Principal Components Analysis

Dr. Myat Htut Nyunt

MBBS, MMedSc (Microbiology), DAP&E (IMR, Malaysia),

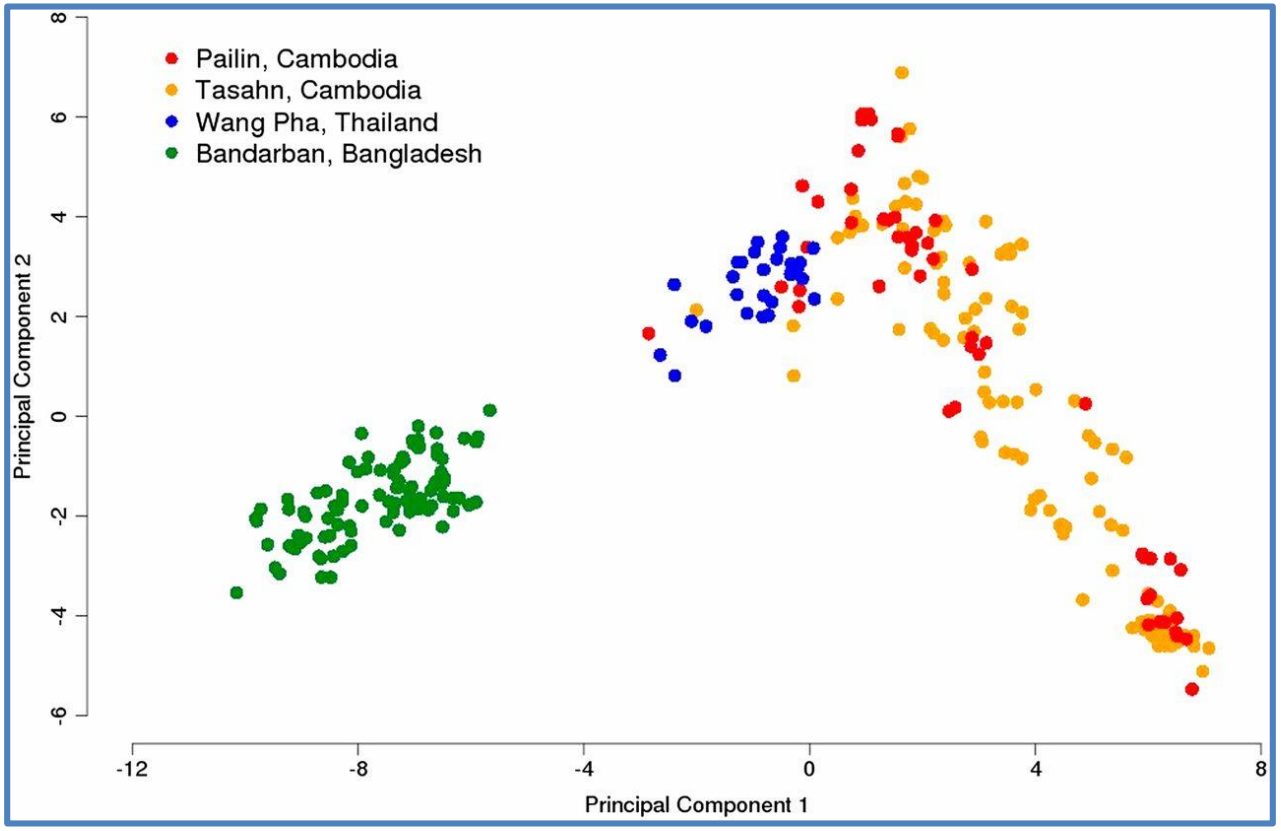
PhD (Parasitology) (KNU, RO Korea)

