SPIDERFINANCIAL
www.spiderfinancial.com

Phone: 1-888-427-9486
1-312-324-0367
Fax: 1-312-238-9092
info@spiderfinancial.com

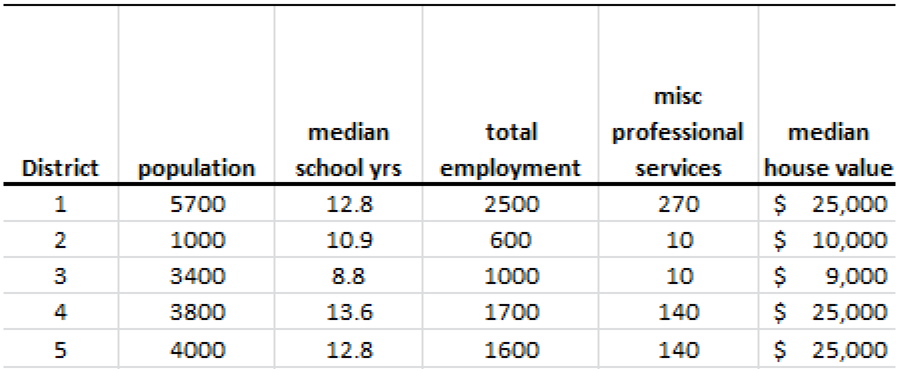
Tutorial: Principal Component 101

This is the first entry in what will become an ongoing series on principal components analysis (PCA). In this tutorial, we will start with the general definition, motivation and applications of a PCA, and then use NumXL to carry on such analysis. Next, we will closely examine the different output elements in an attempt to develop a solid understanding of PCA, which will pave the way to a more advanced treatment in future issues.

In this tutorial, we will use the socioeconomic data provided by Harman (1976). The five variables represent total population (“Population”), median school years (“School”), total employment (“Employment”), miscellaneous professional services (“Services”), and median house value (“House Value”). Each observation represents one of twelve census tracts in the Los Angeles Standard Metropolitan Statistical Area.

Data Preparation

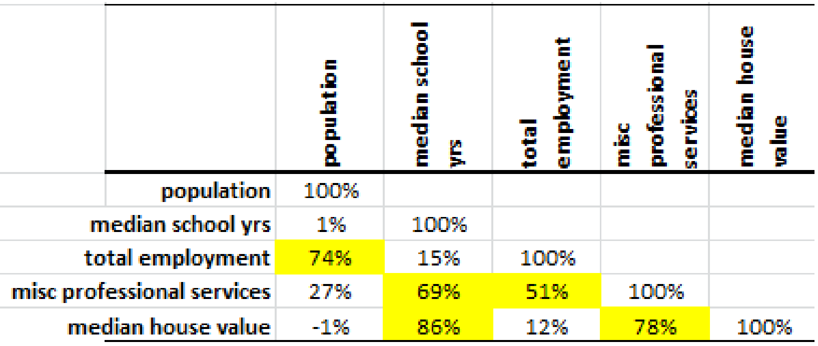
First, let’s organize our input data. First, we place the values of each variable in a separate column, and each observation (i.e. census tract in LA) on a separate row.



District	population	median school yrs	total employment	misc professional services	median house value
1	5700	12.8	2500	270	$ 25,000
2	1000	10.9	600	10	$ 10,000
3	3400	8.8	1000	10	$ 9,000
4	3800	13.6	1700	140	$ 25,000
5	4000	12.8	1600	140	$ 25,000

Note that the scales (i.e. magnitude) of the variables vary significantly, so any analysis of raw data will be biased toward the variables with a larger scale, and downplay the effect of ones with a lower scale.

