Workshop Construction and Analysis of Tests and Questionnaires

Day 2: Validity (Factor analysis)

Simulated data for illustrative purposes

Generated/fictitious dataset with six items X1 upto and including X6 Depression-scale

N respondents = 300

Correlation matrix:

Correlation Matrix

		Anxious	Tense	Restless	Depressed	Useless	Unhappy
Correlation	Anxious	1,000	,449	,443	,296	,314	,326
	Tense	,449	1,000	,446	,312	,264	,250
	Restless	,443	,446	1,000	,279	,258	,282
	Depressed	,296	,312	,279	1,000	,467	,516
	Useless	,314	,264	,258	,467	1,000	,497
	Unhappy	,326	,250	,282	,516	,497	1,000

Correlation Matrix

		Anxious	Tense	Restless	Depressed	Useless	Unhappy
Correlation	Anxious	1,000	,449	,443	,296	,314	,326
	Tense	,449	1,000	,446	,312	,264	,250
	Restless	,443	,446	1,000	,279	,258	,282
	Depressed	,296	,312	,279	1,000	,467	,516
	Useless	,314	,264	,258	,467	1,000	,497
	Unhappy	,326	,250	,282	,516	,497	1,000

Can you see a pattern? Are there multiple dimensions?

Correlation Matrix

		X1: Anxious	X2: Tense	X3: Restless	X4: Depressed	X5: Useless	X6: Unhappy
Correlation	X1: Anxious	1,000	,449	,443	,296	,314	,326
	X2: Tense	,449	1,000	,446	,312	,264	,250
	X3: Restless	,443	,446	1,000	,279	,258	,282
	X4: Depressed	,296	,312	,279	1,000	,467	,516
	X5: Useless	,314	,264	,258	,467	1,000	,497
	X6: Unhappy	,326	,250	,282	,516	,497	1,000

- For now: we just assume that we have two factors.
  We will learn later how we can determine the number of factors

- Two factors -> 2 dimensions -> 2 scales
  - Factor 1: Accounts for most of the variance seen in the six items
  - Factor 2: **given the first factor and in addition to this factor** accounts for most of the *remaining* variance of the six items

Component Matrix [a]

	Component	
	1	2
X1: Anxious	,687	,375
X2: Tense	,654	,467
X3: Restless	,651	,464
X4: Depressed	,708	-,385
X5: Useless	,688	-,415
X6: Unhappy	,710	-,432

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

- Here, in this table we see the factorloadings for every item on factor 1 and factor 2 (denoted as components in the table)
- Note that the second factor accounts for unexplained var/correlation *on top* of the 1st component

- The loadings on the first and second factor are equal to the correlation of an item with these factors (dimensions):

- $a_{j1} = r_{X_j F_1}$
- $a_{j2} = r_{X_j F_2}$

- For example: the correlation of the item 'Anxious' with the second factor = 0,375:

- $r_{Anxious,F2} = a_{12} = 0{,}375$

Component Matrix [a]

	Component	
	1	2
X1: Anxious	,687	,375
X2: Tense	,654	,467
X3: Restless	,651	,464
X4: Depressed	,708	-,385
X5: Useless	,688	-,415
X6: Unhappy	,710	-,432

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

- The eigenvalue λ_2 is equal to the total amount of variance that is accounted for by one of the factors. For example, for the second factor:
- $\lambda_2 = a_{12}^2 + a_{22}^2 + a_{32}^2 + a_{42}^2 + a_{52}^2 + a_{62}^2$
- For example: the second component accounts for 1,081 variance in total
- $\lambda_2 = 1{,}081$
- So, 1,081 of the differences among people (e.g., how differently they filled in the questionnaire) can be accounted by this factor (this dimension).

Component Matrix [a]

	Component	
	1	2
X1: Anxious	,687	,375
X2: Tense	,654	,467
X3: Restless	,651	,464
X4: Depressed	,708	-,385
X5: Useless	,688	-,415
X6: Unhappy	,710	-,432

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

- The proportion of variance accounted for by factor 2 is equal to λ_2 divided by the total variance
- Prop.VAF = $\lambda_2/\mathrm{TotalVar} = \lambda_2/J$

  where J is total number of items
- For example: the 2nd factor accounts for 1,081/6 = 18,0% of the variance in total
- Total (proportion) of variance accounted for by both factors: $\lambda_1 + \lambda_2$ = 3,883 (64,7%)

In SPSS

Total proportion of variance accounted by the 2 factors

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,802	46,700	46,700	2,802	46,700	46,700
2	1,081	18,015	64,715	1,081	18,015	64,715
3	,578	9,639	74,354			
4	,560	9,335	83,690			
5	,522	8,693	92,382			
6	,457	7,618	100,000			

Extraction Method: Principal Component Analysis.