Research Scientist

Department of Medical Research

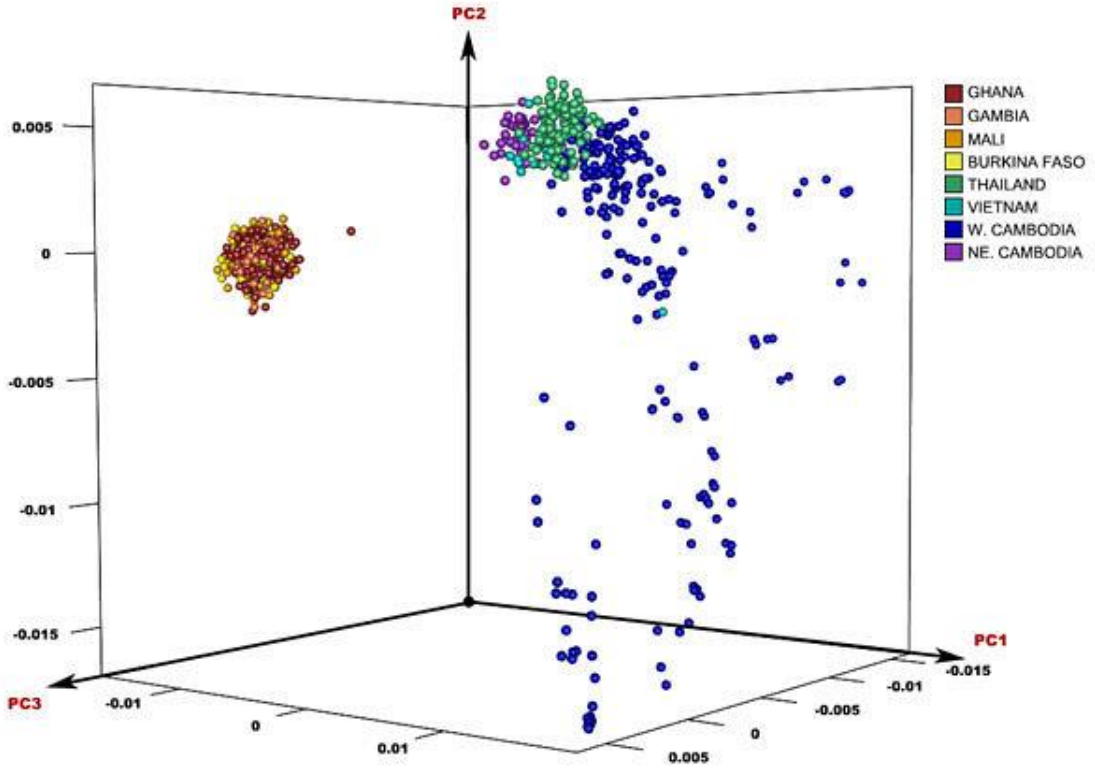
myathtutnyunt@mohs.gov.mm; drmhnyunt@myanmarhsrj.com

Use of PCA



Population structure by geography was examined using principal components analysis https://doi.org/10.1073/pnas.1211205110

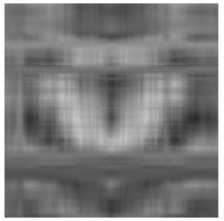
Use of PCA



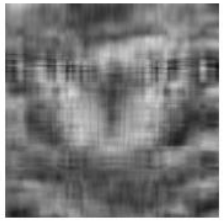
Three-dimensional plot of PCA using all 825 *Plasmodium falciparum* samples. PC1 separates continental clusters in Africa and Asia, whereas PC2 and PC3 largely account for the variability in western Cambodian samples.

Ref: Fingerprinting malaria parasite drug resistance, Sanger's press 2013

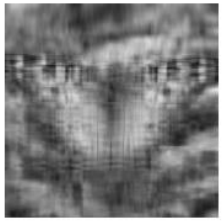
102.4:1 compression
2 principal components



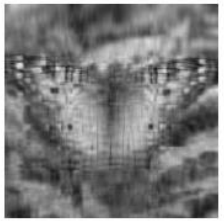
39.4:1 compression
6 principal components



