S1 Sains Data
FMIPA UNESA

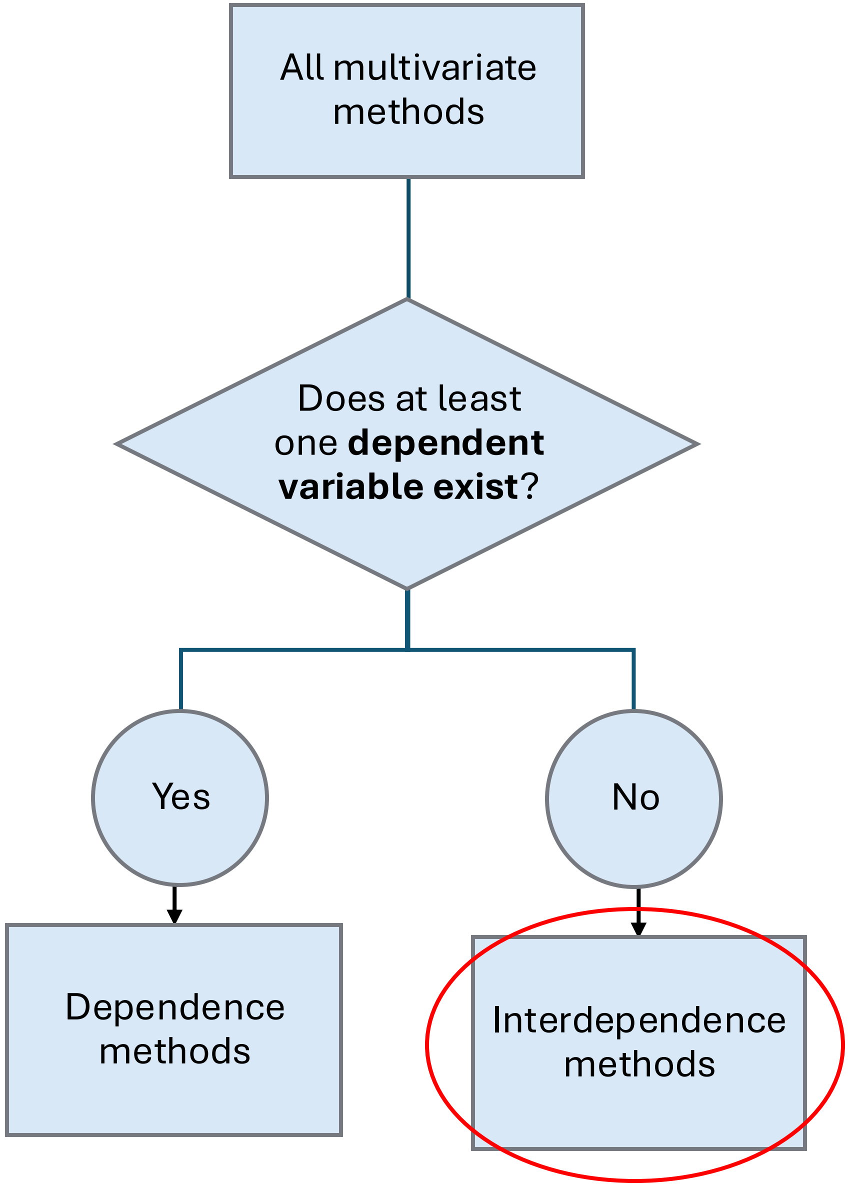
Analisis Multivariat – Materi 04

Principal Component Analysis

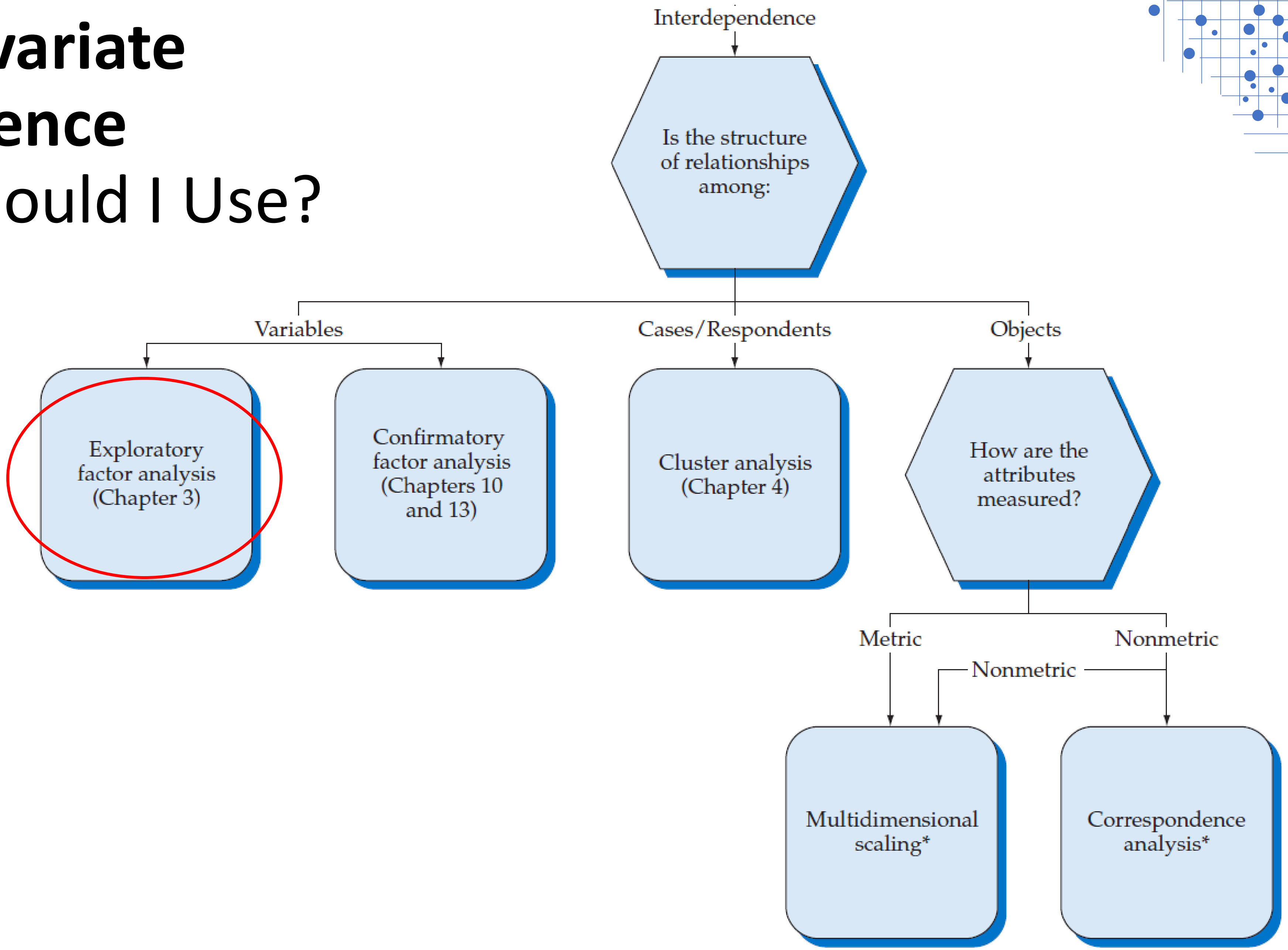
Prodi S1 Sains Data
Universitas Negeri Surabaya
2025

@sainsdata_unesa Sains Data UNESA datascience@unesa.ac.id https://datascience.fmipa.unesa.ac.id

Which **Multivariate Approach** is Appropriate?



Which **Multivariate Interdependence** Technique Should I Use?


Interdependence
Is the structure of relationships among:
Variables
Cases/Respondents
Objects
Exploratory factor analysis (Chapter 3)
Confirmatory factor analysis (Chapters 10 and 13)
Cluster analysis (Chapter 4)
How are the attributes measured?
Metric
Nonmetric
Nonmetric
Multidimensional scaling*
Correspondence analysis*

Basic Concept

- Mereduksi dimensi sejumlah (*p*) variabel asal (**X**) yang saling berkorelasi tinggi, dengan menguraikan sruktur Cov(**X**) atau Cov(**Z**) (=kor(**X**)) ke dalam sejumlah dimensi yang lebih kecil (*m*) variabel baru (***PC***, *principal component*).
- Variabel-variabel asal yang saling berkorelasi lebih tinggi dibandingkan variabel asal lainnya akan dikelompokkan ke dalam *PC* yang sama, **sedemikian hingga setiap *PC* mampu menjelaskan sebesar mungkin proporsi variabilitas dari total variabel asal dan antar *PC* independen.**

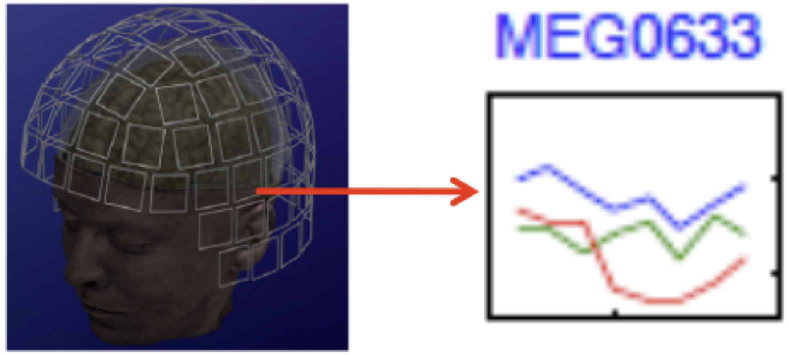
Mengapa
harus
direduksi?

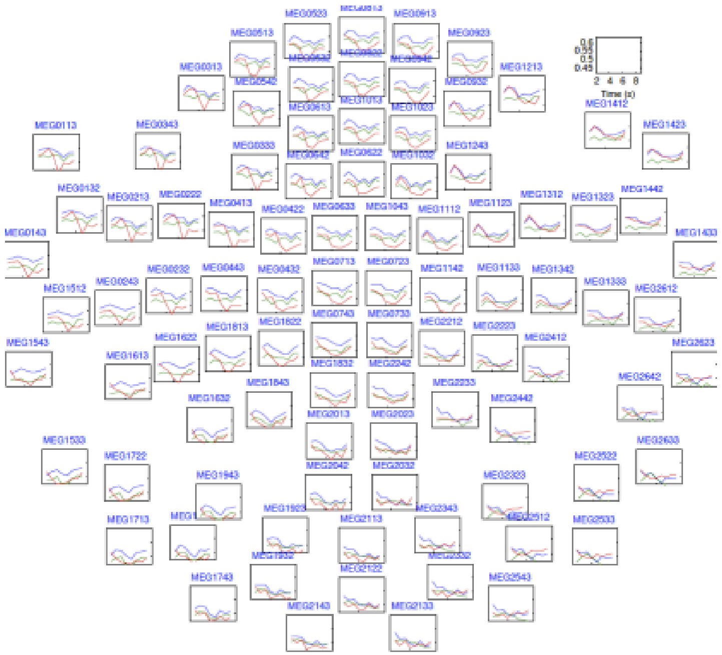
Big & High-Dimensional Data

- High-Dimensions = Lot of Features

MEG Brain Imaging

120 locations x 500 time points
x 20 objects





Or any high-dimensional image data

Big & High-Dimensional Data

- High-Dimensions = Lot of Features

Document classification

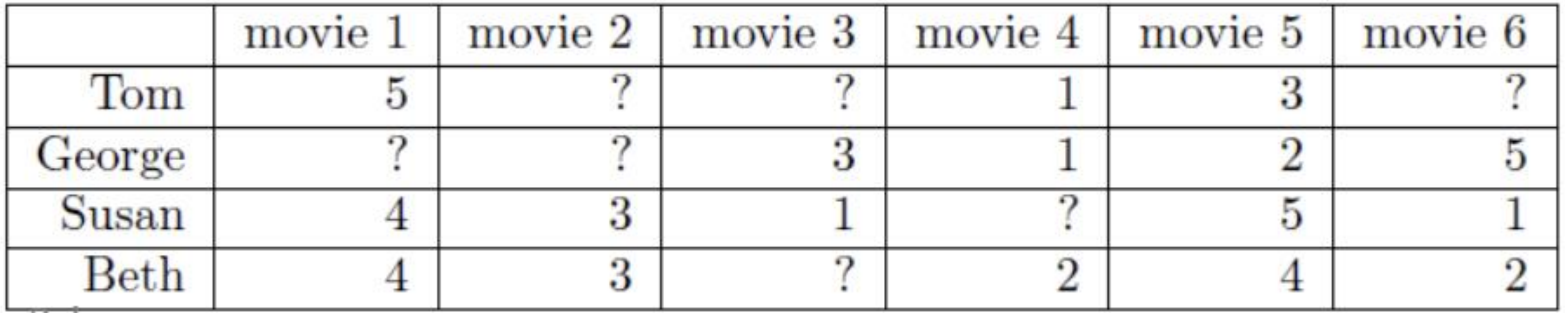
Features per document =

thousands of words/unigrams

millions of bigrams, contextual information

Surveys - Netflix

480189 users x 17770 movies



	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6
Tom	5	?	?	1	3	?
George	?	?	3	1	2	5
Susan	4	3	1	?	5	1
Beth	4	3	?	2	4	2

What if our data have way more than 3-dimensions?

