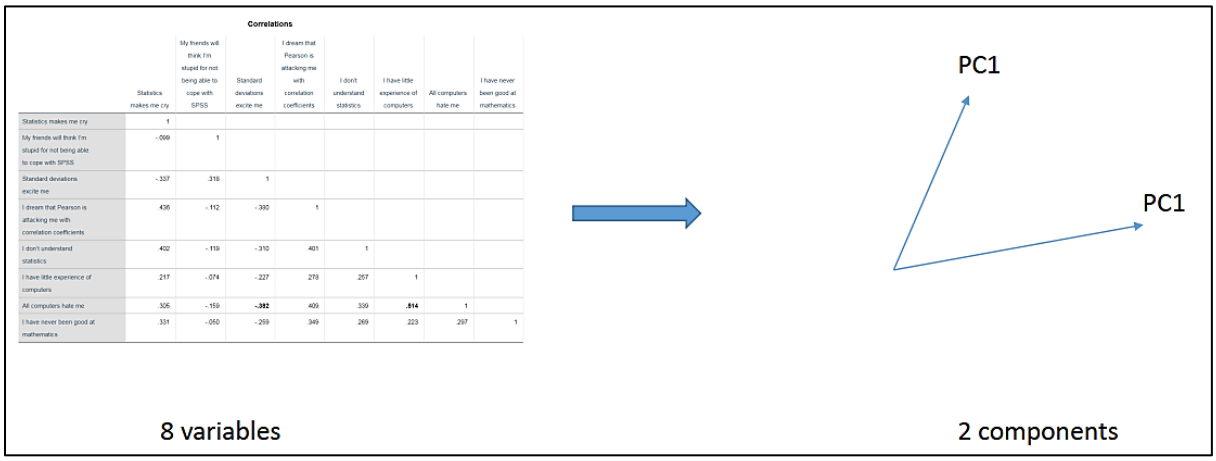
Principal Component Analysis

Dr. Myat Htut Nyunt
MBBS, MMedSc (Microbiology),
DAP&E, PhD (Parasitology)
Research Scientist
Department of Medical Research
myathtutnyunt@mohs.gov.mm

What is PCA?

PCA is a multivariate method used for data reducing purposes to represent a set of variables by a smaller number of variables called "principal components".

Aim of the PCA is to replicate the correlation matrix using a set of components that are fewer in number than the original set of items.



What are the application of PCA?

- Tool for exploratory data analysis

- Predictive model

- Visualize the genetic distance and relatedness between the population

Input Data

- should be quantitative (interval) or ordinal.

- should be linearly related to each other (checked by scatter plot of pairs of variables) or at least moderately correlated to each other.

- preferred minimum sample size requirement of 100 valid cases. While principal component analysis can be conducted on a sample that has fewer than 100 cases, but more than 50 cases, we should be cautious about its interpretation.

- The ratio of cases to variables in a principal component analysis should be at least 5 to 1. With 620 and 12 variables, the ratio of cases to variables is 51.67 to 1, which exceeds the requirement for the ratio of cases to variables.

Five basic steps

1. Data collection and generation of the correlation matrix

2. Partition of variance into common and unique components (unique may include random error variability)

3. Extraction of initial factor solution

4. Rotation and interpretation

5. Construction of scales or factor scores to use in further analyses

Preparation before analysis

- Any computed variables should be excluded. (as they will have a correlation of 1 with the variable from which they were calculated)

- All variables should measure the construct in the same direction. (Check the questions whether it include same or reverse direction)

Check correlation matrix

	wordmean	sentence	paragrap	lozenges	cubes	visperc
wordmean	1.000					
sentence	.696	1.000				
paragrap	.743	.724	1.000			
lozenges	.369	.335	.326	1.000		
cubes	.184	.179	.211	.492	1.000	
visperc	.230	.367	.343	.492	.483	1.000

- few correlation (<0.3) should not be carried out for further analysis.

Name of the matrix	Elements are:	Good signs	Bad signs
Correlation 'R'	correlations	Many above 0.3 and possible clustering	Few above 0.3
Partial correlation		Few above 0.3 and possible clustering	Many above 0.3
Anti-image correlation	Partial correlations - reversed	Few above 0.3 and possible clustering	Many above 0.3

- Two tests can be used

a) The **Bartlett Test of Sphericity** compares the correlation matrix with a matrix of zero correlations (small p-value = unlikely to obtain observed correlation matrix from the population with zero correlation)

b) **Kaiser-Meyer-Olkin Measure of Sampling Adequacy (MSA).** MSA<0.7 should be removed.