24.4:1 compression
10 principal components



17.7:1 compression
14 principal components



12.5:1 compression
20 principal components

8.4:1 compression
30 principal components

6.3:1 compression
40 principal components

4.2:1 compression
60 principal components

2.8:1 compression
90 principal components

2.1:1 compression
120 principal components

1.7:1 compression
150 principal components

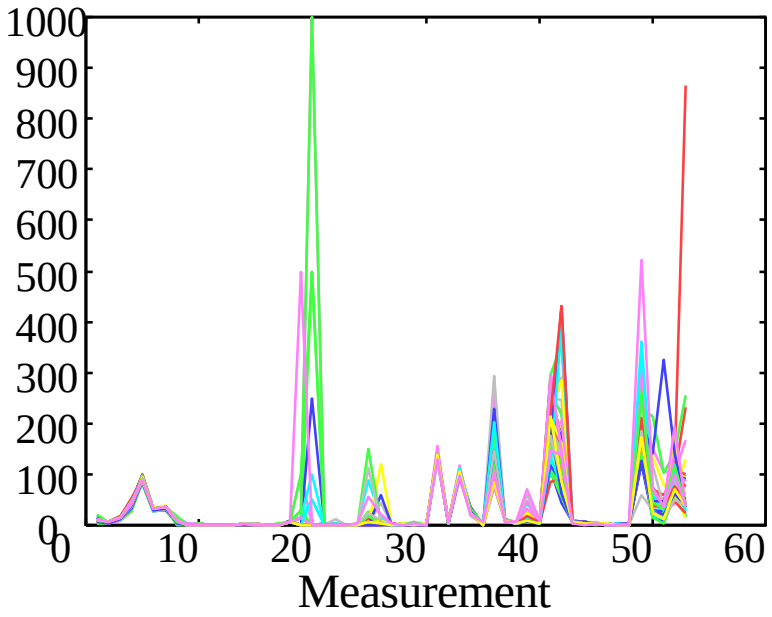
1.4:1 compression
180 principal components

Data Presentation

- Example: 53 Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics).
- Matrix Format

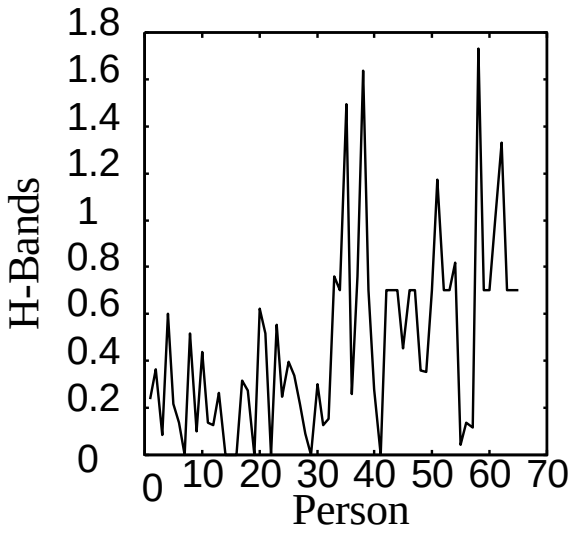
	H-WBC	H-RBC	H-Hgb	H-Hct	H-MCV	H-MCH	H-MCHC
A1	8.0000	4.8200	14.1000	41.0000	85.0000	29.0000	34.0000
A2	7.3000	5.0200	14.7000	43.0000	86.0000	29.0000	34.0000
A3	4.3000	4.4800	14.1000	41.0000	91.0000	32.0000	35.0000
A4	7.5000	4.4700	14.9000	45.0000	101.0000	33.0000	33.0000
A5	7.3000	5.5200	15.4000	46.0000	84.0000	28.0000	33.0000
A6	6.9000	4.8600	16.0000	47.0000	97.0000	33.0000	34.0000
A7	7.8000	4.6800	14.7000	43.0000	92.0000	31.0000	34.0000
A8	8.6000	4.8200	15.8000	42.0000	88.0000	33.0000	37.0000
A9	5.1000	4.7100	14.0000	43.0000	92.0000	30.0000	32.0000