Eigenvalue of factor 2

Total variance accounted by factor 2

Proportion of variance accounted by factor 2

Reproduced correlations

- Also the correlation matrix can be (approximately) reproduced by the solution of a factor analysis. This is the matrix of reproduced correlations, R_{prod}
- The reproduced correlation between two variables is equal to the *sum – over components* – of the product of the variable loadings on the components
- For example:

$$\widehat{r_{12}} = (a_{11}*a_{21}) + (a_{12}*a_{22})$$

$$= (0{,}687*0{,}654) + (0{,}375*0{,}467) = 0{,}624$$

- The residual correlation is the difference between the observed and reproduced correlation. These residual correlations are contained in the matrix with *residual correlations*, R_{res}

- For example: $r_{12}^{res} = r_{12} - \widehat{r_{12}}$ = 0,449- 0,624 = -0,175

Reproduced Correlations

		X1: Anxious	X2: Tense	X3: Restless	X4: Depressed	X5: Useless	X6: Unhappy
Reproduced Correlation	X1: Anxious	,613 [a]	,625	,621	,343	,317	,326
	X2: Tense	,625	,646 [a]	,642	,284	,256	,262
	X3: Restless	,621	,642	,639 [a]	,282	,255	,261
	X4: Depressed	,343	,284	,282	,649 [a]	,647	,669
	X5: Useless	,317	,256	,255	,647	,646 [a]	,668
	X6: Unhappy	,326	,262	,261	,669	,668	,691 [a]
Residual [b]	X1: Anxious		-,175	-,178	-,047	-,004	,000
	X2: Tense	-,175		-,196	,028	,008	-,012
	X3: Restless	-,178	-,196		-,003	,003	,021
	X4: Depressed	-,047	,028	-,003		-,180	-,153
	X5: Useless	-,004	,008	,003	-,180		-,171
	X6: Unhappy	,000	-,012	,021	-,153	-,171	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 6 (40,0%) nonredundant residuals with absolute values greater than 0.05.

- So, how do we know how many factors we need?!

- Are there any rules to decide on the number of factors?

How do we decide how many factors we have?

- How can you determine the number of factors underlying the data?
    - Option 1: Common sense
    - Option 2: Rules based on eigenvalues: Kaiser’s criterion and scree test
    - Option 3: Theory

(1) Common sense

- $K \leq J/3$

- Intuition: a factor with less than three items cannot be reliable

- For the example: $K \leq J/3 = 6/3 = 2$
  (K = total number of factors, J = total number of items)

(2) Kaiser's criterion

- "Greater-than-one" rule
- Number of factors = number of factors having an eigenvalue larger than one
- How many factors would we choose for our example, based on this rule?