To better understand the problem, let’s compute the correlation matrix for the 5 variables:



	population	median school yrs	total employment	misc professional services	median house value
population	100%				
median school yrs	1%	100%			
total employment	74%	15%	100%		
misc professional services	27%	69%	51%	100%	
median house value	-1%	86%	12%	78%	100%

The five (5) variables are highly correlated, so one may wonder:

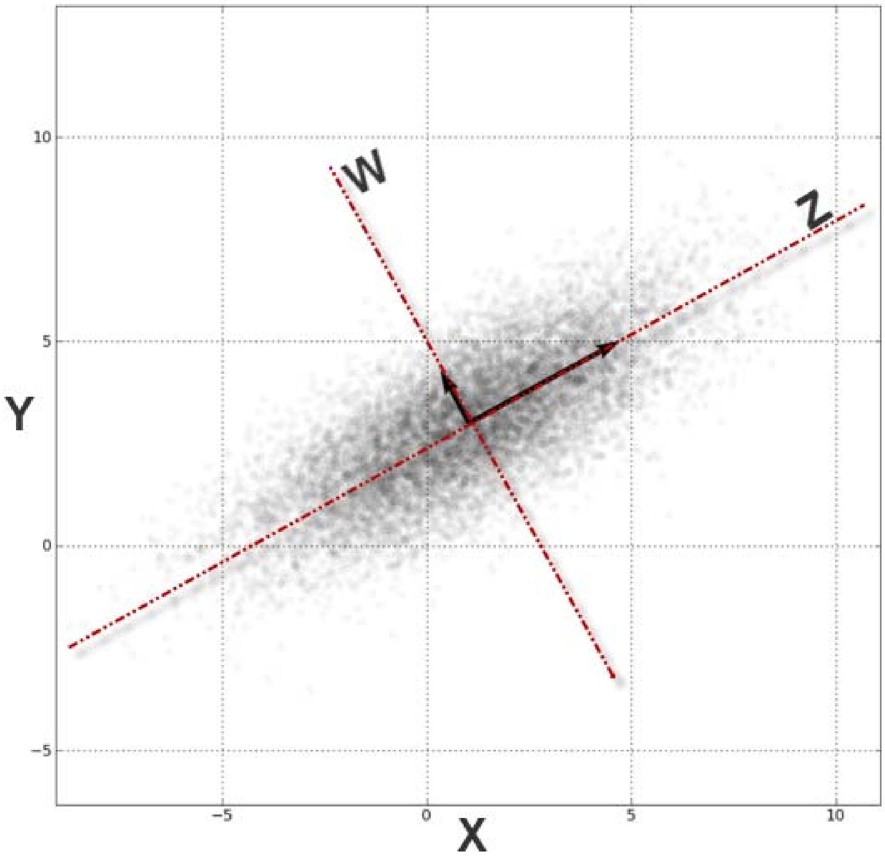
1. If we were to use those variables to predict another variable, do we need the 5 variables?
2. Are there hidden forces (drivers or other factors) that move those 5 variables?

In practice, we often encounter correlated data series: commodity prices in different locations, future prices for different contracts, stock prices, interest rates, etc.

In plain English, what is principal component analysis (PCA)?

PCA is a technique that takes a set of correlated variables and linearly transforms those variables into a set of uncorrelated factors.

To explain it further, you can think about PCA as an axis-system transformation. Let's examine this plot of two correlated variables:



Simply put, from the (X, Y) Cartesian system, the data points are highly correlated. By transforming (rotating) the axis into (Z, W), the data points are no longer correlated.

In theory, the PCA finds that those transformations (of the axis) of data points will look uncorrelated with their respect.

OK, now where are the principal components?

To transform the data points from the (X, Y) Cartesian system to (Z, W), we need to compute the z and w values of each data point:

$$\begin{aligned} z_i &= \alpha_1 x_i + \beta_1 y_i \\ w_i &= \alpha_2 x_i + \beta_2 y_i \end{aligned}$$

In effect, we are replacing the input variables (x_i, y_i) with those of (z_i, w_i). The (z_i, w_i) values are ones we refer to as the principal components.

Alright, how do we reduce the dimensions of the variables?

When we transform the values of the data points (x_i, y_i) into the new axis system (z_i, w_i), we may find that a few axes capture more of the values' variation than others. For instance, in our example above, we may claim that all w_i values are plain zero and don't really matter.

$$\begin{aligned} x_i &= \gamma_1 z_i + \varphi_1 w_i \\ y_i &= \gamma_2 z_i + \varphi_2 w_i \end{aligned} \rightarrow \begin{aligned} x_i &= \gamma_1 z_i \\ y_i &= \gamma_2 z_i \end{aligned}$$

In effect, the two dimensional system (z_i, w_i) is reduced to a one-dimensional system (z_i).