- Spectral Format

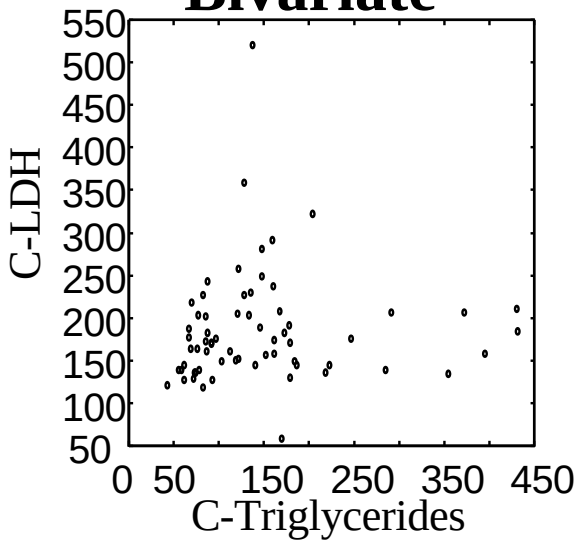


Data Presentation

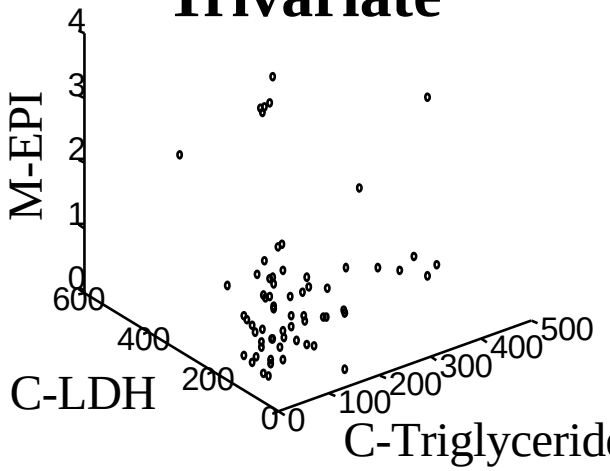
Univariate



Bivariate



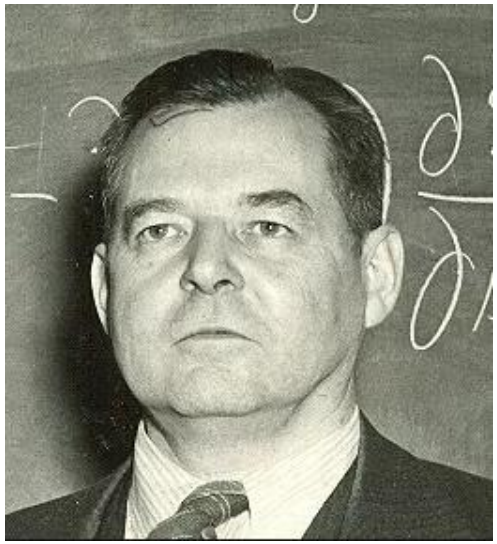
Trivariate



Data Presentation

- Better presentation than ordinate axes?
- Do we need a 53 dimension space to view data?
- How to find the ‘best’ low dimension space that conveys maximum useful information?
- One answer: Find “Principal Components”

What is "PCA"?



- invented by Pearson (1901) and Hotelling (1933)
- first applied in ecology by Goodall (1954) under the name "factor analysis"
- ("principal factor analysis" is a synonym of PCA).

What is "PCA"?

- **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.
- **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.

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".

What is "PCA"?

- To extract the important information from the data,
- To represent it as a set of **new** orthogonal **variables** called **principal components**, and
- To display the pattern of similarity of the observations and of the variables as points in maps.
- Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components

Use of "PCA"

- Tool for exploratory data analysis
- Predictive model
- Data Visualization (Visualize the genetic distance and relatedness between the population)
- Data Reduction
- Data Classification
- Trend Analysis
- Factor Analysis
- Noise Reduction
- Face recognition
- Image compression
- Gene expression analysis

Example use of "PCA"

- How many unique “sub-sets” are in the sample?
- How are they similar / different?
- What are the underlying factors that influence the samples?
- Which time / temporal trends are (anti)correlated?
- Which measurements are needed to differentiate?
- How to best present what is “interesting”?
- Which “sub-set” does this new sample rightfully belong?

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

Geometric Rationale of PCA

- objects are represented as a cloud of *n* points in a multidimensional space with an axis for each of the *p* variables
- the **centroid** of the points is defined by the mean of each variable
- the **variance** of each variable is the average squared deviation of its *n* values around the mean of that variable.

$$V_i = \frac{1}{n-1} \sum_{m=1}^{n} \left(X_{im} - \overline{X}_i \right)^2$$

Geometric Rationale of PCA

degree to which the variables are linearly correlated is represented by their covariances.

$$C_{ij} = \frac{1}{n-1}\sum_{m=1}^{n}\left(X_{im} - \bar{X}_i\right)\left(X_{jm} - \bar{X}_i\right)$$

Covariance of variables *i* and *j*

Sum over all *n* objects

Value of variable *i* in object *m*

Mean of variable *i*

Value of variable *j* in object *m*

Mean of variable *j*

PCA is NOT…

1. Factor Analysis or Principal Coordinates Analysis (PCO)
2. A test of significance
3. No null hypothesis is required
4. Prior to ordination – no way to objectively decide which variables to include
5. After analysis – no way to decide which variables were unimportant
6. Cannot cope with missing values

Calculation in SPSS

Input Data

- should be quantitative (interval) or ordinal.