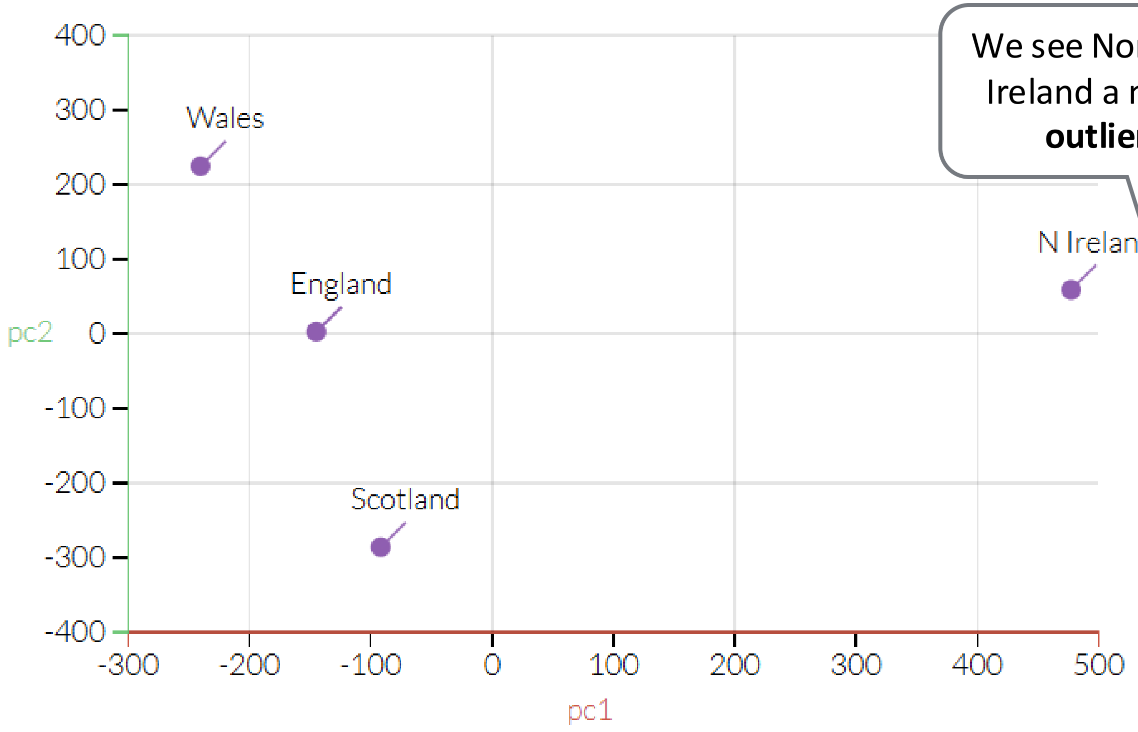
Can you spot the difference?

- The table shows some interesting variations across different food types, but **overall differences aren't so notable.**

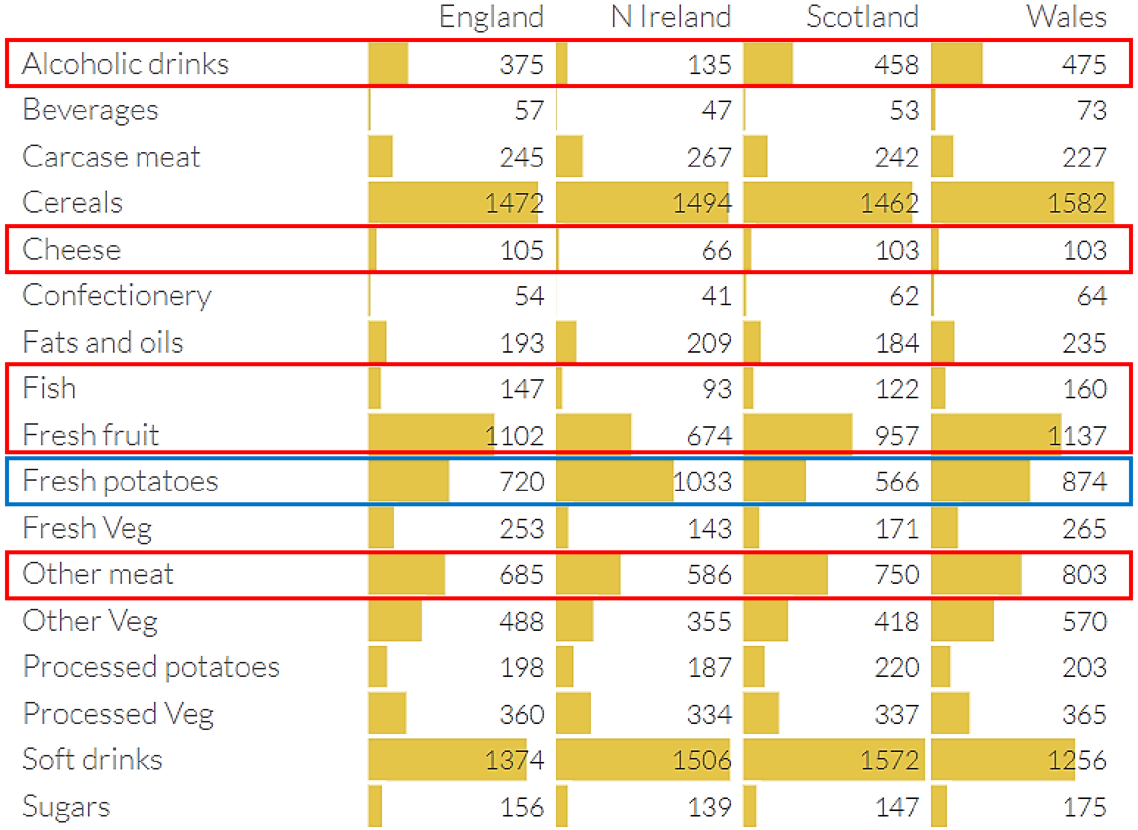
Table. The average consumption of **17 types of food** in grams per person per week for every country in the UK.

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

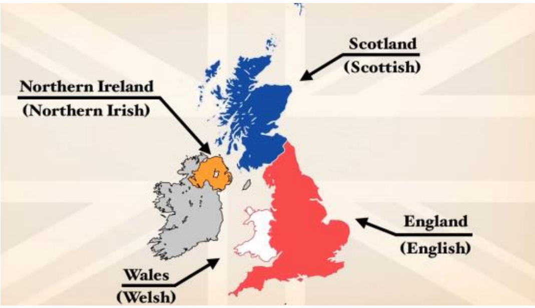
Let's see if **PCA** can eliminate dimensions to emphasize how countries differ.



We see Northern Ireland a major **outlier**.



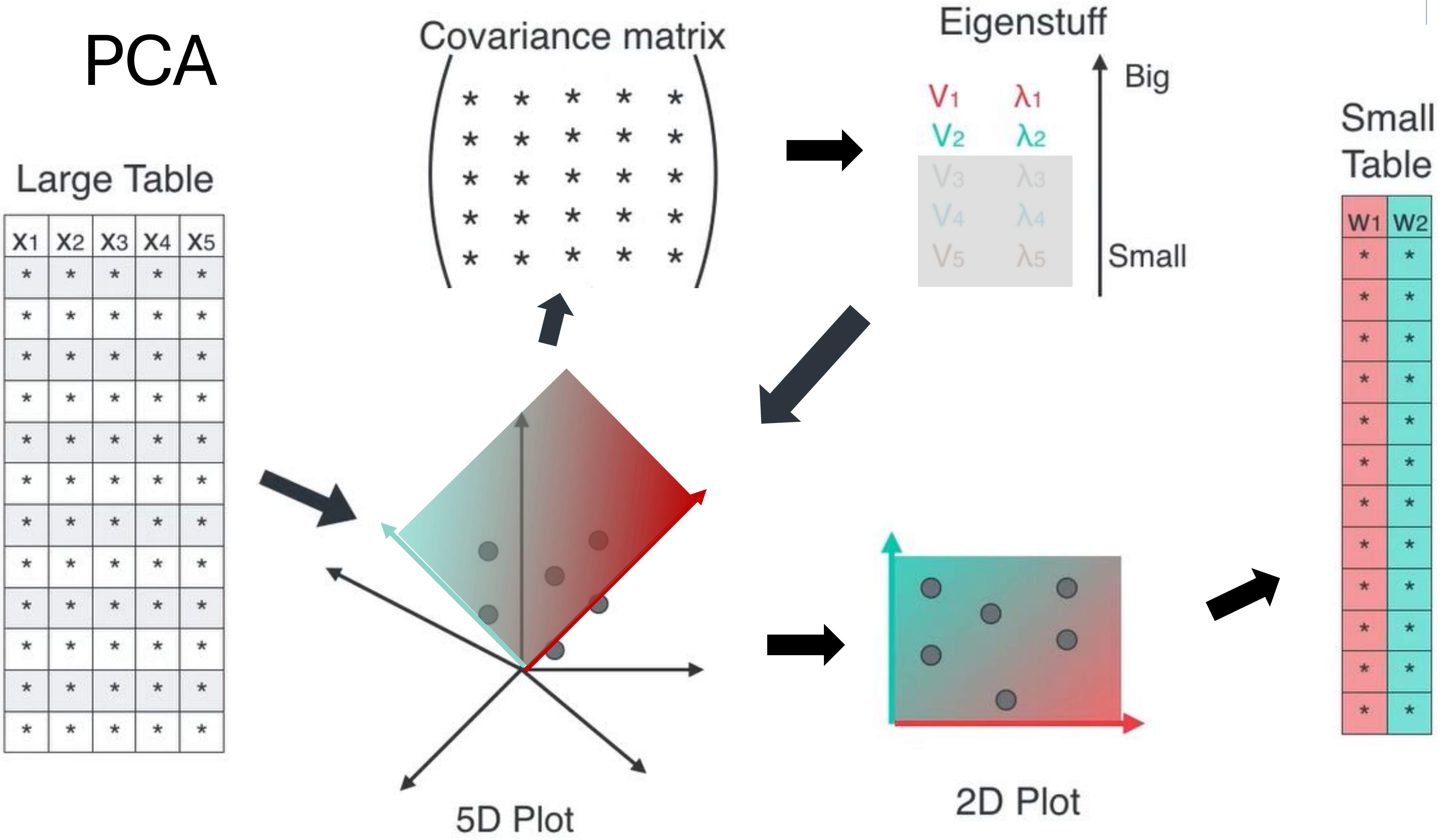
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



The fact: Northern Ireland is the only of the four countries not on the island of Great Britain (England, Scotland, Wales).

The Northern Irish eat way more grams of fresh potatoes and way fewer of fresh fruits, cheese, fish, other meat, and alcoholic drinks.

PCA



Large Table

X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Covariance matrix

$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

Eigenstuff

V_1	λ_1
V_2	λ_2
V_3	λ_3
V_4	λ_4
V_5	λ_5

Small Table

W1	W2
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*

source

PCA Procedure

Suppose that we have a random vector **X** with population variance-covariance matrix **∑.**

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \qquad \text{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

Let **∑** have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then consider the i^{th} principal component is given by.

$$Y_i = \boldsymbol{e_i'X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, i = 1, 2, \dots, p$$

$$\boxed{\begin{array}{rcl} Y_1 & = & e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 & = & e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ & & \vdots \\ Y_p & = & e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{array}} \qquad \mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}$$

Therefore, all principal components are those uncorrelated linear combinations $Y_1, Y_2, \dots, Y_p$ whose variances $Var(Y_i)$ are as large as possible.

with,

$$\text{var}(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{il}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_i = \lambda_i \qquad \mathbf{e}_i'\mathbf{e}_i = \sum_{j=1}^{p} e_{ij}^2 = 1$$

$$\text{cov}(Y_i, Y_j) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{jl}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_j = 0$$

How do we find the coefficients?

Let variance-covariance matrix **∑** have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

$$\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p$$

The elements for these **eigenvectors are the coefficients** of our principal components.

$$var(Y_i) = \mathrm{var}(e_{i1}X_1 + e_{i2}X_2 + \dots e_{ip}X_p) = \lambda_i$$

Total variation of **X** as the trace of the variance-covariance matrix.

Total population variance

$$\begin{aligned} trace(\Sigma) &= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned}$$

The *i*th principal component explains the following proportion of the total variation.

Proportion of total population variance

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Example

Calculate the percentage of variance and cumulative percentage of variance

Suppose we have the following variance-covariance matriks $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

Calculate the percentage of variance and cumulative percentage of variance.

Component	*Eigenvalue*
	Total
1	5.83
2	2.00
3	0.17

Example

Calculate the percentage of variance and cumulative percentage of variance

Suppose we have the following variance-covariance matriks $\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

Calculate the percentage of variance and cumulative percentage of variance.