In SPSS

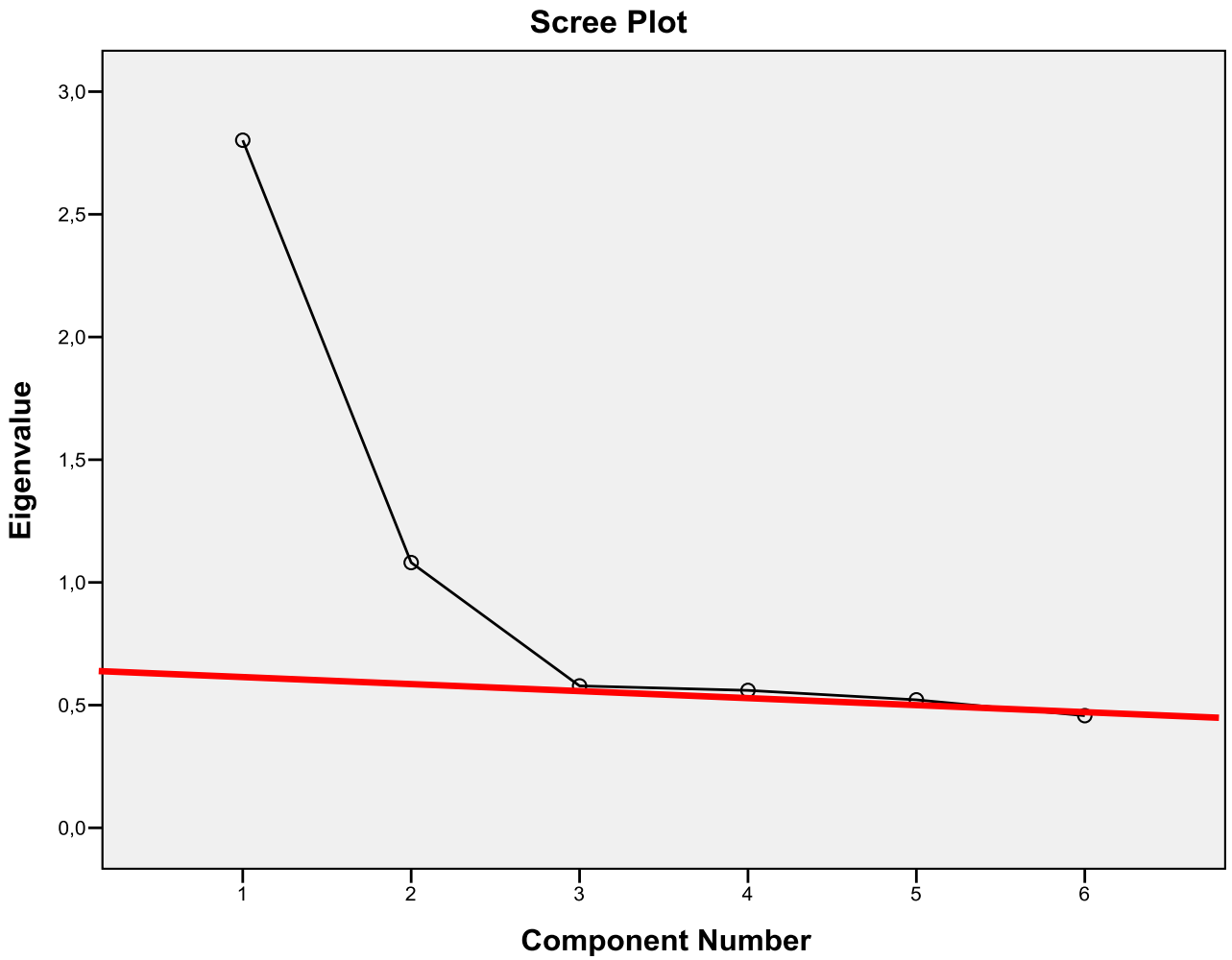
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,802	46,700	46,700	2,802	46,700	46,700
2	1,081	18,015	64,715	1,081	18,015	64,715
3	,578	9,639	74,354			
4	,560	9,335	83,690			
5	,522	8,693	92,382			
6	,457	7,618	100,000			

Extraction Method: Principal Component Analysis.

Eigenvalues:
2 are > 1 => 2 factors

(3) Scree test (Catell)



- Choose the number of factors for which the eigenvalue is above the line that goes true the “scree”
- Often very subjective
- Here: indicates a solution with 2 factors

(4) Theory

- Often there is some theoretical knowledge about the number of constructs/dimensions underlying the data

- Eg, NEO big five

- Use this knowledge to further support other indications of that number of factors or to choose between several options (in case one rule or different rules do not result in a unique number)
    - eg, scree test: often ‘different levels of scree’, or different number of factors based on Kaiser’s criterion and the scree plot

Methods to find easier solutions for a factor analysis

SPSS output in the perfect world

- Ideal would be that we don't have any doubts to which dimension an item belongs to
- E.g., perfect correlation/loading 1 on only ONE dimension + rest of the loadings = 0 (see table on the right)
- *-> easy interpretation!*
- *But in practice, we don't get that (what you see here to the right = hypothetical example)*

Component Matrix[a]

	Component		
	1	**2**	**3**
Item 1	1		
Item 2	1		
Item 3	1		
Item 4		1	
Item 5		1	
Item 6		1	
Item 7			1
Item 8			1
Item 9			1

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

… however, in real life: Interpretation of the output can be difficult

Component Matrix[a]			
	Component		
	1	2	3
+ situatie denk	.837	-.258	
- situatie denk	.819	-.245	
- denk	.762	-.230	
+ denk	.744	-.260	-.106
verkeer	.687		
controleer	.370	.761	
voor mezelf	.265	.745	.133
- uitdrukking	.279	.658	.324
+ uitdrukking	.253	.413	-.721
stressvol	.373		.580
Extraction Method: Principal Component Analysis.			
a. 3 components extracted.			

Component Matrix [a]

	Component	
	1	2
X1: Anxious	,687	,375
X2: Tense	,654	,467
X3: Restless	,651	,464
X4: Depressed	,708	-,385
X5: Useless	,688	-,415
X6: Unhappy	,710	-,432

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

- **Realistic aim: as simple as possible (-> easier interpretation!)**

- **Rules of thumb for what we refer to as “simple structure”:**

  - Every item (variable) measures only ONE construct = ever item correlates >|,30| with only one factor