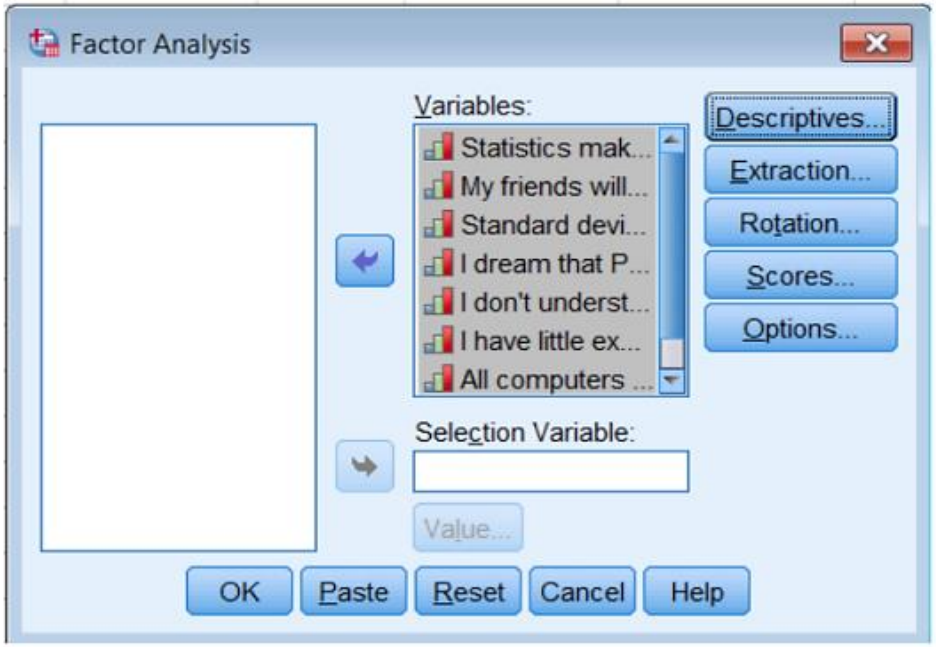
- should be linearly related to each other (checked by scatter plot of pairs of variables) or at least moderately correlated to each other.

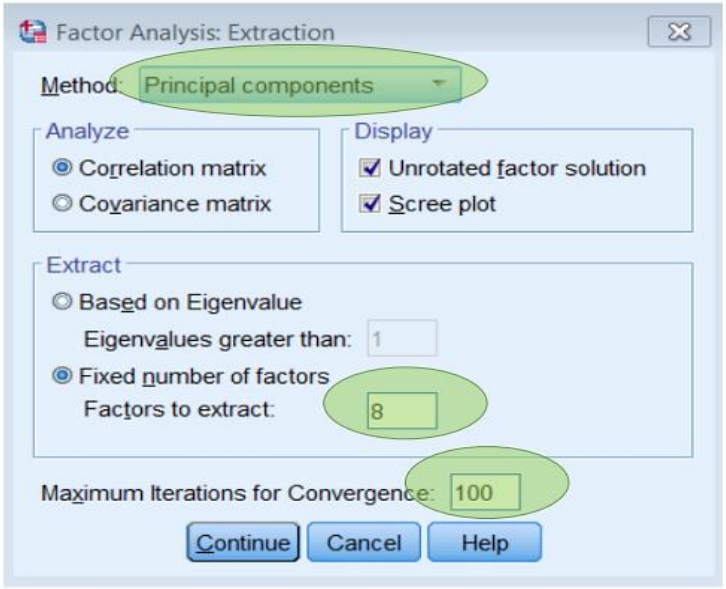
- minimum sample size requirement of 100.

- The ratio of cases to variables should be at least 5 to 1. Example: 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.