Of course, for this example, dropping the W factor distorts our data, but for higher dimensions it may not be so bad.

Which component should we drop?

In practice, we order the components (aka factors) in terms of their variance (highest first) and examine the effect of removing the ones of lower variance (right most) in an effort to reduce the dimension of the data set with minimal loss of information.

Why should we care about principal components?

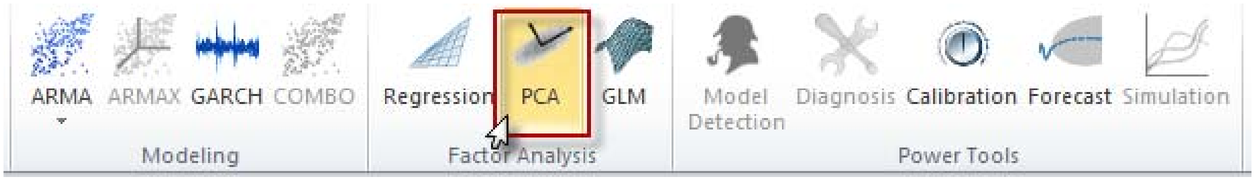
A risk manager can quantify their overall risk in terms of a portfolio aggregate exposure to a handful of drivers, instead of tens of hundreds of correlated securities prices. Furthermore, designing an effective hedging strategy is vastly simplified.

For traders, quantifying trades in terms of their sensitivities (e.g. delta, gamma, etc.) to those drivers gives trader options to substitute (or trade) one security for another, construct a trading strategy, hedge, synthesize a security, etc.

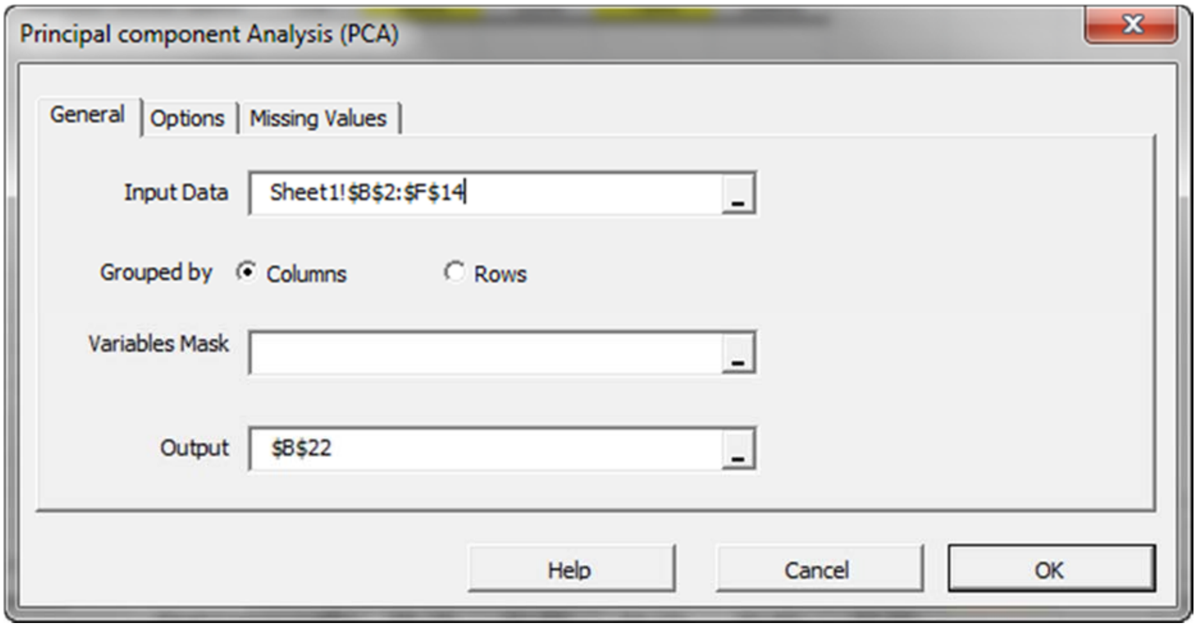
A data modeler can reduce the number of input variables with minimal loss of information.