Eigenvalues

- variances of the principal component (eg. First eigenvalue is the variance of first PC

- Total variance explained by given principal component

- Eigenvalue >0, good

- Negative eigenvalue, ill-conditioned.

- Positive eigenvalue, multicollinearity

- Any PC that account for only a small proportion of the variation in the data (those with small eigenvalues) are discarded.

- Different methods can be used to discard. Example, if correlation matrix is used, choose only those principal components with eigenvalues over 1.

Eigenvectors

- weight for each eigenvalue

- eigenvector times the square root of the eigenvalue → component loadings

Component loading

- correlation of each item with the principal component

- Higher values mean a closer relationship. (Higher the value the better).

Communality for the PA

- Is the total influence on a single observed variable from all the factors associated with it. It is equal to the sum of all the squared factor loadings for all the factors related to the observed variable and this value is the same as R^2 in multiple regression.

- The value ranges from zero to 1 where 1 indicates that the variable can be fully defined by the factors and has no uniqueness. In contrast a value of 0 indicates that the variable cannot be predicted at all from any of the factors.

Calculation in SPSS

Remark!

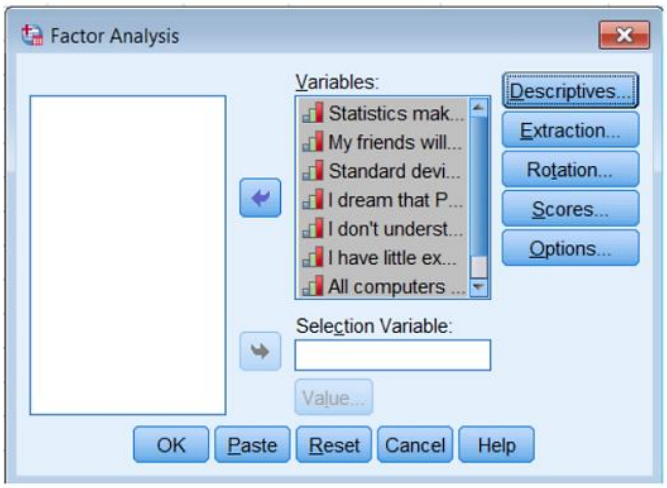
1. You have to install SPSS in your computer first.

2. Noted that SPSS will not give you the actual principal components. However, it can be calculated from the output provided.

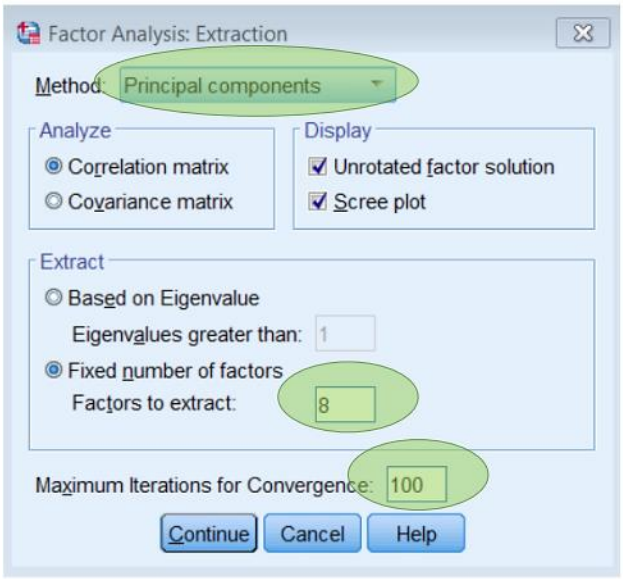
Step by step for calculation in SPSS

- Analyze> Dimension Reduction > Factor

Analyze – Dimension Reduction – Factor



- Select the variables and move into the variables box.
- Select "Principal Components" under "Extraction" window.
- Select "Correlation Matrix" under "Analyze". (Default eigenvalues over 1)
- Click the "Continue"
- In the "Rotation" window, select "None" under "Method". (or select the suitable rotation method)
- Click on "Continue".
- In "Score" window, specify whether PC values will be saved or not. (if save, new variables will be appeared as a new variable.
- Under "Method", choose "Regression".
- Click on "Continue".



Factors= components

8 components is NOT what you typically want to use.