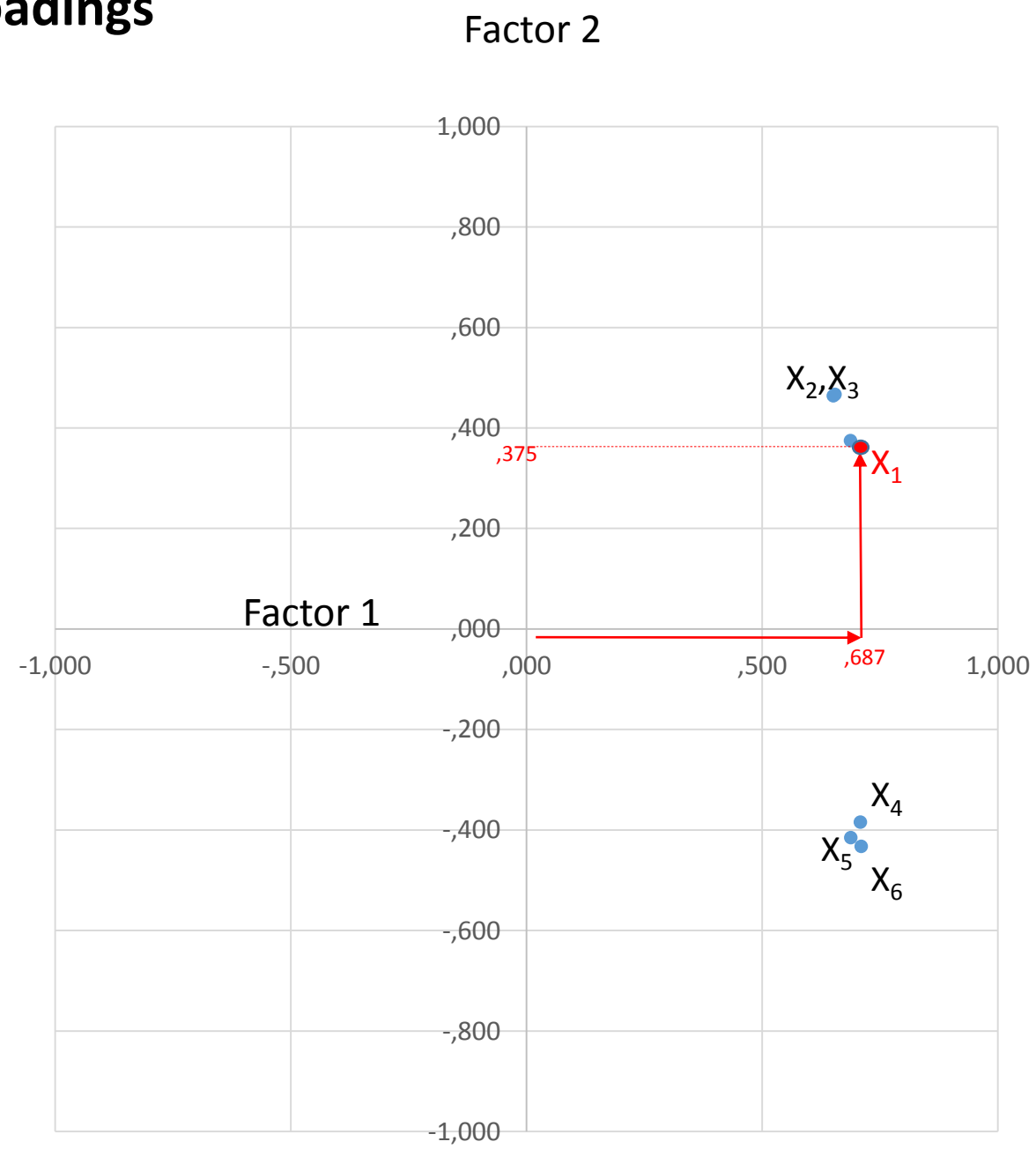
  - Every construct has its own group of items

Get closer to a simple structure: ROTATION

- The solution we get from a factor analysis is an estimation and therefore not unique:
  - There are endless possible combinations of factorloadings that are all equally good : - they explain the same amount of variance. Compare: 5 + 5 = 10, but 4 + 6 is also 10
  - -> *ROTATION* of the factors.
  - -> Statistical method that makes our lives easier
  - *Very complex mathematically, so you don't need to know the technical details but just remember that rotation is a technical method that eases interpretation of the results of a factor analysis (FA)*