Process

Now we are ready to conduct our PCA. First, select an empty cell in your worksheet where you wish the output to be generated, then locate and click on the “PCA” icon in the NumXL tab (or toolbar).



The principal component analysis Wizard pops up.



Input Data Sheet1!$B$2:$F$14

Grouped by ● Columns ○ Rows

Variables Mask

Output $B$22

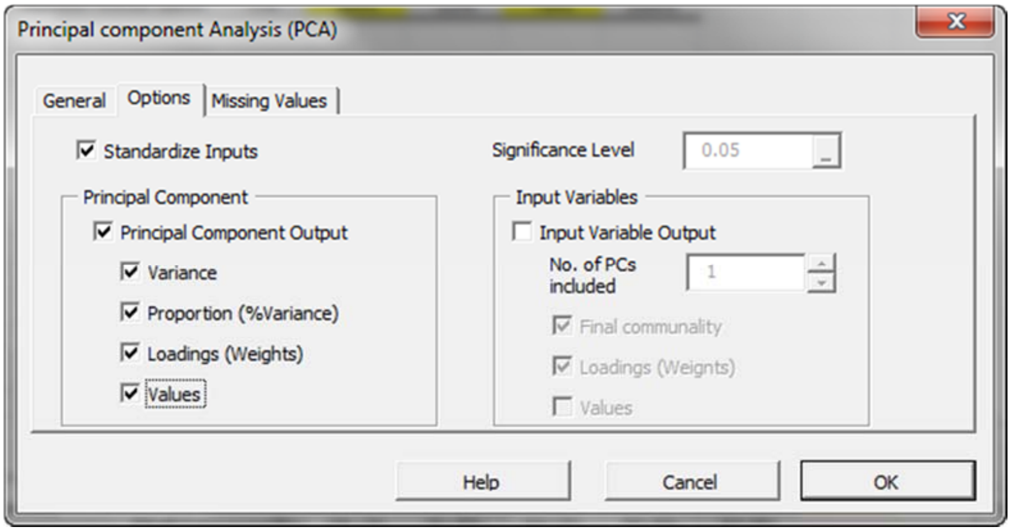
Select the cells range for the five input variable values.

Notes:

1. The cells range includes (optional) the heading (“Label”) cell, which would be used in the output tables where it references those variables.
2. The input variables (i.e. X) are already grouped by columns (each column represents a variable), so we don’t need to change that.
3. Leave the “Variable Mask” field blank for now. We will revisit this field in later entries.
4. By default, the output cells range is set to the current selected cell in your worksheet.

Finally, once we select the Input data (X) cells range, the “Options” and “Missing Values” tabs become available (enabled).

Next, select the “Options” tab.



Principal component Analysis (PCA)

☑ Standardize Inputs

Significance Level 0.05

Principal Component

☑ Principal Component Output

☑ Variance

☑ Proportion (%Variance)

☑ Loadings (Weights)