Remark

- ✓ Univariate descriptive: mark this check box to get a tally of valid cases.
- ✓ Initial solution: to get the statistics needed to determine the number of factors to extract.
- ✓ Coefficient: checkbox to get a correlation matrix, one of the outputs needed to assess the appropriateness of factor analysis for the variables.
- ✓ KMO and Bartlett's test of sphericity checkbox to get more outputs used to assess the appropriateness of factor analysis for the variables.
- ✓ The Anti-image checkbox to get more outputs used to assess the appropriateness of factor analysis for the variables.
- ✓ The extraction method refers to the mathematical method that SPSS uses to compute the factors or components.
- ✓ Rotation… button to specify statistics to include in the output. The rotation method refers to the mathematical method that SPSS rotate the axes in geometric space. This makes it easier to determine which variables are loaded on which components. Varimax method as the type of rotation to be used in the analysis.

Component loadings

correlation of each item with the principal component

Component Matrix[a]

	Component							
	1	2	3	4	5	6	7	8
Statistics makes me cry	.659	.136	-.398	.160	-.064	.568	-.177	.068
My friends will think I'm stupid for not being able to cope with SPSS	-.300	.866	-.025	.092	-.290	-.170	-.193	-.001
Standard deviations excite me	-.653	.409	.081	.064	.410	.254	.378	.142
I dream that Pearson is attacking me with correlation coefficients	.720	.119	-.192	.064	-.288	-.089	.563	-.137
I don't understand statistics	.650	.096	-.215	.460	.443	-.326	-.092	-.010
I have little experience of computers	.572	.185	.675	.031	.107	.176	-.058	-.369
All computers hate me	.718	.044	.453	-.006	-.090	-.051	.025	.516
I have never been good at mathematics	.568	.267	-.221	-.694	.258	-.084	-.043	-.012

Extraction Method: Principal Component Analysis.

a. 8 components extracted. 3.057 1.067 0.958 0.736 0.622 0.571 0.543 0.446

Sum of squared loadings across components is the communality

Q: why is it 1?

$0.659^2 = 0.434$

43.4% of the variance explained by first component

$0.136^2 = 0.018$

1.8% of the variance explained by second component

Sum squared loadings down each column (component) = eigenvalues

3.057 1.067 0.958 0.736 0.622 0.571 0.543 0.446

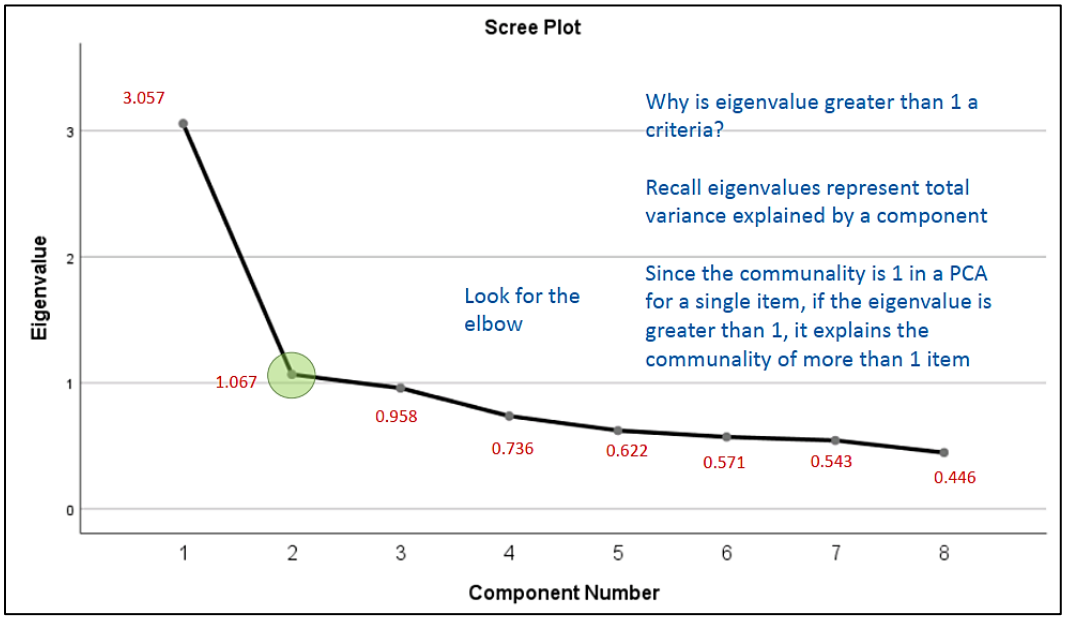
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.057	38.206	38.206	3.057	38.206	38.206
2	1.067	13.336	51.543	1.067	13.336	51.543
3	.958	11.980	63.523	.958	11.980	63.523
4	.736	9.205	72.728	.736	9.205	72.728
5	.622	7.770	80.498	.622	7.770	80.498
6	.571	7.135	87.632	.571	7.135	87.632
7	.543	6.788	94.420	.543	6.788	94.420
8	.446	5.580	100.000	.446	5.580	100.000

Extraction Method: Principal Component Analysis.