Figure of the loadings

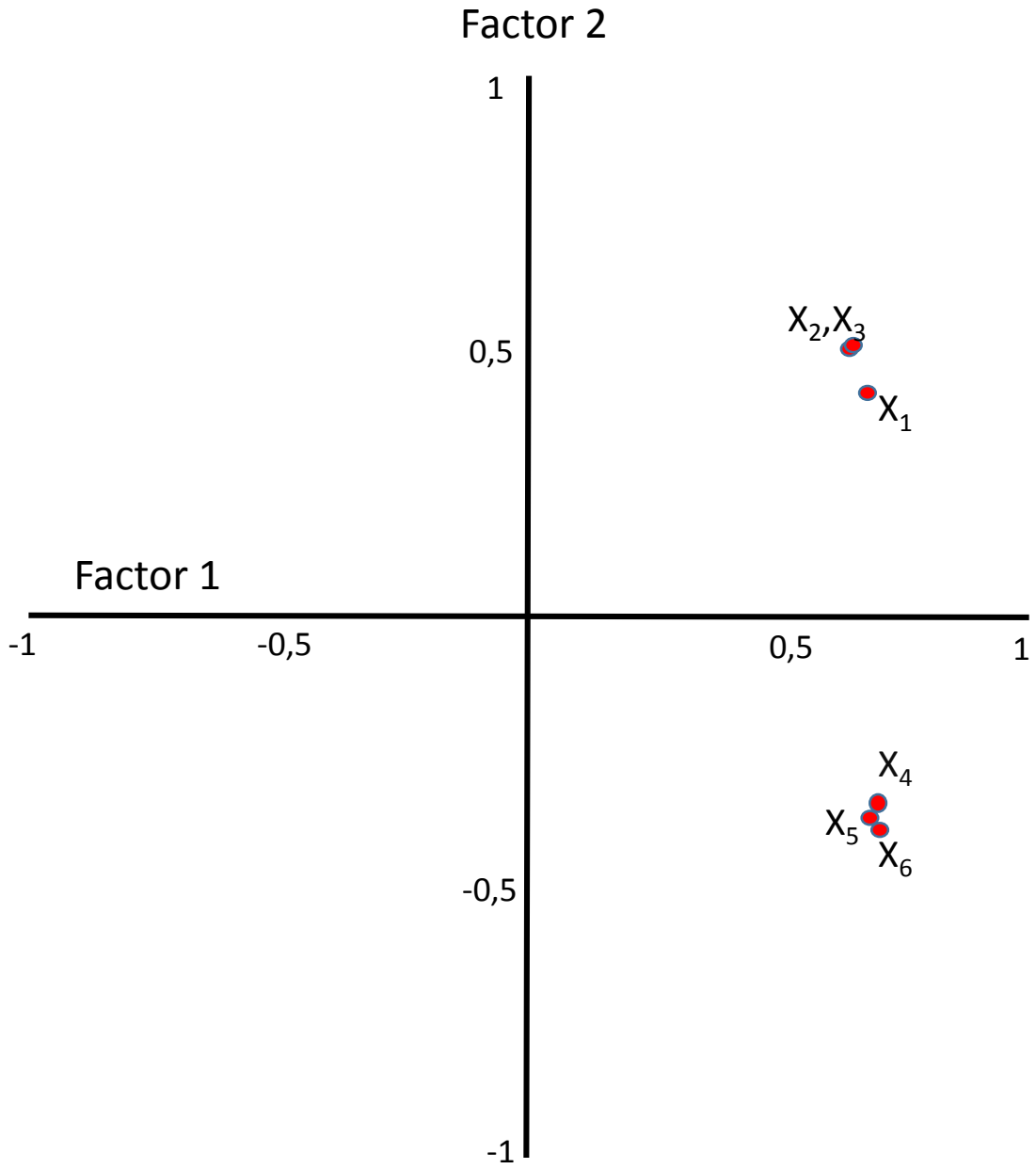
	Factor 1	Factor 2
X_1	0.687	0.375
X_2	0.654	0.467
X_3	0.651	0.464
X_4	0.708	-0.385
X_5	0.688	-0.415
X_6	0.710	-0.432



- Factor analysis with 2 factors, based on a dataset with 6 variables (X_1, X_2, X_3, X_4, X_5 & X_6)

- X axis = loading of every variable on the first factor (factor 1)