Component	*Eigenvalues*		
	Total	*% of variance*	*Cumulative %*
1	5.83	0.73	0.73
2	2.00	0.25	0.98
3	0.17	0.02	1.00

Total population variance

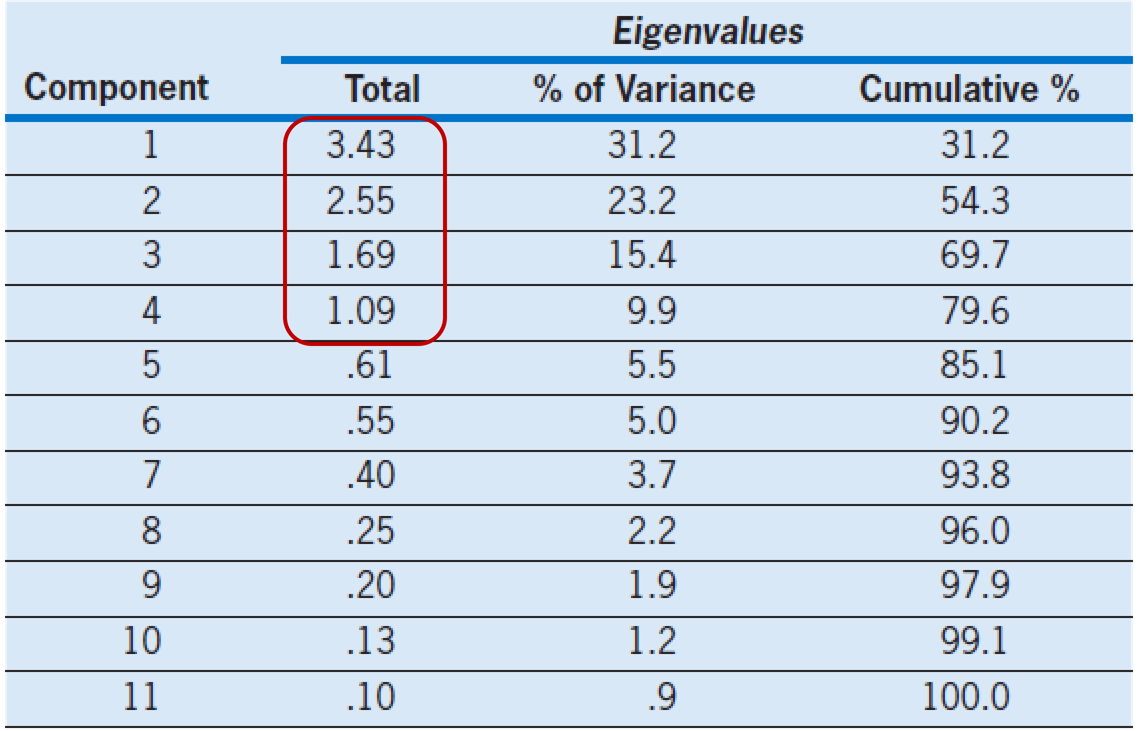
How many components to extract or retain?

Latent Root Criterion (Kaiser Rule)

- **Factors with eigenvalues greater than 1.0.**
- Working less accurately with a small number of variables or lower communalities.
- The most reliable when the number of variables is between 20 and 50 and communalities above 0.40.
  - If the number of variables is less than 20, the tendency is for this method to extract a conservative number of factors (too few).
  - In contrast, if more than 50 variables are involved, it is not uncommon for too many factors to be extracted.

The perceptions of HBAT on 13 variables are examined:

Tabel: Results for the extraction of component Factors



	Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	3.43	31.2	31.2
2	2.55	23.2	54.3
3	1.69	15.4	69.7
4	1.09	9.9	79.6
5	.61	5.5	85.1
6	.55	5.0	90.2
7	.40	3.7	93.8
8	.25	2.2	96.0
9	.20	1.9	97.9
10	.13	1.2	99.1
11	.10	.9	100.0

The four components will be retained based on latent root criterion

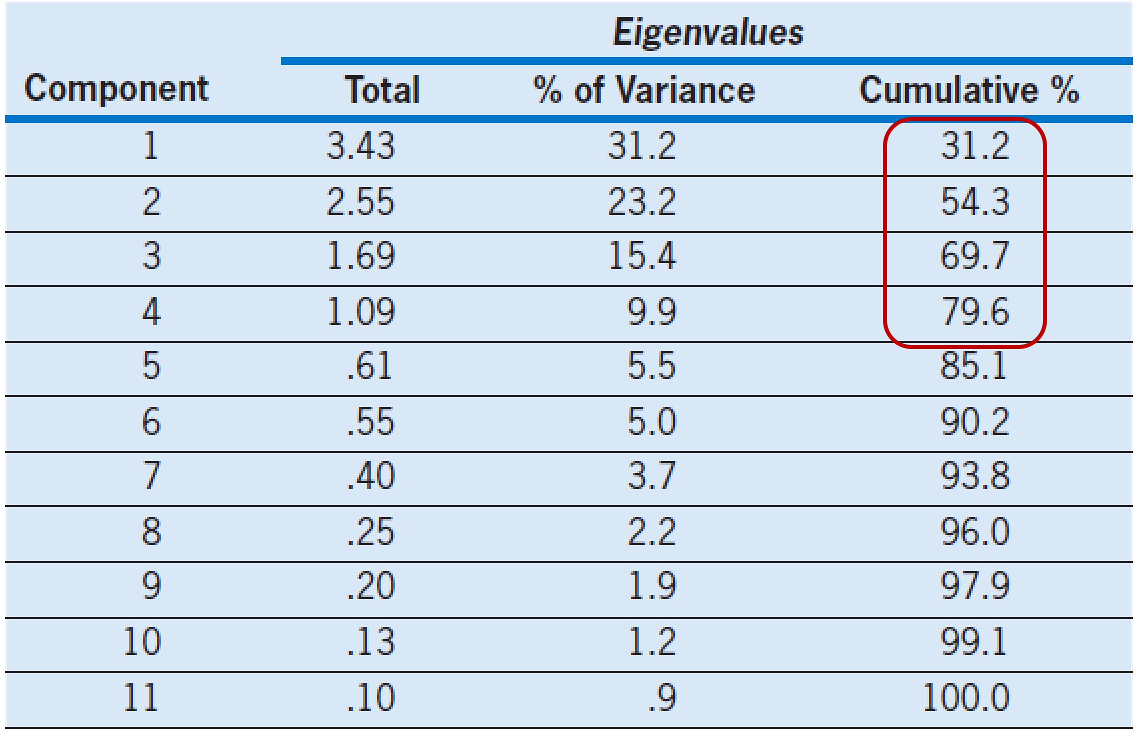
How many components to extract or retain?

Percentage of Variance Criterion, usually (60 % or higher) or (80 to 90%) .

- An approach based on achieving a specified cumulative percentage of total variance extracted by successive factors.

The perceptions of HBAT on 13 variables are examined:

Tabel. Results for the extraction of component Factors



	Eigenvalues		
Component	**Total**	**% of Variance**	**Cumulative %**
1	3.43	31.2	31.2
2	2.55	23.2	54.3
3	1.69	15.4	69.7
4	1.09	9.9	79.6
5	.61	5.5	85.1
6	.55	5.0	90.2
7	.40	3.7	93.8
8	.25	2.2	96.0
9	.20	1.9	97.9
10	.13	1.2	99.1
11	.10	.9	100.0

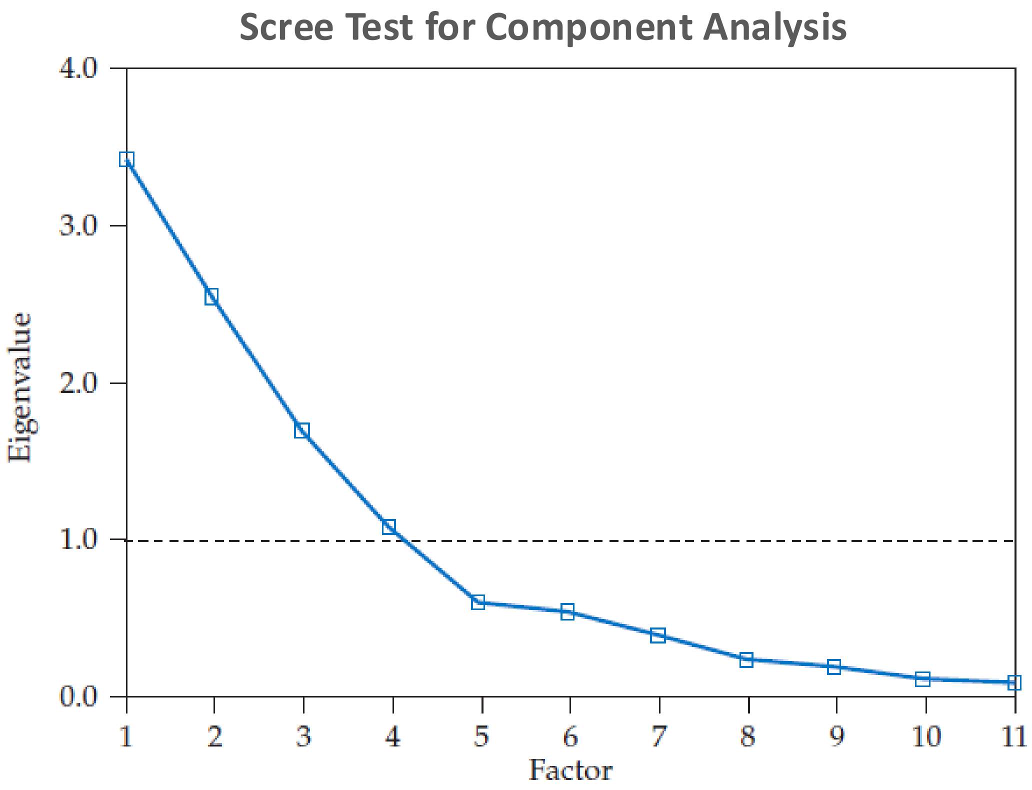
The **four-component** retained represent 79.6 % of the variance of the 11 variables, deemed sufficient in terms of total variance explained.

How many components to extract or retain?

Scree Test Criterion

- Identify the optimum number of factors that can be extracted before the amount of unique variance begins to dominate the common variance structure
- Derived by **plotting the latent roots against the number of factors** in their order of extraction
- The number of components is taken to be point at which the remaining eigenvalues are relatively small and all about the same size

The perceptions of HBAT on 13 variables are examined:



The scree test indicates that **four or perhaps five components** may be appropriate when considering the changes in eigenvalues

How many components to extract or retain?

A Priori Criterion

- The researcher already knows how many factors to extract before undertaking the factor analysis.
- A predetermined number of factors based on research objectives and/or prior research.
- This approach is useful when testing a theory or hypothesis about the number of factors to be extracted.

Parallel analysis

- To form a stopping rule based on the specific characteristics (i.e., **number of variables** and **sample size**) of the dataset being analyzed
- Each of these simulated datasets is then factor analyzed, either with principal components or common factor methods, and the eigenvalues are averaged for each factor across all the datasets.
  - The result is the average eigenvalue for the first factor, the second factor and so on across the set of simulated datasets.
  - These values are then compared to the eigenvalues extracted for the original data and all factors with eigenvalues above those of the simulated datasets are retained.

Correlation between Variables and Principal Components

Let consider the i^{th} principal component is given by.

$$Y_i = \boldsymbol{e}_i' \boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, i = 1, 2, \ldots, p$$

So, we have the **correlation coefficients** between the components *Yi* and the variables *Xk*.

$$\rho_{Y_i, X_k} = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}}$$

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, where\ i, k = 1, 2, \ldots, p$$

- This **correlation** measure only the univariate contribution of an individual *X* to a component *Y*.
- They do not indicate the importance of an *X* to a component *Y*.
- The **coefficients** $\boldsymbol{e_{ik}}$ be used to interpret the components.
- We recommend that **both measure** be examined to help interpret the principal components.

<u>proof</u>

$Cov(X_k, Y_i) = Cov(\boldsymbol{a}_k'\boldsymbol{X}, \boldsymbol{e}_i'\boldsymbol{X}) = \boldsymbol{a}_k'\boldsymbol{\Sigma}\boldsymbol{e}_i$

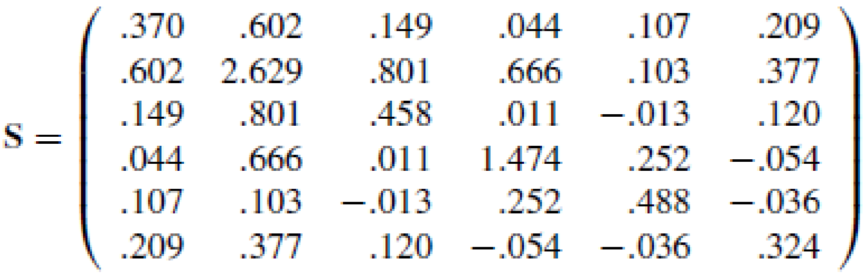
since $\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$

so $Cov(X_k, Y_i) = \boldsymbol{a}_k'\lambda_i\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_{ik}$

Example

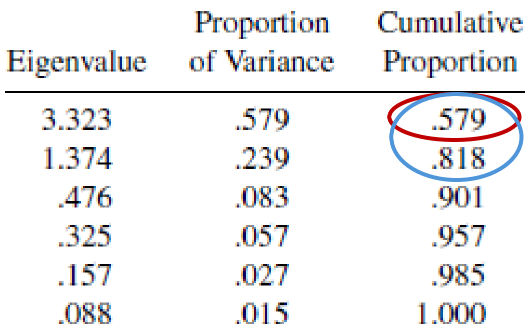
Correlation between Variables and Principal Components

Suppose we have the following variance-covariance matrix *S*



$$\mathbf{S} = \begin{pmatrix} .370 & .602 & .149 & .044 & .107 & .209 \\ .602 & 2.629 & .801 & .666 & .103 & .377 \\ .149 & .801 & .458 & .011 & -.013 & .120 \\ .044 & .666 & .011 & 1.474 & .252 & -.054 \\ .107 & .103 & -.013 & .252 & .488 & -.036 \\ .209 & .377 & .120 & -.054 & -.036 & .324 \end{pmatrix}$$

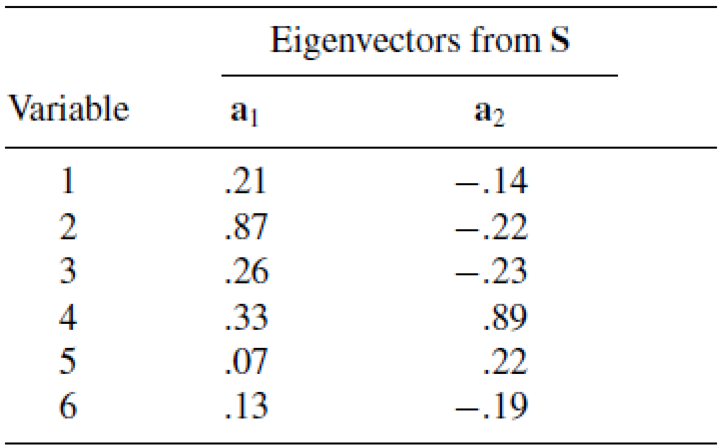
Then we calculate the eigenvalue



Eigenvalue	Proportion of Variance	Cumulative Proportion
3.323	.579	.579
1.374	.239	.818
.476	.083	.901
.325	.057	.957
.157	.027	.985
.088	.015	1.000