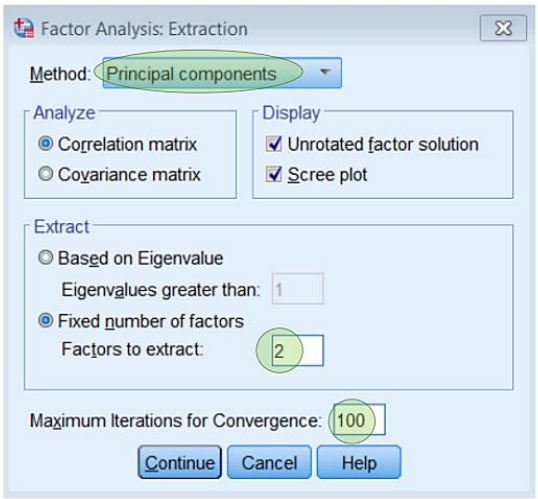
Look familiar? Extraction Sums of Squared Loadings = Eigenvalues

Choosing the number of the components to extract



Running a PCA with two components

Analyze – Dimension Reduction – Factor



Goal of PCA is dimension reduction

This is more realistic than an 8-component solution

Recall these numbers from the 8-component solution

3.057 1.067 0.958 0.736 0.622 0.571 0.543 0.446

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.057	38.206	38.206	3.057	38.206	38.206
2	1.067	13.336	51.543	1.067	13.336	51.543
3	.958	11.980	63.523			
4	.736	9.205	72.728			
5	.622	7.770	80.498			
6	.571	7.135	87.632			
7	.543	6.788	94.420			
8	.446	5.580	100.000			

Notice only two eigenvalues

Extraction Method: Principal Component Analysis.

Notice communalities not equal 1

Communalities

	Initial	Extraction
Statistics makes me cry	1.000	.453
My friends will think I'm stupid for not being able to cope with SPSS	1.000	.840
Standard deviations excite me	1.000	.594
I dream that Pearson is attacking me with correlation coefficients	1.000	.532
I don't understand statistics	1.000	.431
I have little experience of computers	1.000	.361
All computers hate me	1.000	.517
I have never been good at mathematics	1.000	.394

Extraction Method: Principal Component Analysis.

Calculation for coefficients of the PC

- SPSS will give you a table entitled the "Component Matrix". The components are listed as columns in the table; the variables are listed as rows.
- For the coefficient, you must divide the values in the table by the SQUARE ROOT of the corresponding eigenvalue. Example, you have to divide each number in the first column by the square root of the first (largest) eigenvalue. Similarly, the second column should be divided by the square root of the eigenvalue corresponding to component 2, and so on.
- To get the values of the principal components as new variables in the data set:
- SPSS will have saved variable called FAC1_1, FAC2_1, and so on. These are NOT the value of the PC. To get the PC, you have to MULTIFY these factor scores by the square root of the corresponding eigenvalue.
- For example, if the eigenvalue for the first principal component was 3.65, you would compute the first principal component in SPSS as follows:
  - ✓ Transform > Compute variable.
  - ✓ Under "Target variable", type "PC1" and under "Numeric Expression", type "FAC1_1*sqrt(3.65)
  - ✓ Click on "OK".
  - ✓ (Calculate the second PC, 3rd PC etc by replacing 3.65 by whatever the second, third eigenvalue is)

Glossary

Principal component analysis: Factor model in which the factors are based on summarizing the total variance. With PCA, unities are used in the diagonal of the correlation matrix computationally implying that all the variance is common or shared. Algorithm lacking underlying model.

Common factor analysis: Factor model explores a reduced correlation matrix. That is, communalities (r2) are inserted on the diagonal of the correlation matrix, and the extracted factors are based only on the common variance, with specific and error variances excluded. Explores underlying "latent" structure of data. Model assumes variability partitionable into common and unique components.

Common variance: Variance shared with other variables in the factor analysis.

Specific or unique variance: Variance of each variable unique to that variable and not explained or associated with other variables in the factor analysis.

Communality: Total amount of variance an original variable shares with all other variables included in the analysis.