☑ Values

Input Variables

☐ Input Variable Output

No. of PCs included 1

☑ Final communality

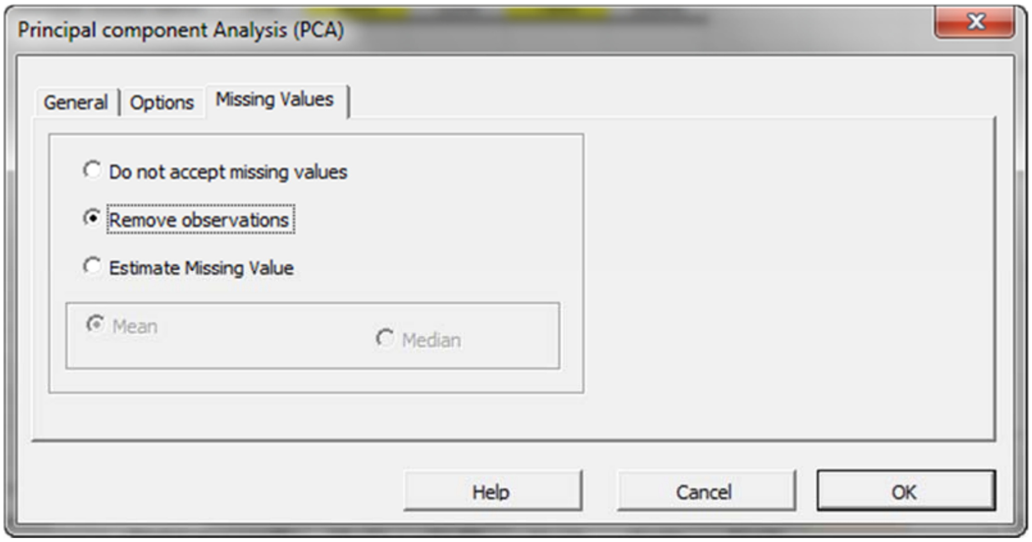
☑ Loadings (Weights)

☐ Values

Initially, the tab is set to the following values:

- "Standardize Input" is checked. This option in effect replace the values of each variable with its standardized version (i.e. subtract the mean and divide by standard deviation). This option overcomes the bias issue when the values of the input variables have different magnitude scales. **Leave this option checked.**
- "Principal Component Output" is checked. This option instructs the wizard to generate PCA related tables. **Leave it checked.**
- Under "Principal Component," check the "Values" option to display the values for each principal component.
- The significance level (aka α) is set to 5%.
- The "Input Variables" is unchecked. **Leave it unchecked for now.**

Now, click on the "Missing Values" tab.



Principal component Analysis (PCA)