- The first principal component explains 57.9% of the total variance.
- The first two principal components, collectively, explain 81.8% of the total variance.
- *Consequently, sample variation is summarized very well by two pc*

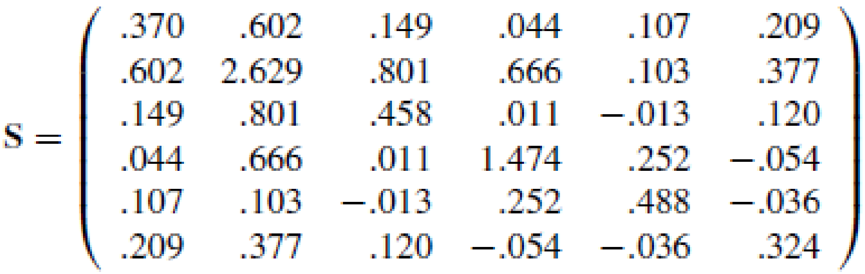
Table. Eigenvectors



	Eigenvectors from S	
Variable	$\mathbf{a}_1$	$\mathbf{a}_2$
1	.21	−.14
2	.87	−.22
3	.26	−.23
4	.33	.89
5	.07	.22
6	.13	−.19

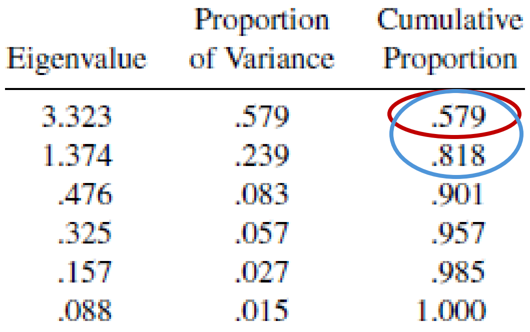
Example
Correlation between Variables and Principal Components

Suppose we have the following variance-covariance matrix *S*



$$
\mathbf{S} = \begin{pmatrix}
.370 & .602 & .149 & .044 & .107 & .209 \\
.602 & 2.629 & .801 & .666 & .103 & .377 \\
.149 & .801 & .458 & .011 & -.013 & .120 \\
.044 & .666 & .011 & 1.474 & .252 & -.054 \\
.107 & .103 & -.013 & .252 & .488 & -.036 \\
.209 & .377 & .120 & -.054 & -.036 & .324
\end{pmatrix}
$$

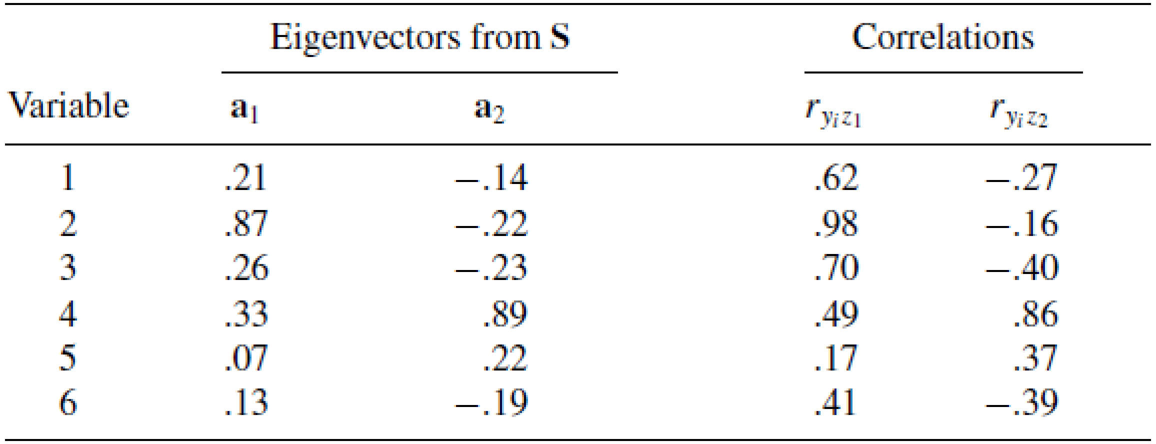
Then we calculate the eigenvalue



Eigenvalue	Proportion of Variance	Cumulative Proportion
3.323	.579	.579
1.374	.239	.818
.476	.083	.901
.325	.057	.957
.157	.027	.985
.088	.015	1.000

- The first principal component explains 57.9% of the total variance.
- The first two principal components, collectively, explain 81.8% of the total variance.
- *Consequently, sample variation is summarized very well by two pc*

Table. Eigenvectors



	Eigenvectors from S		Correlations	
Variable	$\mathbf{a}_1$	$\mathbf{a}_2$	$r_{y_i z_1}$	$r_{y_i z_2}$
1	.21	−.14	.62	−.27
2	.87	−.22	.98	−.16
3	.26	−.23	.70	−.40
4	.33	.89	.49	.86
5	.07	.22	.17	.37
6	.13	−.19	.41	−.39

$$\rho_{Y_i,Z_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

Principal Components from The Correlation Matrix

- Generally, extracting components from **S** rather than **R** remains closer to the spirit and intent of principal component analysis, especially if the components are to be used in further computations.
- However, in some cases, the principal components will be more interpretable if **R** is used.
- For example, if the variances differ widely or if the measurement units are not commensurate, the components of **S** will be dominated by the variables with large variances. The other variables will contribute very little.

Principal components may also be obtained for the standardized variables

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$

$$\vdots$$

$$Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

In matrix notation,

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu})$$

where $\mathbf{V}^{1/2}$ the diagonal standard deviation matrix and

E(**Z**) = 0 and

$$\text{Cov}\ (\mathbf{Z}) = (\mathbf{V}^{1/2})^{-1}\ \boldsymbol{\Sigma}\ (\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho}$$

Principal Components from The Correlation Matrix

The principal components of Z may be obtained from the eigenvectors of the correlation matrix **ρ** of **X**. The ith principal component of the standardized variables $Z = [Z_1, Z_2, \dots, Z_p]$ with Cov(Z) = **ρ**, is given by

$$\boxed{Y_i = \boldsymbol{e}_i'\boldsymbol{Z} = \boldsymbol{e}_i'\left(\boldsymbol{V}^{1/2}\right)^{-1}(\boldsymbol{X} - \boldsymbol{\mu}), i = 1, 2, \dots, p}$$

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$$

$$\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i}, i, k = 1, 2, \dots, p$$

In this case $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ are the eigenvalue-eigenvector pairs for **ρ** where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

$$\begin{pmatrix} proporsi\ (distandarisasi) \\ varians\ populasi\ karena \\ komponen\ utama\ ke - k \end{pmatrix} = \frac{\lambda_k}{p}, k = 1, 2, \dots, p$$

Example

Principal Components obtained from Correlation Matrices

Consider the covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ and the derived correlation matrix $\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$

We can calculate

eigenvalue-eigenvector from Σ

$\lambda_1 = 100.16$ with $e_1^1 = [0.040, 0.999]$

$\lambda_2 = 0.84$ with $e_2^1 = [0.999, -0.040]$

The principal components become

$Y_1 = 0.040X_1 + 0.999X_2$

$Y_2 = 0.999X_1 - 0.040X_2$

X2 dominates the first principal component. Moreover, the 1st pc explains a proportion 0.992 of total variance.

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{100.16 + 0.84} = 0.992$$

Example

Principal Components obtained from Correlation Matrices

Consider the covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ and the derived correlation matrix $\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$

We can calculate

Eigenvalue-eigenvector from $\boldsymbol{\rho}$

$\lambda_1 = 1.4 \text{ with } e_1^1 = [0.707, 0.707]$

$\lambda_2 = 0.6 \text{ with } e_2^1 = [0.707, -0.707]$

The principal components become

$$Y_1 = 0.707Z_1 + 0.707Z_2 = 0.707\left(\frac{X_1 - \mu_1}{1}\right) + 0.707\left(\frac{X_2 - \mu_2}{10}\right)$$
$$= 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2)$$
$$Y_2 = 0.707Z_1 - 0.707Z_2 = 0.707\left(\frac{X_1 - \mu_1}{1}\right) - 0.707\left(\frac{X_2 - \mu_2}{10}\right)$$
$$= 0.707(X_1 - \mu_1) - 0.0707(X_2 - \mu_2)$$

The resulting variables contribute equally to the principal components determined from $\boldsymbol{\rho}$

$$\rho_{Y_1,Z_1} = e_{11}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$$
$$\rho_{Y_1,Z_2} = e_{21}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$$

The 1st pc explains a proportion 0.992 of total variance

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = 0.7$$

of the total (standardized) variance

When the 1st principal component obtained from $\boldsymbol{\rho}$ is expressed in terms of X_1 and X_2, the relative magnitude of the weights 0.707 and 0.0707 are in direct opposition to those of the weights 0.040 and 0.999 attached to these variables in the principal component obtained from $\boldsymbol{\Sigma}$

Exercise

Principal Components from The Covariance Matrix (Z) Standardized Data

The weekly temperatures (in degrees Celsius) for five cities (Jakarta, Surabaya, Bandung, Medan, and Makassar) were recorded from January 2004 through December 2005. The temperature for each week is defined as the average of daily recorded temperatures for that week. The observations for 103 consecutive weeks appear to be independently distributed, but the temperatures across cities are correlated because, as expected, cities in the same region tend to experience similar weather patterns due to seasonal and geographical influences.

Exercise

Principal Components from The Covariance Matrix (Z) Standardized Data

We note the covariance matrix from the standardized data.

$$\Sigma = \begin{bmatrix} 1.000 & 0.632 & 0.511 & 0.115 & 0.155 \\ 0.632 & 1.000 & 0.574 & 0.322 & 0.213 \\ 0.511 & 0.574 & 1.000 & 0.183 & 0.146 \\ 0.115 & 0.322 & 0.183 & 1.000 & 0.683 \\ 0.155 & 0.2133 & 0.146 & 0.683 & 1.000 \end{bmatrix}$$

$$Z_1 = \frac{(X_1 - \overline{X_1})}{\sqrt{s_{11}}}$$

$$Z_1 = \frac{(X_1 - \overline{X_2})}{\sqrt{s_{22}}}$$

$$\vdots$$

$$Z_5 = \frac{(X_1 - \overline{X_5})}{\sqrt{s_{55}}}$$

$$\hat{e}^5 = [\ 0.384, \quad -0.496, \quad 0.071, \quad 0.595, \quad -0.498\ ]\hat{\lambda}^5 = 0.255,$$

$$\hat{e}^4 = [\ 0.363, \quad -0.629, \quad 0.289, \quad -0.381, \quad 0.493\ ]\hat{\lambda}^4 = 0.400,$$

$$\hat{e}^3 = [-0.604, \quad -0.136, \quad 0.772, \quad 0.093, \quad -0.109\ ]\hat{\lambda}^3 = 0.501,$$

$$\hat{e}^2 = [-0.368, \quad -0.236, \quad -0.315, \quad 0.585, \quad 0.606\ ]\hat{\lambda}^2 = 1.407$$

$$\hat{e}_1 = [\ 0.469, \quad 0.532, \quad 0.465, \quad 0.387, \quad 0.361\ ]\hat{\lambda}^1 = 2.437,$$

Calculate the principal components, percentage of variance, correlation between variables and pc!

Fungsi PCA

1. **Metode statistik yang berdiri sendiri**
   - Penentuan m kelompok variabel baru (PC) dominan (proporsi variabilitas yang dijelaskan sebesar mungkin dari p variabel asal) dan saling independen . Dalam hal ini , keanggotaan variabel asal pada kelompok variabel baru tertentu , pada perkembangannya diperjelas dikonfirmasi oleh metode Analisis Faktor (dengan rotasi)
   - Pengelompokan observasi secara visual melalui skor PCA

2. **Intermediate Step**
   - sebagai penghasil input metode statistik lainnya , misalnya dalam rangka untuk pemenuhan asumsi metode statistik lainnya (Regresi Linier, tidak terjadi multikolinieritas atau saling independen antar variabel prediktor , p<n)

Principal Component Analysis Decision Process

- **Stage 1:** make objective
- **Stage 2:** design
- **Stage 3:** assumptions

Principal component analysis: stages 4–7

- **Stage 4:** deriving factors and assessing overall fit
- **Stage 5:** interpreting the factors
- **Stage 6:** validation of principal components analysis
- **Stage 7:** additional uses of the exploratory factor analysis results

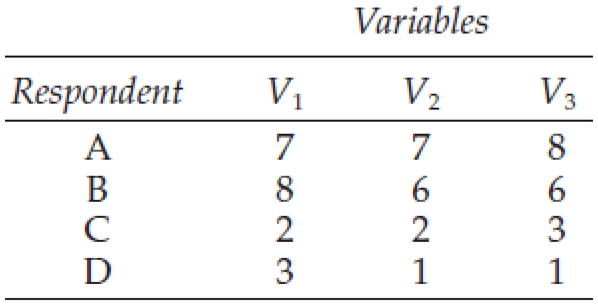
Stage 1: make objective

The perceptions of HBAT on 13 attributes (X6 to X18) are examined for the following reasons

- **Data Summarization with Interpretation**—Understand whether these perceptions can be “grouped”. By grouping the perceptions and then engaging in the steps of factor interpretation, HBAT will be able to see the big picture in terms of understanding its customers and what dimensions the customers think about HBAT.
- **Data Reduction**—If the 13 variables can be represented in a smaller number of composite variables, then the other multivariate techniques can be made more parsimonious.