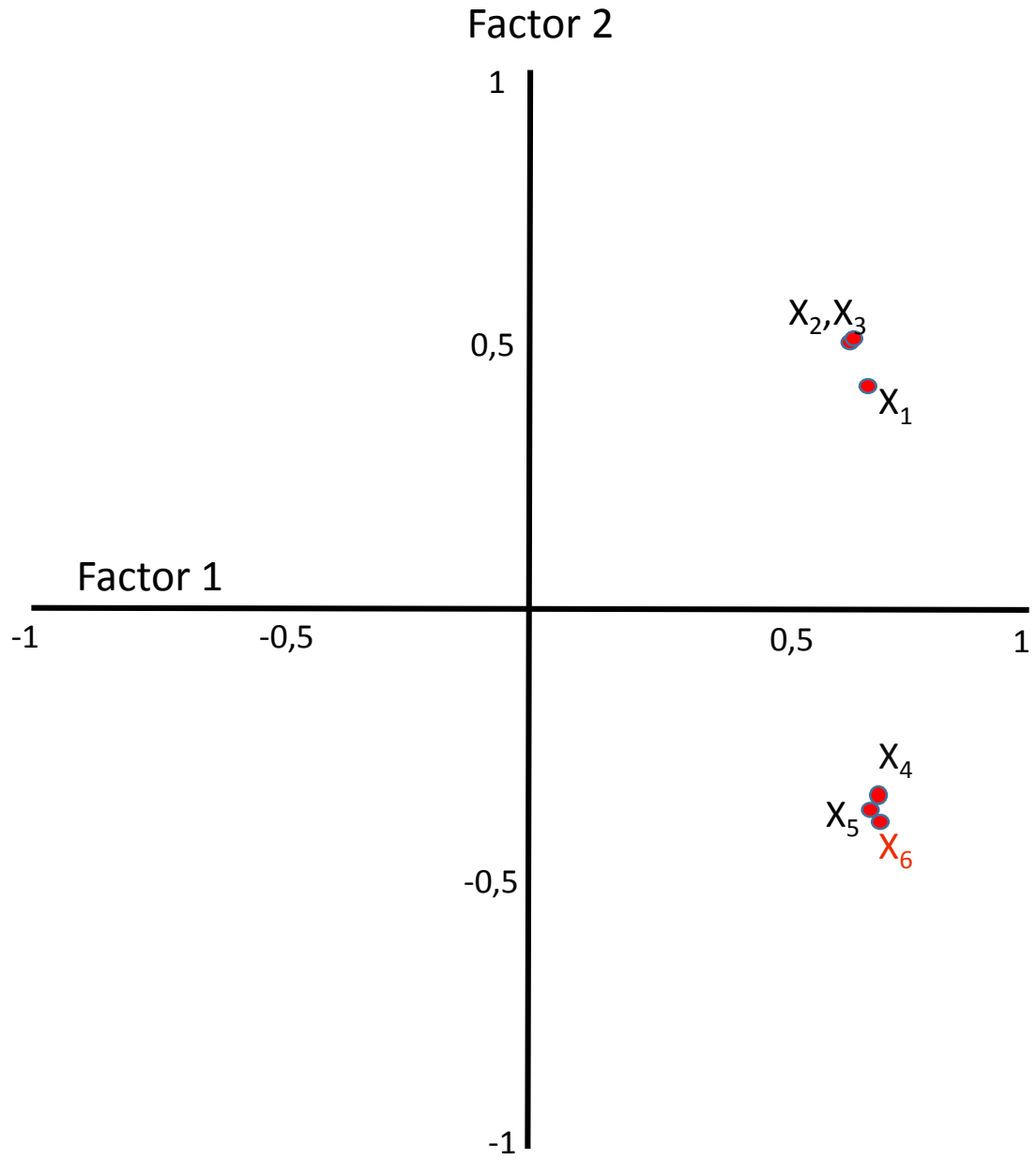
- Y axis = loading of every variable on the second factor (factor 2)



- Now: high loadings (in the absolute sense) on both factors

- For example:

- Variable $X_{6:}$ loading of ~ -.432 on component 2 and lading of .710 factor 1