○ Do not accept missing values

● Remove observations

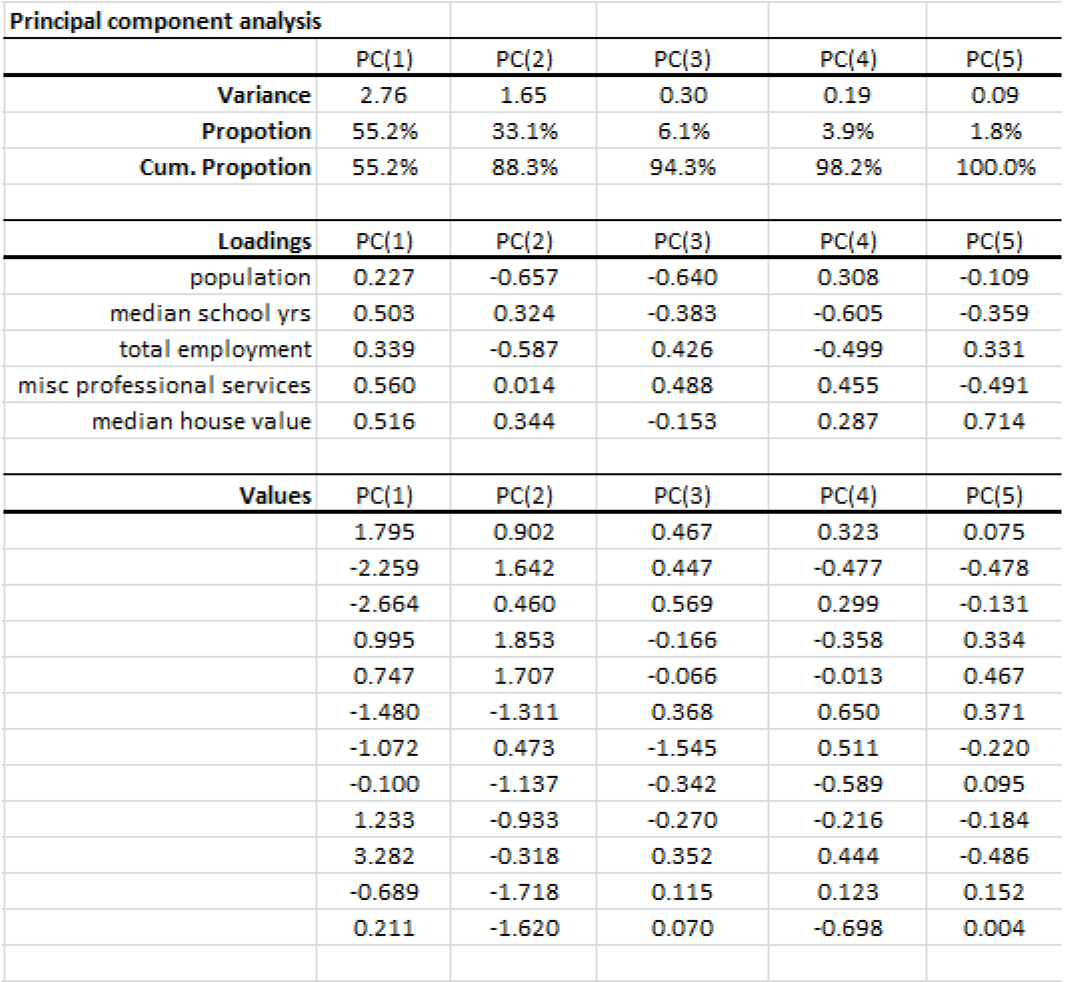
○ Estimate Missing Value

● Mean ○ Median

In this tab, you can select an approach to handle missing values in the data set (X's). By default, any missing value found in any of input variables (X) in any observation would exclude the observation from the analysis.

This treatment is a good approach for our analysis, so let's leave it unchanged.

Now, click "OK" to generate the output tables.



Principal component analysis

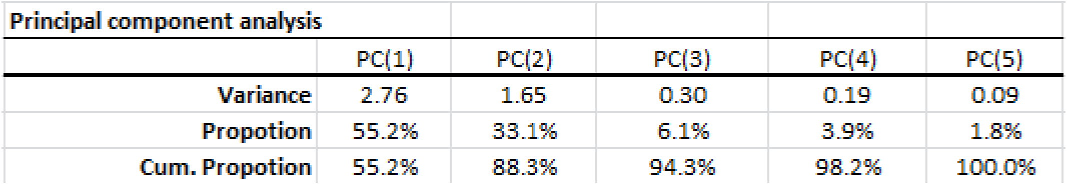
	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
Variance	2.76	1.65	0.30	0.19	0.09
Propotion	55.2%	33.1%	6.1%	3.9%	1.8%
Cum. Propotion	55.2%	88.3%	94.3%	98.2%	100.0%

Loadings	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
population	0.227	-0.657	-0.640	0.308	-0.109
median school yrs	0.503	0.324	-0.383	-0.605	-0.359
total employment	0.339	-0.587	0.426	-0.499	0.331
misc professional services	0.560	0.014	0.488	0.455	-0.491
median house value	0.516	0.344	-0.153	0.287	0.714

Values	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
	1.795	0.902	0.467	0.323	0.075
	-2.259	1.642	0.447	-0.477	-0.478
	-2.664	0.460	0.569	0.299	-0.131
	0.995	1.853	-0.166	-0.358	0.334
	0.747	1.707	-0.066	-0.013	0.467
	-1.480	-1.311	0.368	0.650	0.371
	-1.072	0.473	-1.545	0.511	-0.220
	-0.100	-1.137	-0.342	-0.589	0.095
	1.233	-0.933	-0.270	-0.216	-0.184
	3.282	-0.318	0.352	0.444	-0.486
	-0.689	-1.718	0.115	0.123	0.152
	0.211	-1.620	0.070	-0.698	0.004