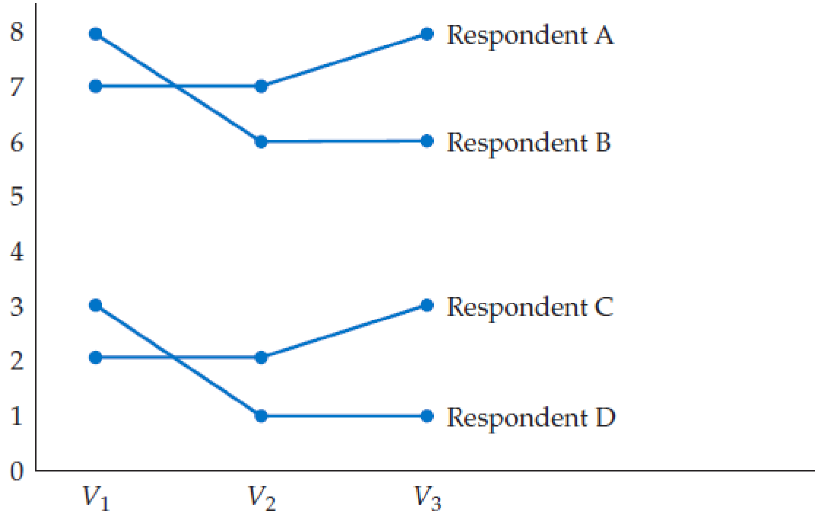
Stage 2: design

- **Variable selection and measurement issues**
  - what types of variables can be used?
  - how many variables should be included?
- **Sample size**
  - The sample must have more observations than variables.
  - The minimum absolute sample size should be 50 observations, with 100 observations the preferred minimum.
  - Increase the sample as the complexity of the factor analysis increases (i.e., number of variables and/or factors retained).
  - Strive to maximize the number of observations per variable, with a desired ratio of at least 5 Observations per variable.
  - Higher communalities among the variables provides support for smaller samples sizes, all other things equal.
- **Correlations among variables or respondents**
  - R-type factor analysis, use a traditional correlation matrix (correlations among variables) as input
  - Q-type factor analysis, the results would be a factor matrix that would identify similar individuals.



	Variables		
Respondent	V_1	V_2	V_3
A	7	7	8
B	8	6	6
C	2	2	3
D	3	1	1

comparisons of *Q*-type Factor Analysis and cluster Analysis



Stage 2: design

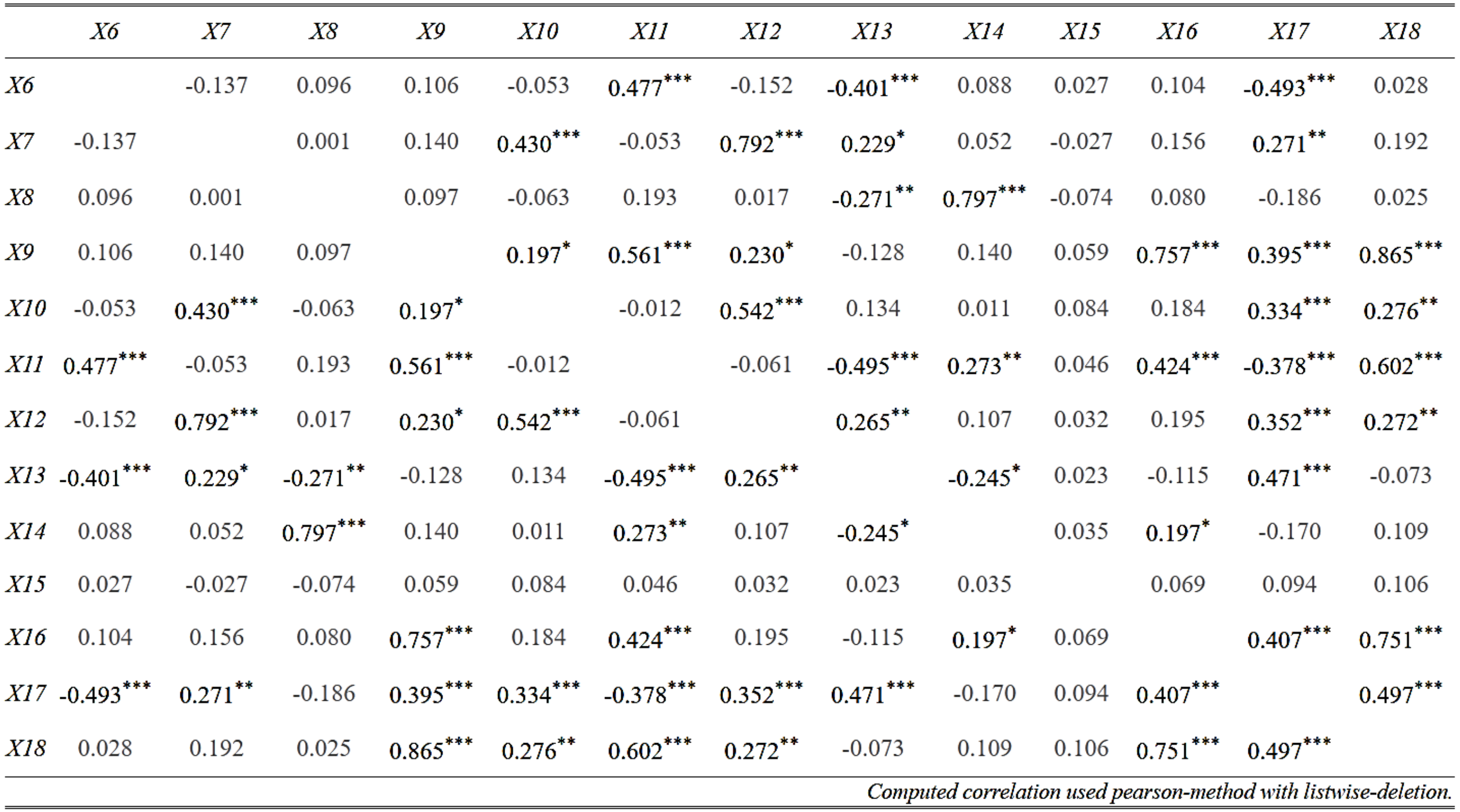
- Understanding the structure of the perceptions of variables requires **R-type factor analysis** and a correlation matrix between variables, not respondents.
- All the variables are metric and constitute a homogeneous set of perceptions appropriate for exploratory factor analysis.
- The sample size in this example is an 7:1 ratio of observations to variables, which falls within acceptable limits.
- Also, the **sample size of 100** provides an adequate basis for the calculation of the correlations between variables.

Stage 3: assumptions

- Visual examination of the correlation
- A statistically significant Bartlett's test of sphericity (sig. , 0.50) indicates that sufficient correlations exist among the variables to proceed.
- **Measure of sampling adequacy (MSA) values must exceed 0.50** for both the overall test and each individual variable;
  - variables with values less than 0.50 should be omitted from the factor analysis one at a time, with the smallest one being omitted each time.

Stage 3: assumptions

Correlations Among Variables



	X6	*X7*	*X8*	*X9*	*X10*	*X11*	*X12*	*X13*	*X14*	*X15*	*X16*	*X17*	*X18*
X6		-0.137	0.096	0.106	-0.053	0.477***	-0.152	-0.401***	0.088	0.027	0.104	-0.493***	0.028
X7	-0.137		0.001	0.140	0.430***	-0.053	0.792***	0.229*	0.052	-0.027	0.156	0.271**	0.192
X8	0.096	0.001		0.097	-0.063	0.193	0.017	-0.271**	0.797***	-0.074	0.080	-0.186	0.025
X9	0.106	0.140	0.097		0.197*	0.561***	0.230*	-0.128	0.140	0.059	0.757***	0.395***	0.865***
X10	-0.053	0.430***	-0.063	0.197*		-0.012	0.542***	0.134	0.011	0.084	0.184	0.334***	0.276**
X11	0.477***	-0.053	0.193	0.561***	-0.012		-0.061	-0.495***	0.273**	0.046	0.424***	-0.378***	0.602***
X12	-0.152	0.792***	0.017	0.230*	0.542***	-0.061		0.265**	0.107	0.032	0.195	0.352***	0.272**
X13	-0.401***	0.229*	-0.271**	-0.128	0.134	-0.495***	0.265**		-0.245*	0.023	-0.115	0.471***	-0.073
X14	0.088	0.052	0.797***	0.140	0.011	0.273**	0.107	-0.245*		0.035	0.197*	-0.170	0.109
X15	0.027	-0.027	-0.074	0.059	0.084	0.046	0.032	0.023	0.035		0.069	0.094	0.106
X16	0.104	0.156	0.080	0.757***	0.184	0.424***	0.195	-0.115	0.197*	0.069		0.407***	0.751***
X17	-0.493***	0.271**	-0.186	0.395***	0.334***	-0.378***	0.352***	0.471***	-0.170	0.094	0.407***		0.497***
X18	0.028	0.192	0.025	0.865***	0.276**	0.602***	0.272**	-0.073	0.109	0.106	0.751***	0.497***	

Computed correlation used pearson-method with listwise-deletion.

```
mat_corr <- round(cor(data_pca),3)
library(sjPlot)tab_corr(data_pca)
```

- If **visual inspection** reveals a small number of correlations among the variables greater than 30%, then exploratory factor analysis is probably inappropriate.
- 29 of the 78 correlations (37%) are significant at the .01 level, which provides an adequate basis for proceeding to an empirical examination of adequacy for factor analysis on both an overall basis and for each variable.

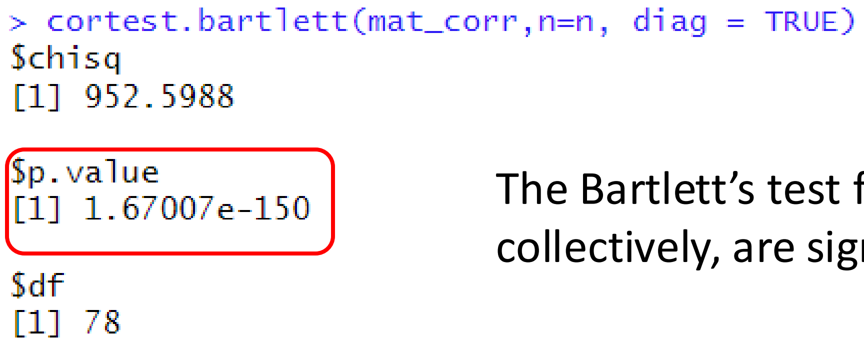
Stage 3: assumptions

Bartlett Test

- The **Bartlett test of sphericity** is a statistical test for the presence of correlations among the variables.
- A statistically significant Bartlett's test of sphericity (sig. , 0.50) indicates that sufficient correlations exist among the variables to proceed.

```
mat_corr <- round(cor(data_pca),3)
#Perform Bartlett test
library(psych)
n = nrow(data_pca)
p = ncol(data_pca)
cortest.bartlett(mat_corr,n=n, diag = TRUE)
```

$$\chi^2_{obs} = -\left[(n-1) - \frac{(2p+5)}{6}\right]\ln|R|$$



```
> cortest.bartlett(mat_corr,n=n, diag = TRUE)
$chisq
[1] 952.5988

$p.value
[1] 1.67007e-150

$df
[1] 78
```

The Bartlett’s test finds that the correlations, when taken collectively, are significant at the 0.05 level

Stage 3: assumptions

Measure of Sampling Adequacy (MSA)

- A measure to quantify the degree of intercorrelations among the variables and the appropriateness of exploratory factor analysis.
- The MSA increases as
  - the sample size increases
  - the average correlations increase
  - the number of variables increases
  - the number of factors decreases.
- The result should always have **an overall MSA value of above 0.50** before proceeding with the factor analysis. If the MSA value falls below 0.50, then the variable-specific MSA values can identify variables for deletion to achieve an overall value of .50.