-> difficult interpretation

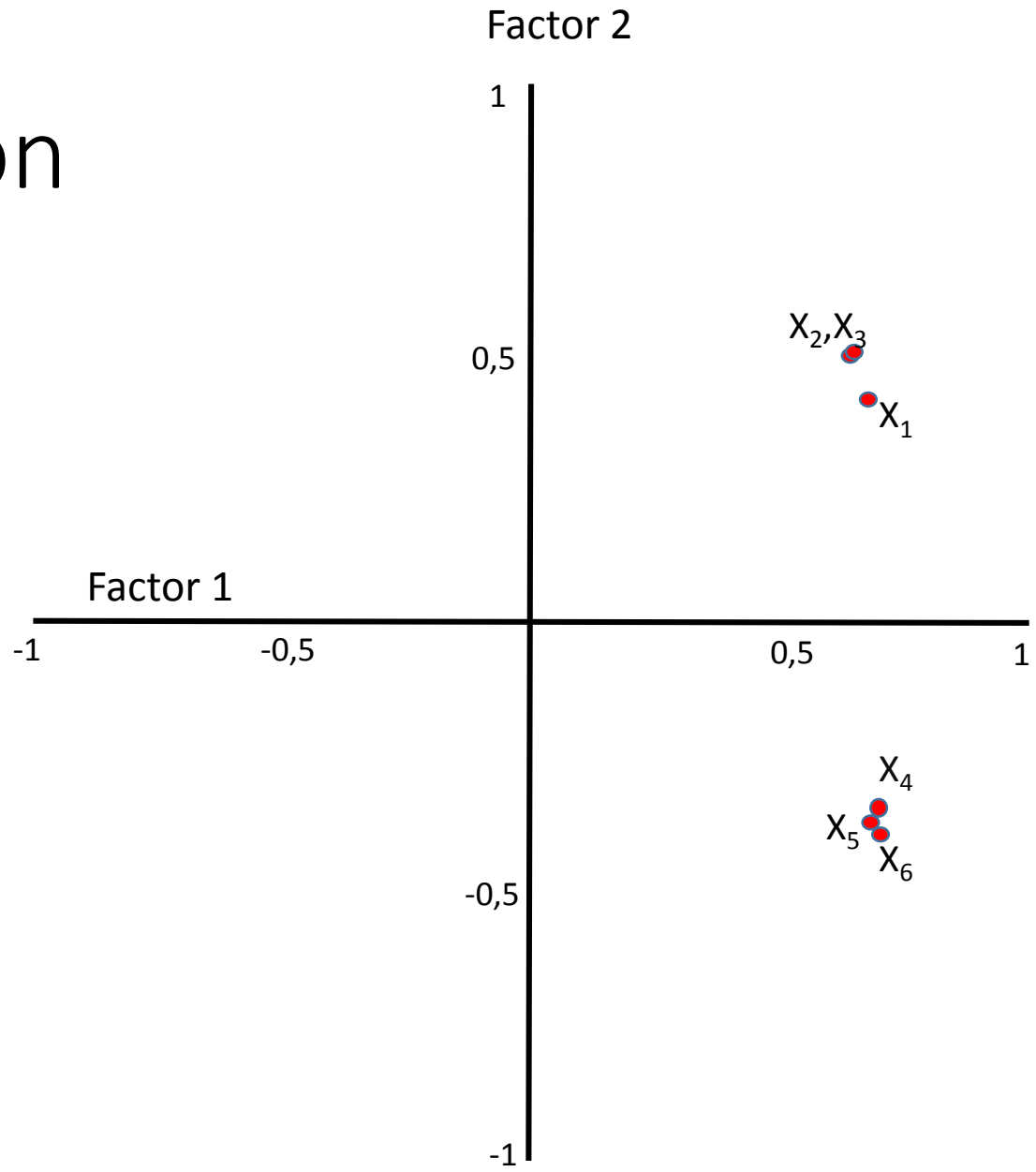
Rotation: demonstration

Rotation

= *changing the orientation of the x axis and y axis*

This changes the loadings

The quality of the solution remains the same, but the interpretation gets easier.