Calculation SPSS

Analyze – Dimension Reduction – Factor





Component Matrix of the 8-component PCA

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 squared loadings down each column (component) = eigenvalues

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

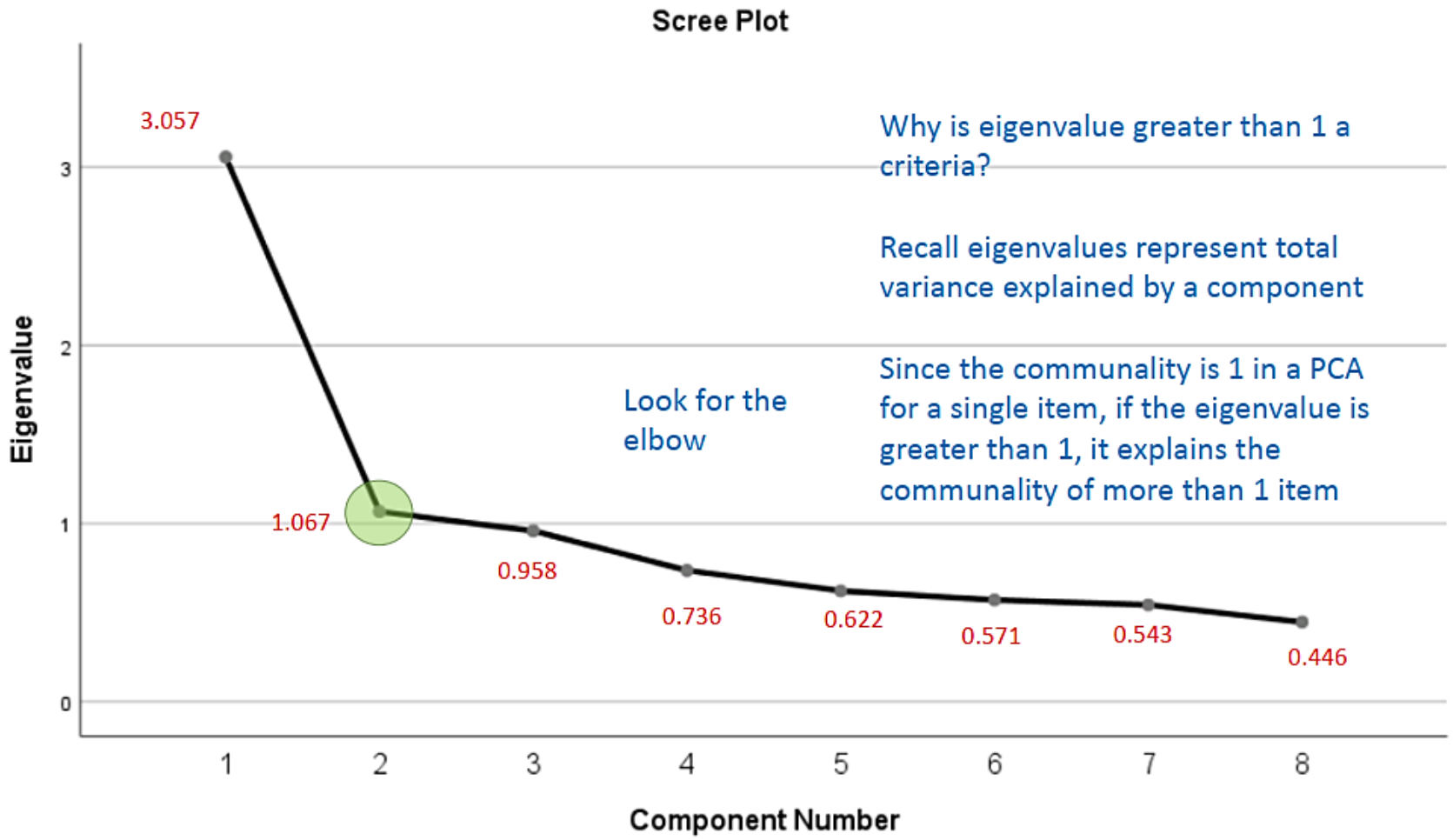
3.057 1.067 0.958 0.736 0.622 0.571 0.543 0.446

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	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


Scree Plot
Eigenvalue
Component Number
3.057
1.067
0.958
0.736
0.622
0.571
0.543
0.446
Why is eigenvalue greater than 1 a criteria?
Recall eigenvalues represent total variance explained by a component
Look for the elbow
Since the communality is 1 in a PCA for a single item, if the eigenvalue is greater than 1, it explains the communality of more than 1 item

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			

Extraction Method: Principal Component Analysis.

Notice only two eigenvalues

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.

How would you derive these communalities?

Dr. Myat Htut Nyunt
Email: myathtutnyunt@mohs.gov.mm; drmhnyunt@myanmarhsrj.com
HP: 0943106659