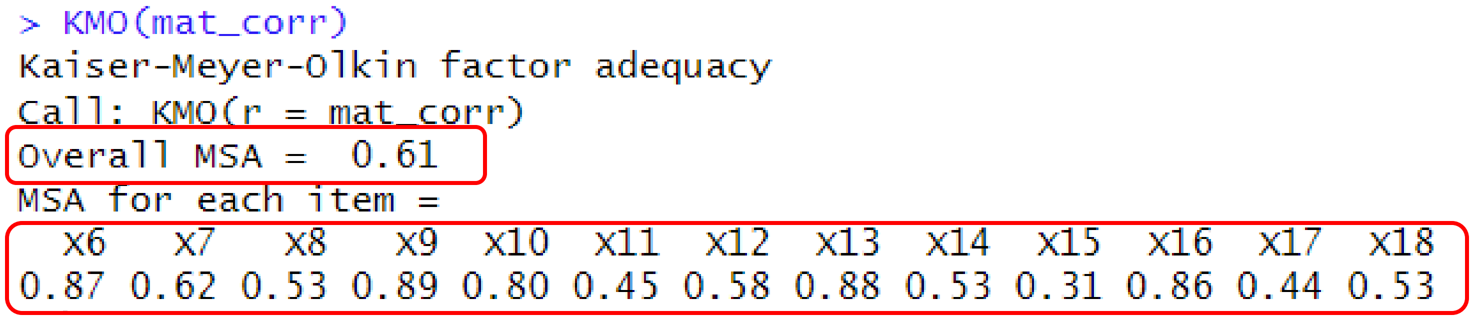
This index ranges from 0 to 1

0.80 or above	: meritorious
0.70 or above	: middling
0.60 or above	: mediocre
0.50 or above	: miserable
below 0.50	: unacceptable

```
mat_corr <- round(cor(data_pca),3)
KMO(mat_corr)
```

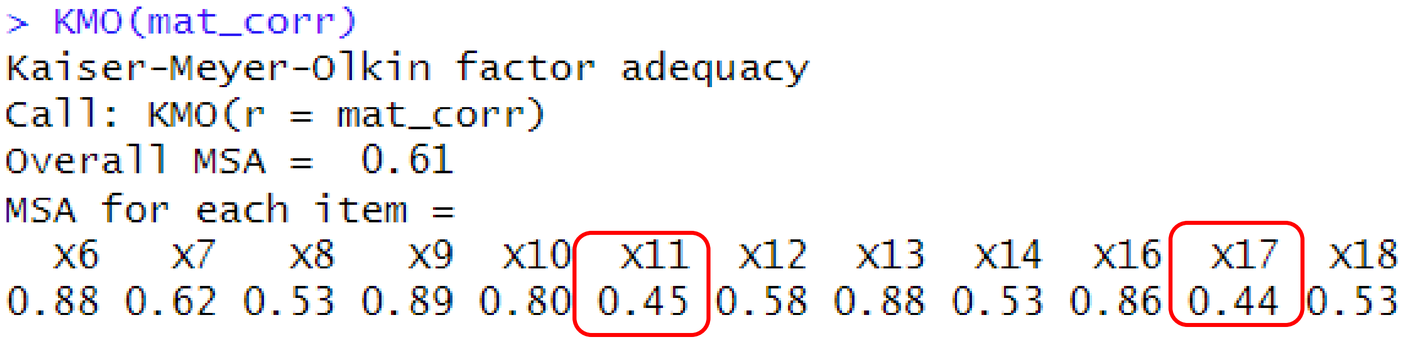
Stage 3: assumptions

Measure of Sampling Adequacy (MSA)



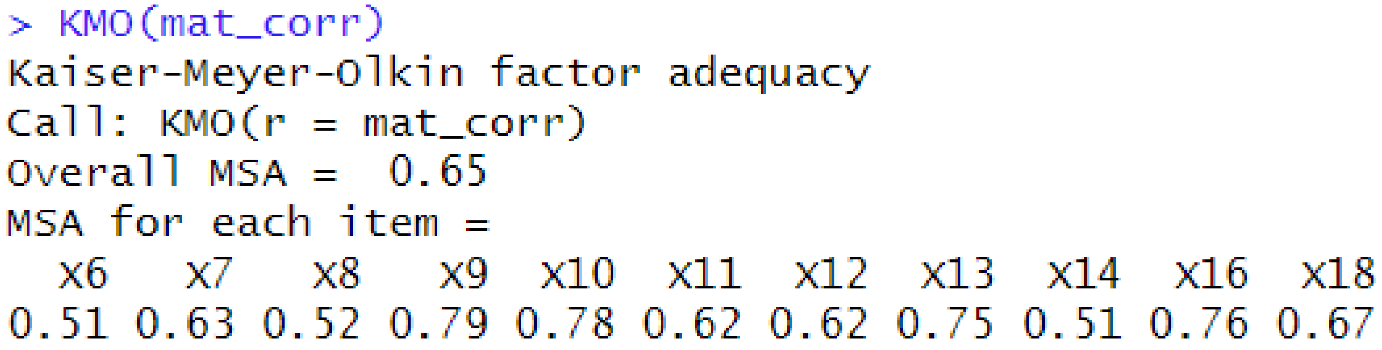
```
> KMO(mat_corr)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = mat_corr)
Overall MSA =  0.61
MSA for each item =
  x6   x7   x8   x9  x10  x11  x12  x13  x14  x15  x16  x17  x18
0.87 0.62 0.53 0.89 0.80 0.45 0.58 0.88 0.53 0.31 0.86 0.44 0.53
```

- Overall MSA value falls in the acceptable range (above 0.50)
- Three variables (X11, X15, and X17) with MSA values under 0.50.
- Because X15 has the lowest MSA value, it will be omitted to obtain a set of variables that can exceed the minimum acceptable MSA levels.
- Recalculating the MSA values after excluding X15



```
> KMO(mat_corr)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = mat_corr)
Overall MSA =  0.61
MSA for each item =
  x6   x7   x8   x9  x10  x11  x12  x13  x14  x16  x17  x18
0.88 0.62 0.53 0.89 0.80 0.45 0.58 0.88 0.53 0.86 0.44 0.53
```

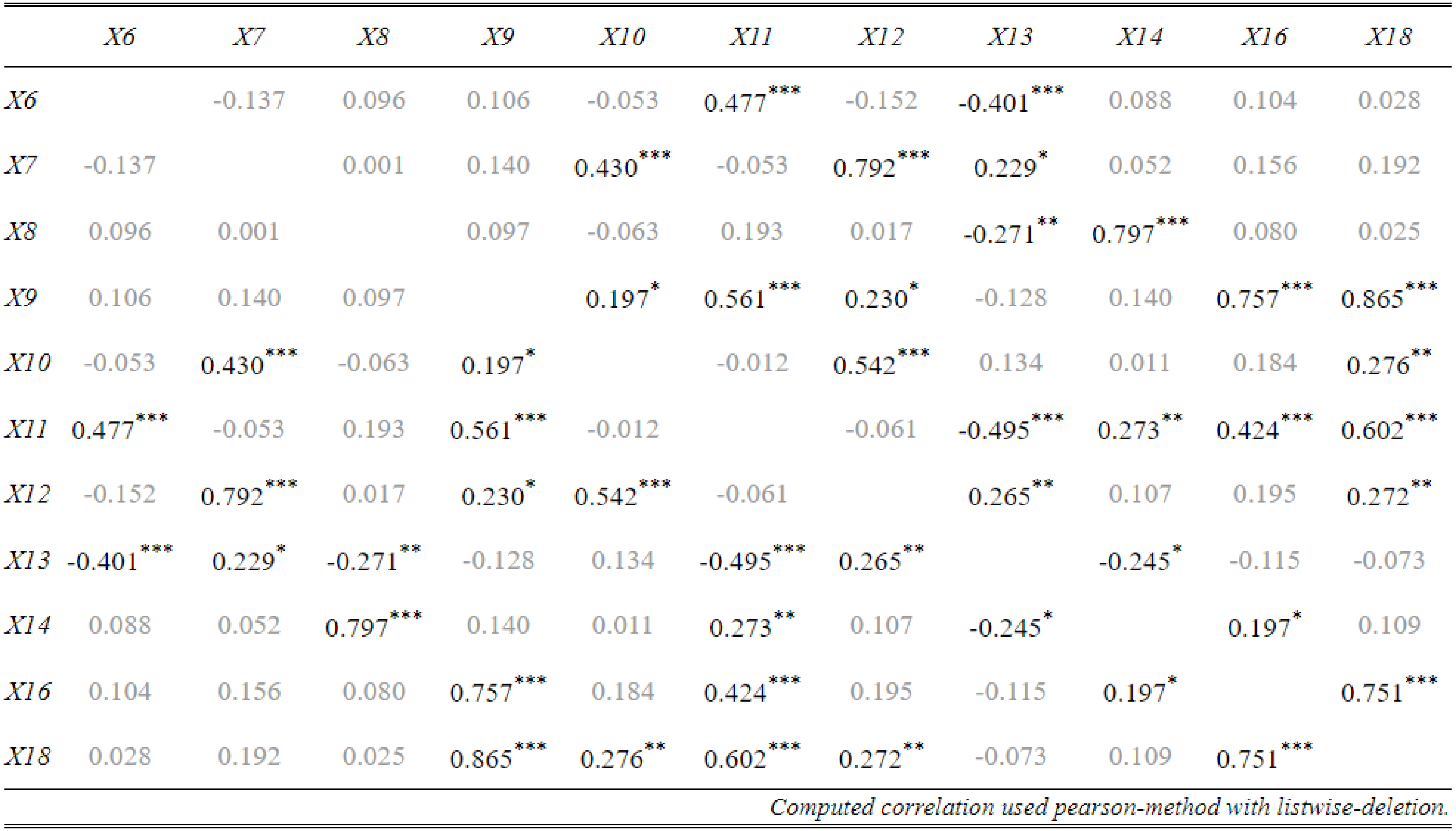
- X15 deleted



```
> KMO(mat_corr)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = mat_corr)
Overall MSA =  0.65
MSA for each item =
  x6   x7   x8   x9  x10  x11  x12  x13  x14  x16  x18
0.51 0.63 0.52 0.79 0.78 0.62 0.62 0.75 0.51 0.76 0.67
```

- X15 and X17 deleted
- Overall MSA value falls in the acceptable range (above 0.50)

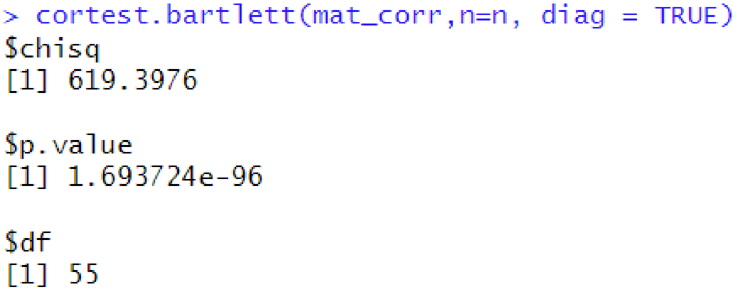
Correlation matrix after removing X15 and X17



	X6	*X7*	*X8*	*X9*	*X10*	*X11*	*X12*	*X13*	*X14*	*X16*	*X18*
X6		-0.137	0.096	0.106	-0.053	0.477***	-0.152	-0.401***	0.088	0.104	0.028
X7	-0.137		0.001	0.140	0.430***	-0.053	0.792***	0.229*	0.052	0.156	0.192
X8	0.096	0.001		0.097	-0.063	0.193	0.017	-0.271**	0.797***	0.080	0.025
X9	0.106	0.140	0.097		0.197*	0.561***	0.230*	-0.128	0.140	0.757***	0.865***
X10	-0.053	0.430***	-0.063	0.197*		-0.012	0.542***	0.134	0.011	0.184	0.276**
X11	0.477***	-0.053	0.193	0.561***	-0.012		-0.061	-0.495***	0.273**	0.424***	0.602***
X12	-0.152	0.792***	0.017	0.230*	0.542***	-0.061		0.265**	0.107	0.195	0.272**
X13	-0.401***	0.229*	-0.271**	-0.128	0.134	-0.495***	0.265**		-0.245*	-0.115	-0.073
X14	0.088	0.052	0.797***	0.140	0.011	0.273**	0.107	-0.245*		0.197*	0.109
X16	0.104	0.156	0.080	0.757***	0.184	0.424***	0.195	-0.115	0.197*		0.751***
X18	0.028	0.192	0.025	0.865***	0.276**	0.602***	0.272**	-0.073	0.109	0.751***	

Computed correlation used pearson-method with listwise-deletion.

Bartlett test after removing X15 and X17

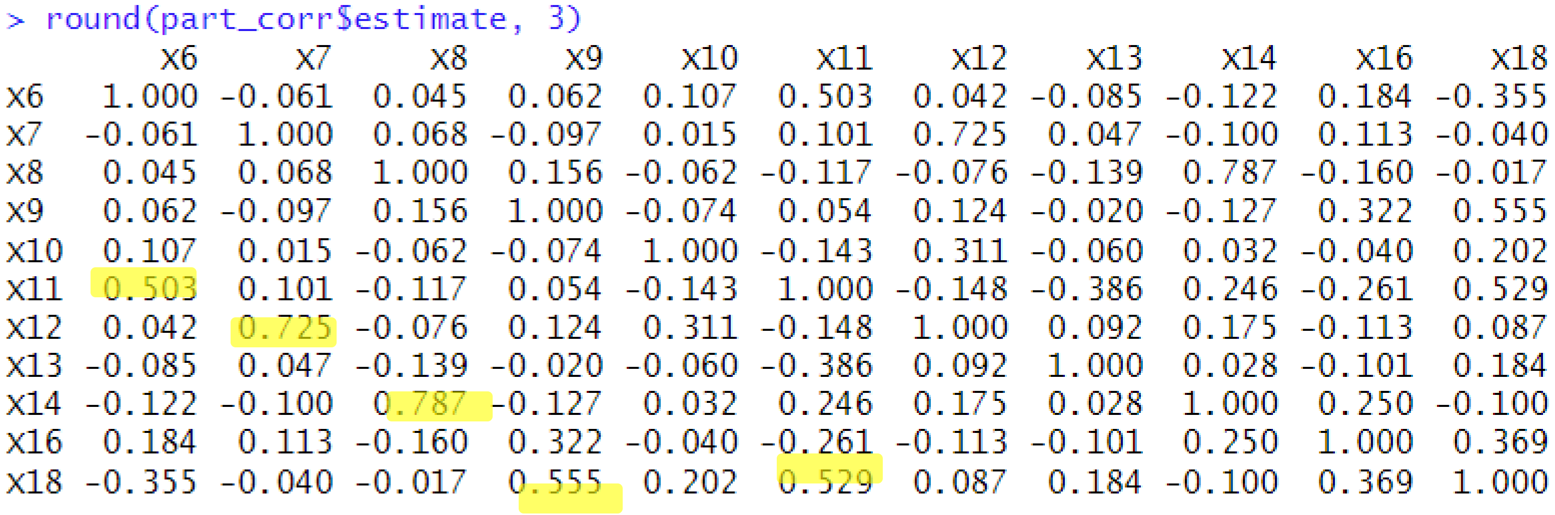


```
> cortest.bartlett(mat_corr,n=n, diag = TRUE)
$chisq
[1] 619.3976

$p.value
[1] 1.693724e-96

$df
[1] 55
```

- 20 of the 55 correlations are statistically significant.
- As with the full set of variables, the Bartlett test shows that non-zero correlations exist at the significance level of .0001.