Analysis

1. PCA Statistics



Principal component analysis

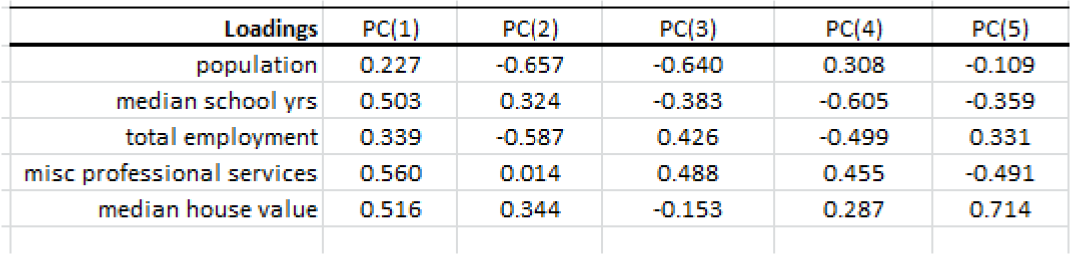
	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
Variance	2.76	1.65	0.30	0.19	0.09
Propotion	55.2%	33.1%	6.1%	3.9%	1.8%
Cum. Propotion	55.2%	88.3%	94.3%	98.2%	100.0%

1. The principal components are ordered (and named) according to their variance in a descending order, i.e. PC(1) has the highest variance.
2. In the second row, the proportion statistics explain the percentage of variation in the original data set (5 variables combined) that each principal component captures or accounts for.
3. The cumulative proportion is a measure of total variation explained by the principal components up to this current component.

   Note: In our example, the first three PC account for 94.3% of the variation of the 5 variables.
4. Note that the sum of variances of the PC should yield the number of input variables, which in this case is five (5).

2. Loadings

In the loading table, we outline the weights of a linear transformation from the input variable (standardized) coordinate system to the principal components.



Loadings	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
population	0.227	-0.657	-0.640	0.308	-0.109
median school yrs	0.503	0.324	-0.383	-0.605	-0.359
total employment	0.339	-0.587	0.426	-0.499	0.331
misc professional services	0.560	0.014	0.488	0.455	-0.491
median house value	0.516	0.344	-0.153	0.287	0.714

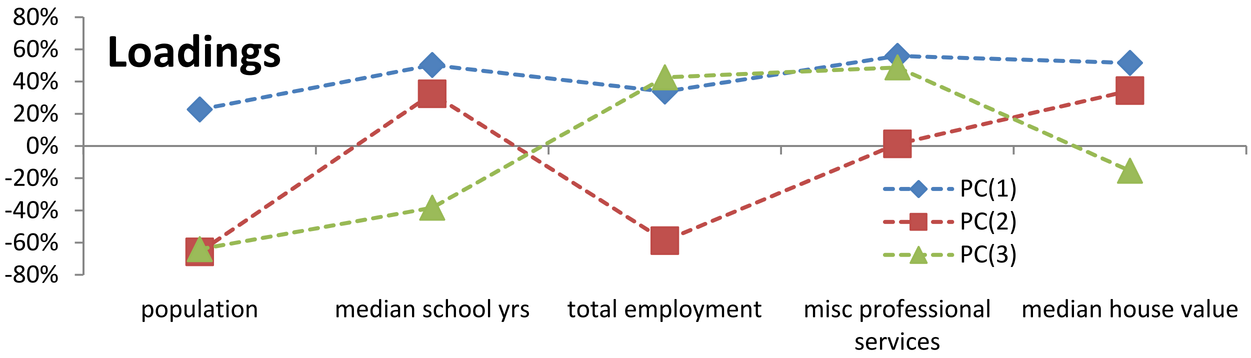
For example, the linear transformation for PC_1 is expressed as follows:

$$PC_1 = 0.27X_1 + 0.503X_2 + 0.339X_3 + 0.56X_4 + 0.516X_5$$

Note:

1. The squared loadings (column) adds up to one.

$$\sum_{j=1}^{5} \beta_j^2 = 1$$

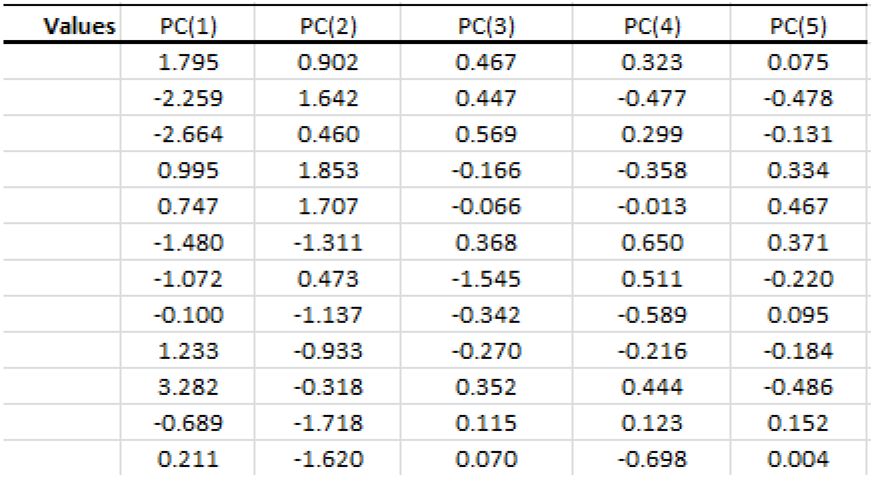


2. In the graph above, we plotted the loadings for our input variables in the first three components.
3. The median school years, misc. professional services and median house value variables have comparable loadings in PC(1), next comes total employment loading and finally, population. One may propose this as a **proxy for the wealth/income factor**.
4. Interpreting the loadings for the input variables in the remaining components prove to be more difficult, and require a deeper level of domain expertise.
5. Finally, computing the input variables back from the PC can be easily done by applying the weights in the row instead of the column. For example, the population factor is expressed as follows:

$$X_1 = 0.227PC_1 - 0.657PC_2 - 0.64PC_3 + 0.308PC_4 - 0.109PC_5$$

6. We'll discuss the PC loading later in this tutorial.

3. Principal Component Values

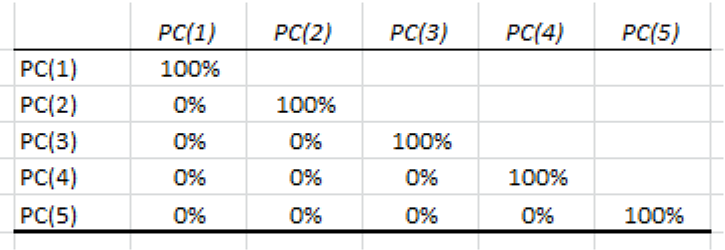


Values	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
	1.795	0.902	0.467	0.323	0.075
	-2.259	1.642	0.447	-0.477	-0.478
	-2.664	0.460	0.569	0.299	-0.131
	0.995	1.853	-0.166	-0.358	0.334
	0.747	1.707	-0.066	-0.013	0.467
	-1.480	-1.311	0.368	0.650	0.371
	-1.072	0.473	-1.545	0.511	-0.220
	-0.100	-1.137	-0.342	-0.589	0.095
	1.233	-0.933	-0.270	-0.216	-0.184
	3.282	-0.318	0.352	0.444	-0.486
	-0.689	-1.718	0.115	0.123	0.152
	0.211	-1.620	0.070	-0.698	0.004

In the PC values table, we calculate the transformation output value for each dimension (i.e. component), so the 1st row corresponds to the 1st data point, and so on.

The variance of each column matches the value in the PCA statistics table. Using Excel, compute the biased version of the variance function (VARA).

By definition, the values in the PCs are uncorrelated. To verify, we can calculate the correlation matrix:



	PC(1)	PC(2)	PC(3)	PC(4)	PC(5)
PC(1)	100%				
PC(2)	0%	100%			
PC(3)	0%	0%	100%		
PC(4)	0%	0%	0%	100%	
PC(5)	0%	0%	0%	0%	100%

Conclusion

In this tutorial, we converted a set of five correlated variables into five uncorrelated variables without any loss of information.

Furthermore, we examined the proportion (and cumulative proportion) of each component as a measure of variance captured by each component, and we found that the first three factors (components) account for 94.3% of the five variables variation, and the first four components account for 98%.

What do we do now?

One of the applications of PCA is dimension reduction; as in, can we drop one or more components and yet retain the information in the original data set for modeling purposes?

In our second entry, we will look at the variation of each input variable captured by principal components (micro-level) and compute the fitted values using a reduced set of PCs.

We will cover this particular issue in a separate entry of our series.