Eigenvalue: Column sum of squared loadings for a factor, i.e., the latent root. It conceptually represents that amount of variance accounted for by a factor.

Sphericity test: Statistical test for the overall significance of all correlations within a correlation matrix.

Factor: Linear combination (variate) of the original variables. Factors also represent the underlying dimensions (constructs) that summarize or account for the original set of observed variables.

Factor loadings: Correlation between the original variables and the factors, and the key to understanding the underlying nature of a particular factor. Squared factor loadings indicate what percentage of the variance in an original variable is explained by a factor.

Factor matrix: Table displaying the factor loadings of all variables on each factor.

Factor score: Composite measure created for each observation on each factor extracted in the factor analysis. The factor weights are used in conjunction with the original variable values to calculate each observation's score. The factor scores are standardized to reflect a z-score. Factor scores place each variable in a plane of multivariate variability.

Principal components analysis (PCA): PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables. It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance, and so on. This is called the principal axis method and results in orthogonal (uncorrelated) factors. PCA analyzes total (common and unique) variance.

Eigenvectors: Principal components (from PCA - principal components analysis) reflect both common and unique variance of the variables and may be seen as a variance-focused approach seeking to reproduce both the total variable variance with all components and to reproduce the correlations. PCA is far more common than PFA, however, and it is common to use "factors" interchangeably with "components." The principal components are linear combinations of the original variables weighted by their contribution to explaining the variance in a particular orthogonal dimension.

Eigenvalues: Also called characteristic roots. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor. The ratio of eigenvalues is the

ratio of explanatory importance of the factors with respect to the variables. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors. Eigenvalues measure the amount of variation in the total sample accounted for by each factor. A factor's eigenvalue may be computed as the sum of its square factor loadings for all the variables. Note that the eigenvalues associated with the unrotated and rotated solution will differ, though their total will be the same.

Factor loadings (factor or component coefficients): The factor loadings, also called component loadings in PCA, are the correlation coefficients between the variables (rows) and factors (columns). Analogous to Pearson's r, the squared factor loading is the percent of variance in that variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables. (Note the number of variables equals the sum of their variances as the variance of a standardized variable is 1.) This is the same as dividing the factor's eigenvalue by the number of variables.

PC scores: Also called component scores in PCA, these scores are the scores of each case (row) on each factor (column). To compute the factor score for a given case for a given factor, one takes the case's standardized score on each variable, multiplies by the corresponding factor loading of the variable for the given factor, and sums these products.

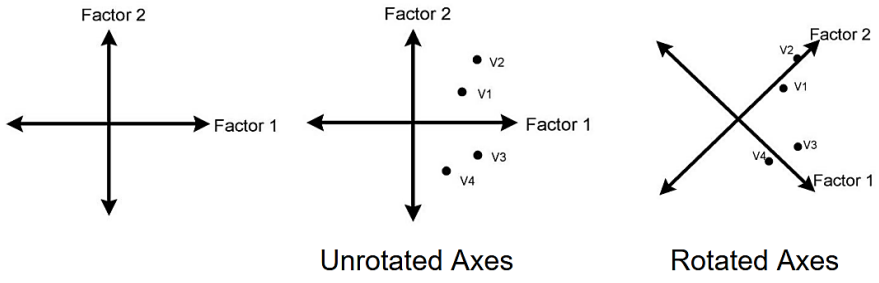
Factor rotation: Process of manipulation or adjusting the factor axes to achieve a simpler and pragmatically more meaningful factor solution.

Oblique factor rotation: Factor rotation computed so that the extracted factors are correlated. Rather than arbitrarily constraining the factor rotation to an orthogonal (90 degree angle) solution, the oblique solution identifies the extent to which each of the factors are correlated.

Orthogonal factor rotation: Factor rotation in which the factors are extracted so that their axes are maintained at 90 degrees. Each factor is independent of, or orthogonal to, all other factors. The correlation between teh factors is determined to be zero.

VARIMAX: One of the most popular orthogonal factor rotation methods

Factor rotation



Unrotated Axes Rotated Axes

- Each variable lies somewhere in the plane formed by these two factors. The factor loadings, which represent the correlation between the factor and the variable, can also be thought of as the variable's coordinates on this plane.
- In unrotated factor solution the Factor "axes" may not line up very well with the pattern of variables and the loadings may show no clear pattern. Factor axes can be rotated to more closely correspond to the variables and therefore become more meaningful. Relative relationships between variables are preserved.
- The rotation can be either orthogonal or oblique.