Partial correlation matrix after removing variable X15 and X17

library(ppcor)
part_corr <- pcor(data_pca)round(part_corr$estimate, 3)

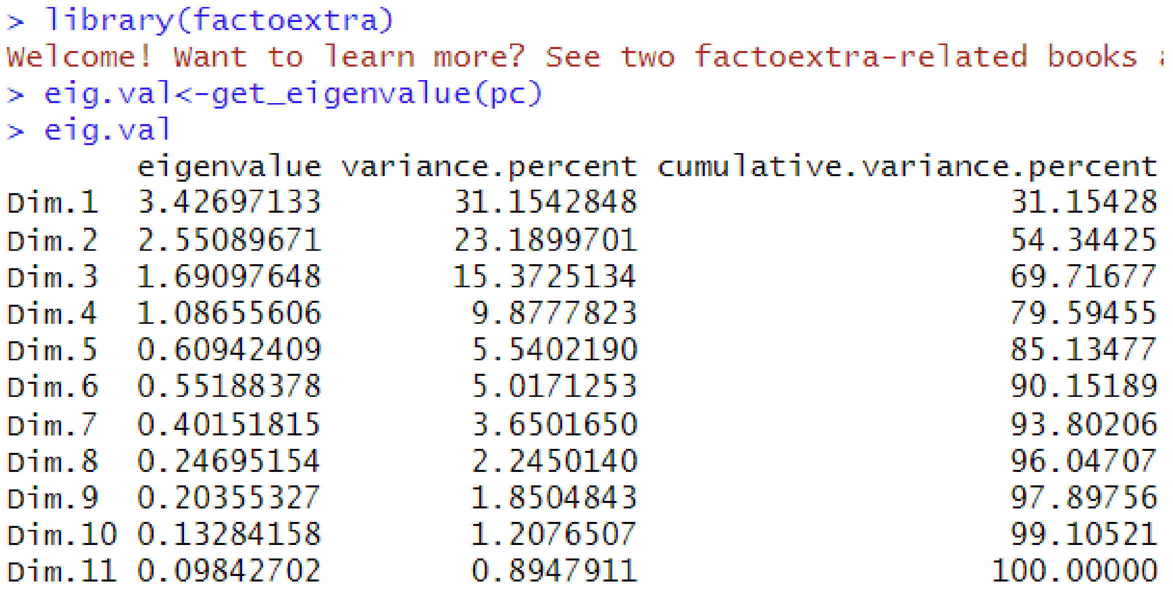


```
> round(part_corr$estimate, 3)
        x6     x7     x8     x9    x10    x11    x12    x13    x14    x16    x18
x6   1.000 -0.061  0.045  0.062  0.107  0.503  0.042 -0.085 -0.122  0.184 -0.355
x7  -0.061  1.000  0.068 -0.097  0.015  0.101  0.725  0.047 -0.100  0.113 -0.040
x8   0.045  0.068  1.000  0.156 -0.062 -0.117 -0.076 -0.139  0.787 -0.160 -0.017
x9   0.062 -0.097  0.156  1.000 -0.074  0.054  0.124 -0.020 -0.127  0.322  0.555
x10  0.107  0.015 -0.062 -0.074  1.000 -0.143  0.311 -0.060  0.032 -0.040  0.202
x11  0.503  0.101 -0.117  0.054 -0.143  1.000 -0.148 -0.386  0.246 -0.261  0.529
x12  0.042  0.725 -0.076  0.124  0.311 -0.148  1.000  0.092  0.175 -0.113  0.087
x13 -0.085  0.047 -0.139 -0.020 -0.060 -0.386  0.092  1.000  0.028 -0.101  0.184
x14 -0.122 -0.100  0.787 -0.127  0.032  0.246  0.175  0.028  1.000  0.250 -0.100
x16  0.184  0.113 -0.160  0.322 -0.040 -0.261 -0.113 -0.101  0.250  1.000  0.369
x18 -0.355 -0.040 -0.017  0.555  0.202  0.529  0.087  0.184 -0.100  0.369  1.000
```

- examining the partial correlations shows only five with values greater than 0.50 (X6–X11, X7–X12, X8–X14, X9–X18, and X11–X18), which is another indicator of the strength of the interrelationships among the variables in the reduced set.
- Both X11 and X18 are involved in two of the high partial correlations.
- Collectively, these measures all indicate that the reduced set of variables is appropriate for factor analysis, and can proceed to the next stages.

Stage 4: deriving factors and assessing overall fit

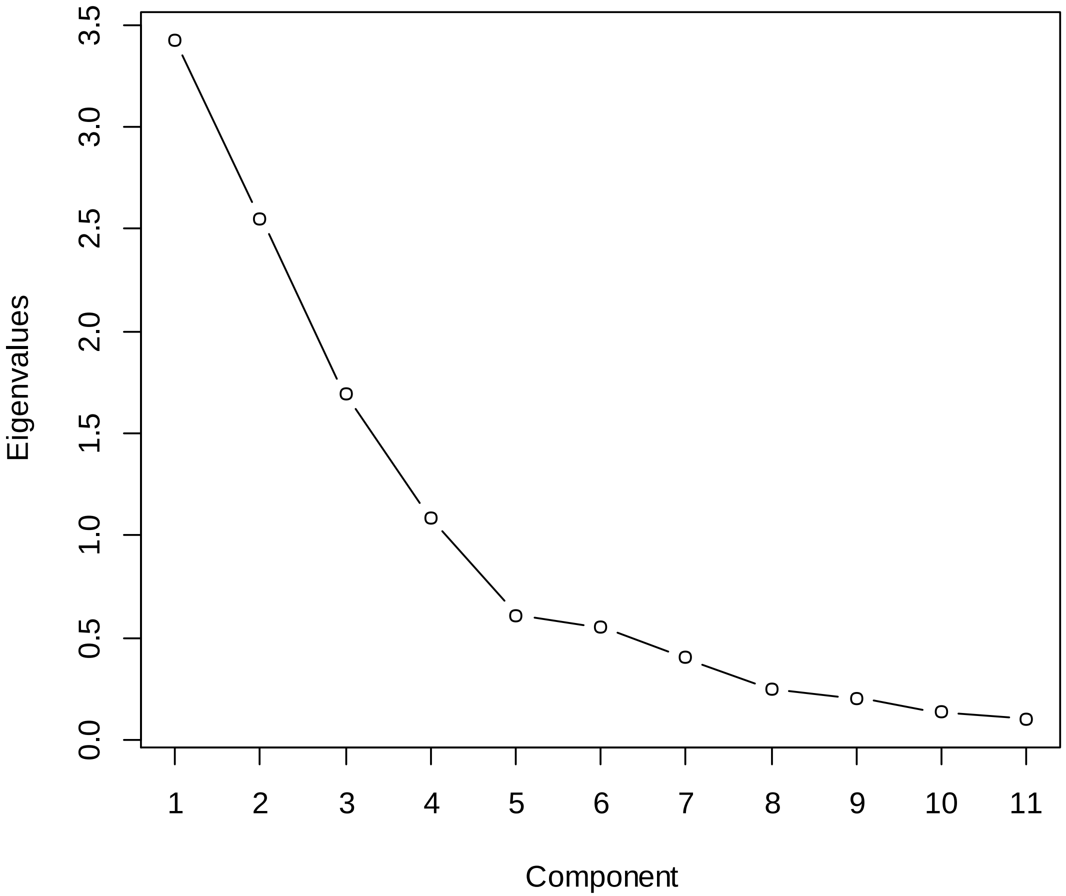
- **A priori criterion**. Practical reasons of desiring multiple measures per factor (at least 2 and preferably 3) dictate that between **three and five components** would be best given the 11 variables to be analyzed.



```
> library(factoextra)
Welcome! Want to learn more? See two factoextra-related books
> eig.val<-get_eigenvalue(pc)
> eig.val
       eigenvalue variance.percent cumulative.variance.percent
Dim.1  3.42697133       31.1542848                    31.15428
Dim.2  2.55089671       23.1899701                    54.34425
Dim.3  1.69097648       15.3725134                    69.71677
Dim.4  1.08655606        9.8777823                    79.59455
Dim.5  0.60942409        5.5402190                    85.13477
Dim.6  0.55188378        5.0171253                    90.15189
Dim.7  0.40151815        3.6501650                    93.80206
Dim.8  0.24695154        2.2450140                    96.04707
Dim.9  0.20355327        1.8504843                    97.89756
Dim.10 0.13284158        1.2076507                    99.10521
Dim.11 0.09842702        0.8947911                   100.00000
```

- **Latent root criterion**. If we retain factors with eigenvalues greater than 1.0, **four components** will be retained
- **Percentage of variance criterion**. The **four components** retained represent 79.6 % of the variance of the 11 variables, deemed sufficient in terms of total variance explained.

Stage 4: deriving factors and assessing overall fit

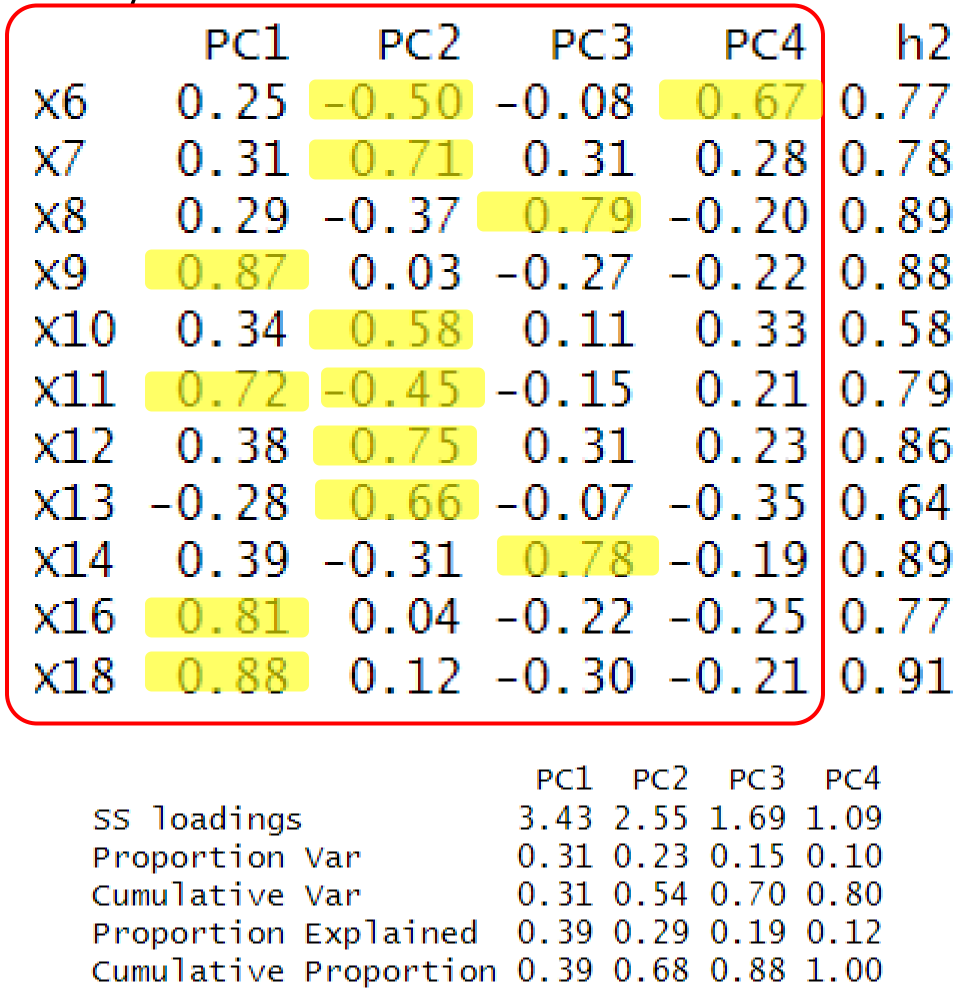


- The scree test indicates that **four or perhaps five components** may be appropriate when considering the changes in eigenvalues (i.e., identifying the "elbow" in the eigenvalues at the fifth factor).

- Combining all these criteria together is essential given that there is no single best method for determining the number of factors.
- **In this case it leads to the conclusion to retain four components for further analysis.**

Stage 5: interpreting the factors

Output below presents the unrotated principal component analysis factor matrix.

