well linked with an axis, take a look at the squared cosines table: the greater the squared cosine, the greater the link with the corresponding axis. The closer the squared cosine of a given variable is to zero, the more careful you have to be when interpreting the results in terms of trends on the corresponding axis. Looking at this table we can see that the trends for international migration would be best viewed on a F2/F3 map.

Squared cosines of the variables:

	F1	F2	F3	F4
Net Domes	0.026	0.707	0.175	0.029
Federal/Civ	0.280	0.044	0.041	0.623
Net Int. Mi	0.174	0.317	0.463	0.017
Period Birt	0.559	0.043	0.043	0.074
Period Dea	0.780	0.026	0.002	0.002
< 65 Pop.	0.874	0.017	0.055	0.015
> 65 Pop.	0.874	0.017	0.055	0.015


The next chart can be the ultimate goal of the Principal Component Analysis (PCA). It enables you to look at the data on a two-dimensional map, and to identify trends. We can see that the demographics of Nevada and Florida are unique, as are the demographics of Utah and Alaska, two states that share common characteristics. Going back to the table, we can confirm that Utah and Alaska have a low population rate of people over age 65. Utah has the highest birth rate in the U.S., and Alaska ranks high as well.



Watch this video to see how the settings were performed.

[http://www.youtube.com/watch?v=ca0OBsml79o&feature=player\\_embedded](http://www.youtube.com/watch?v=ca0OBsml79o&feature=player_embedded)



Click  to view a 3D visualization on the first three axes generated by XLSTAT-3DPlot.

## **Note on the usage of Principal Component Analysis**

Principal component analysis is often performed before a regression, to avoid using correlated variables, or before clustering the data, to have a better overview of the variables. The number of clusters might sometimes be a simple guess based on the maps. The above demographic data have also been used in the tutorial on hierarchical clustering. The ">65 pop" variable has been removed as its inclusion would double the weight of the age variables in the analysis.