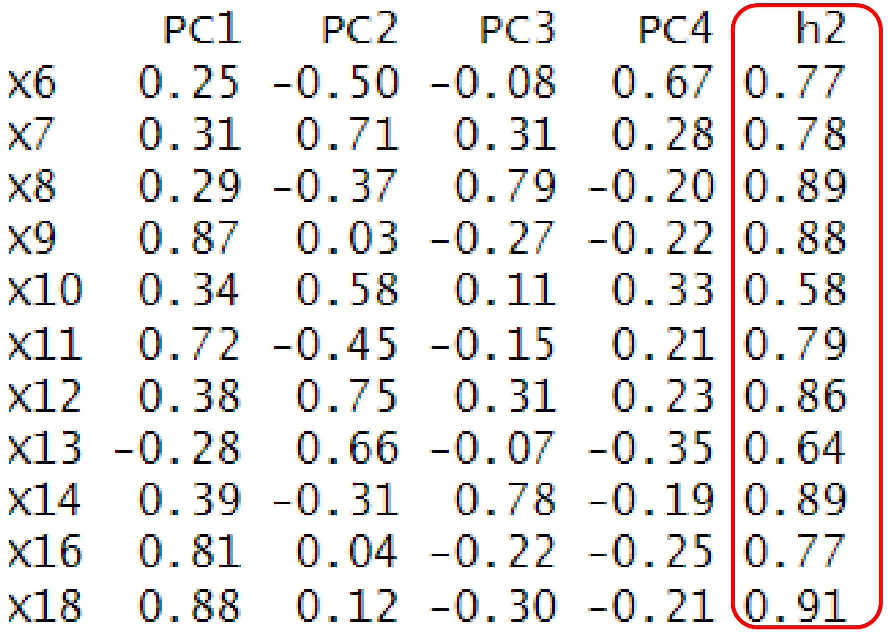


```
       PC1   PC2   PC3   PC4   h2
x6    0.25 -0.50 -0.08  0.67 0.77
x7    0.31  0.71  0.31  0.28 0.78
x8    0.29 -0.37  0.79 -0.20 0.89
x9    0.87  0.03 -0.27 -0.22 0.88
x10   0.34  0.58  0.11  0.33 0.58
x11   0.72 -0.45 -0.15  0.21 0.79
x12   0.38  0.75  0.31  0.23 0.86
x13  -0.28  0.66 -0.07 -0.35 0.64
x14   0.39 -0.31  0.78 -0.19 0.89
x16   0.81  0.04 -0.22 -0.25 0.77
x18   0.88  0.12 -0.30 -0.21 0.91

                       PC1  PC2  PC3  PC4
SS loadings           3.43 2.55 1.69 1.09
Proportion Var        0.31 0.23 0.15 0.10
Cumulative Var        0.31 0.54 0.70 0.80
Proportion Explained  0.39 0.29 0.19 0.12
Cumulative Proportion 0.39 0.68 0.88 1.00
```

- **The first factor** accounts for the largest amount of variance
- **The second factor** is somewhat of a general factor, with half of the variables having a high loading (high loading is defined as greater than 0.40).
- **The third factor** has two high loadings, whereas the fourth factor only has one high loading.
- Based on this factor-loading pattern with a relatively large number of high loadings on factor 2 and only one high loading on factor 4, interpretation would be difficult and theoretically less meaningful.
- Therefore, we should proceed to rotate the factor matrix to redistribute the variance from the earlier factors to the later factors.
- However, before proceeding with the rotation process, we must examine the communalities to see whether any variables have communalities so low that they should be eliminated.

Stage 5: interpreting the factors

Output below presents the unrotated principal component analysis factor matrix.

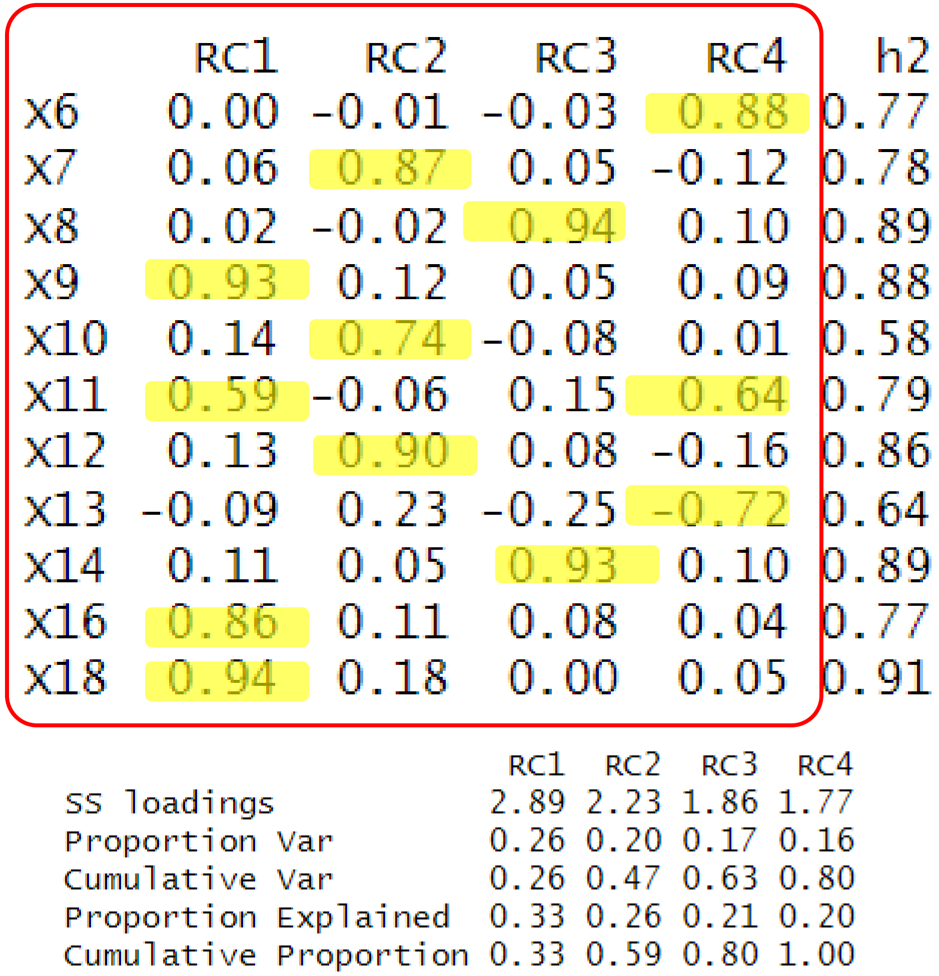


```
      PC1   PC2   PC3   PC4   h2
x6   0.25 -0.50 -0.08  0.67 0.77
x7   0.31  0.71  0.31  0.28 0.78
x8   0.29 -0.37  0.79 -0.20 0.89
x9   0.87  0.03 -0.27 -0.22 0.88
x10  0.34  0.58  0.11  0.33 0.58
x11  0.72 -0.45 -0.15  0.21 0.79
x12  0.38  0.75  0.31  0.23 0.86
x13 -0.28  0.66 -0.07 -0.35 0.64
x14  0.39 -0.31  0.78 -0.19 0.89
x16  0.81  0.04 -0.22 -0.25 0.77
x18  0.88  0.12 -0.30 -0.21 0.91
```

- **The row sum of squared factor loadings are referred to as communalities.**
- The communalities show **the amount of variance in a variable that is accounted for by all of the retained factors taken together**.
- Higher communality values indicate that a large amount of the variance in a variable has been extracted by the factor solution.
- Small communalities show that a substantial portion of the variable’s variance is not accounted for by the factors.
- Although no statistical guidelines indicate exactly what is “large” or “small,” practical considerations are consistent with a lower level of .50 for communalities in this analysis.
- For instance, the communality value **of 0.58 for variable X10 indicates that it has less in common** with the other variables included in the analysis than does variable X8, which has a communality of 0.89. Both variables, however, still share more than one-half of their variance with the four factors.
- All of the communalities are sufficiently high to proceed with the rotation of the factor matrix.

Stage 5: interpreting the factors

Output below presents the varimax rotated principal component analysis factor matrix.

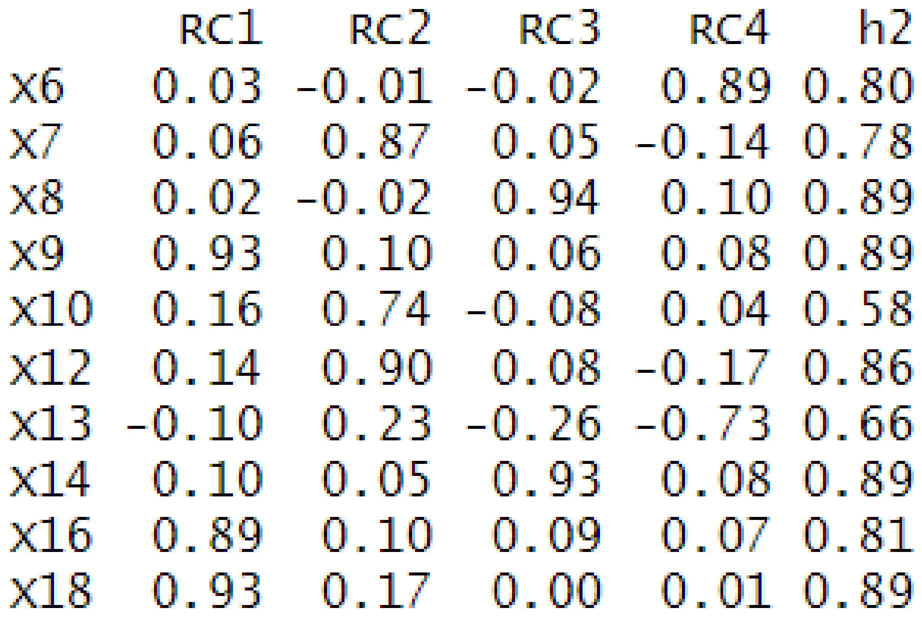


```
      RC1   RC2   RC3   RC4   h2
x6   0.00 -0.01 -0.03  0.88 0.77
x7   0.06  0.87  0.05 -0.12 0.78
x8   0.02 -0.02  0.94  0.10 0.89
x9   0.93  0.12  0.05  0.09 0.88
x10  0.14  0.74 -0.08  0.01 0.58
x11  0.59 -0.06  0.15  0.64 0.79
x12  0.13  0.90  0.08 -0.16 0.86
x13 -0.09  0.23 -0.25 -0.72 0.64
x14  0.11  0.05  0.93  0.10 0.89
x16  0.86  0.11  0.08  0.04 0.77
x18  0.94  0.18  0.00  0.05 0.91

                       RC1  RC2  RC3  RC4
SS loadings           2.89 2.23 1.86 1.77
Proportion Var        0.26 0.20 0.17 0.16
Cumulative Var        0.26 0.47 0.63 0.80
Proportion Explained  0.33 0.26 0.21 0.20
Cumulative Proportion 0.33 0.59 0.80 1.00
```

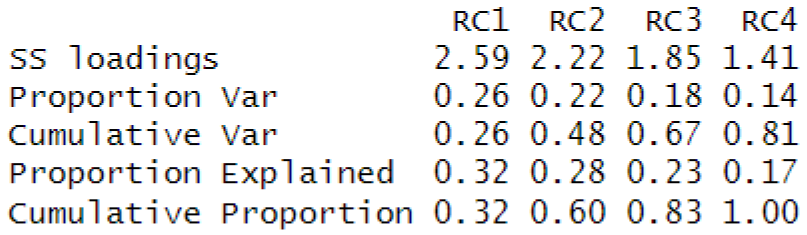
- total amount of variance extracted is the same in the rotated solution as it was in the unrotated solution, 80%.
- the communalities for each variable do not change when a rotation technique is applied.
- Specifically, in the VARIMAX-rotated factor the first factor accounts for 26.3 percent of the variance, compared to 31.2 percent in the unrotated solution.
- Likewise, the other factors also change, the largest change being the fourth factor, increasing from 9.9 percent in the unrotated solution to 16.1 percent in the rotated solution.
- the factor loadings for each variable are maximized for each variable on one factor.
- In the rotated factor solution, each of the variables has a significant loading (defined as a loading above .40) on only one factor, except for X11, which cross-loads on two factors (factors 1 and 4). Moreover, all the loadings are above 0.70, meaning that more than one-half of the variance is accounted for by the loading on a single factor.
- The only remaining decision is to determine the action to be taken for X11.

Stage 5: interpreting the factors

Output below presents the varimax rotated principal component analysis factor matrix.



```
      RC1   RC2   RC3   RC4   h2
X6   0.03 -0.01 -0.02  0.89 0.80
X7   0.06  0.87  0.05 -0.14 0.78
X8   0.02 -0.02  0.94  0.10 0.89
X9   0.93  0.10  0.06  0.08 0.89
X10  0.16  0.74 -0.08  0.04 0.58
X12  0.14  0.90  0.08 -0.17 0.86
X13 -0.10  0.23 -0.26 -0.73 0.66
X14  0.10  0.05  0.93  0.08 0.89
X16  0.89  0.10  0.09  0.07 0.81
X18  0.93  0.17  0.00  0.01 0.89
```



```
                       RC1  RC2  RC3  RC4
SS loadings           2.59 2.22 1.85 1.41
Proportion Var        0.26 0.22 0.18 0.14
Cumulative Var        0.26 0.48 0.67 0.81
Proportion Explained  0.32 0.28 0.23 0.17
Cumulative Proportion 0.32 0.60 0.83 1.00
```

- **Factor 1 Postsale Customer Service**: X9 complaint resolution; X18, delivery speed; and X16, order and billing;
- **Factor 2 Marketing**: X12 salesforce image; X7, e-commerce presence; and X10, advertising;
- **Factor 3 Technical Support**: X8 technical support; and X14, warranty and claims;
- **Factor 4 Product Value**: X6 product quality; and X13, competitive pricing.

Stage 6: validation of principal components analysis

- Optimally, we would always follow our use of exploratory factor analysis with some form of confirmatory factor analysis, such as structure equation modeling, but this type of follow-up is often not feasible.
- We must look to other means, such as split sample analysis or application to entirely new samples. In this example, we split the sample into two equal samples of 50 respondents and re-estimate the factor models to test for comparability.

Stage 7: additional uses of the exploratory factor analysis results

- Selecting surrogate variables for subsequent analysis
- Creating summated scales
- Use a factor scores

Thank you