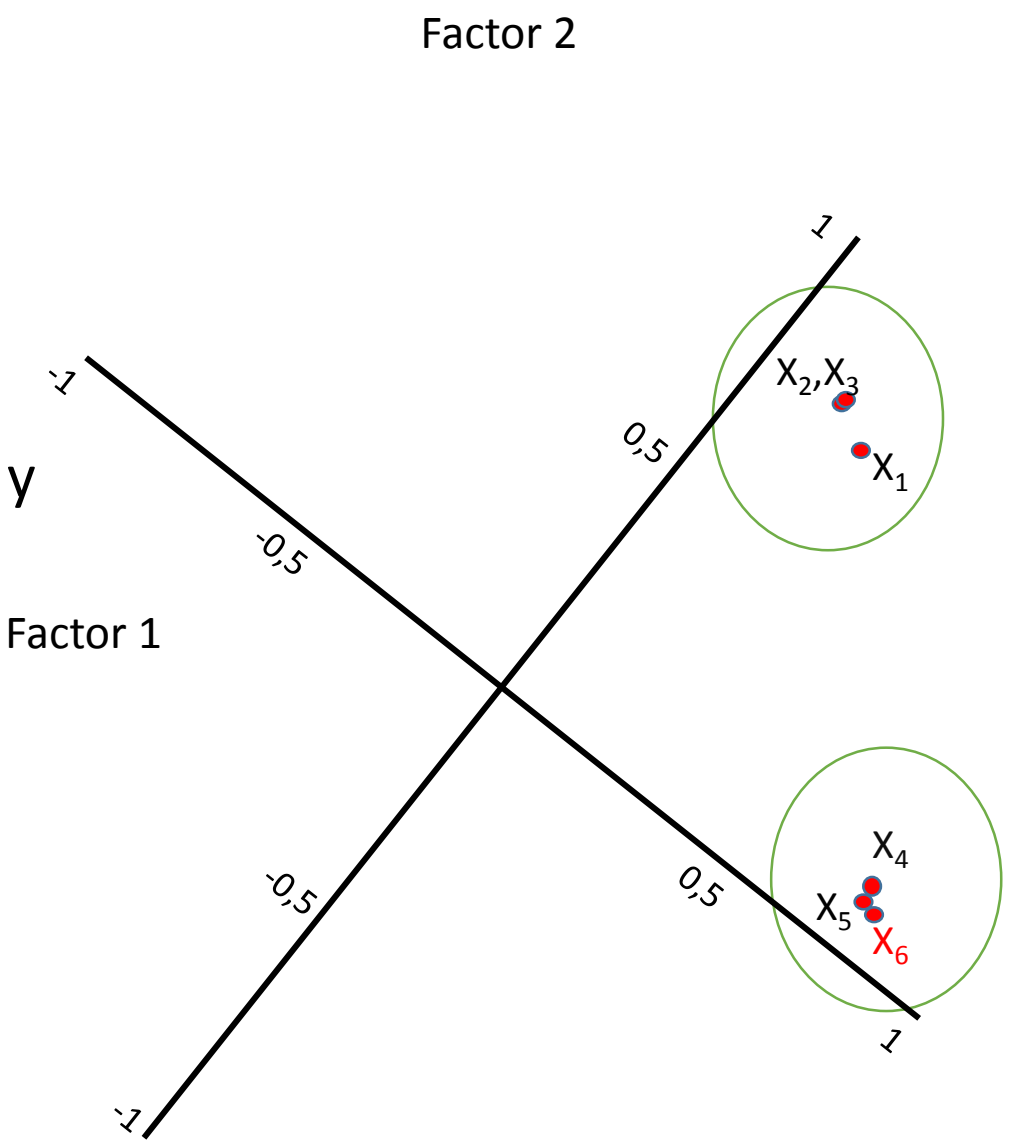


Rotation:demonstration

Rotation

= changing the orientation of the x axis and y axis

-> nou we can clearly see **2 groups of items** e.g., loadings of ~ .8 on factor 1 and .1 on factor 2 for item 6



Rotation: Summary of what happens

Question: I am confused, what is this rotation about?

Answer: We are interested in another/alternative (easier to interpret) solution.

We want items that follow a **simple structure,** meaning that an item loads strongly only on **one** factor (rule of thumb: loading of >|,30| with only *one* factor)

Through rotation, the loadings change such that we get closer to a simple structure, but the quality of the solution (the results of the analysis) remain the same (we still explain as much of the variance as we did before)

Pragmatically said:
Rotation = "***Changing" the loadings such that we can interpret the results easier***

Different sorts of rotations:

Orthogonal rotation (“VARIMAX” rotation)
Oblique rotation (“OBLIMIN” rotation)

Two types of rotation:

- **Orthogonal**:
  Factors *cannot* correlate with each other: $r_{F_m F_n} = 0$
  Many different techniques, but most commonly used: **VARIMAX** rotation

Two types of rotation:

- **Orthogonal**:
  Factors *cannot* correlate with each other: $r_{F_m F_n} = 0$
  Many different techniques, but most commonly used: **VARIMAX** rotation

- **Oblique**:
  Factors *can* correlate with each other: $r_{F_m F_n} \neq 0$
  Many different techniques, but most commonly used: **OBLIMIN** rotation

1. VARIMAX

In SPSS: VARIMAX

Component Matrix [a]

	Component	
	1	2
X1: Anxious	,687	,375
X2: Tense	,654	,467
X3: Restless	,651	,464
X4: Depressed	,708	-,385
X5: Useless	,688	-,415
X6: Unhappy	,710	-,432

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix [a]

	Component	
	1	2
X1: Anxious		,745
X2: Tense		,789
X3: Restless		,784
X4: Depressed	,779	
X5: Useless	,785	
X6: Unhappy	,813	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

- Simple structure?
- Easier interpretation than before rotation
- Reflects also what we expected from the correlation matrix (2 groups of items that have high inter-correlations and not very correlated with items of other group)

Rotated Component Matrix [a]

	Component	
	1	2
X1: Anxious		,745
X2: Tense		,789
X3: Restless		,784
X4: Depressed	,779	
X5: Useless	,785	
X6: Unhappy	,813	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

2. OBLIMIN

In SPSS: OBLIMIN

Pattern Matrix [a]

	Component	
	1	2
X1: Anxious	,076	,747
X2: Tense	-,030	,816
X3: Restless	-,030	,812
X4: Depressed	,792	,030
X5: Useless	,808	-,010
X6: Unhappy	,837	-,014

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 5 iterations.

- Loadings are now in *pattern* matrix instead of (rotated) component matrix
- Stronger *simple structure* than with VARIMAX (loadings are more extreme / closer to one and zero)

In SPSS: OBLIMIN

Component Correlation Matrix

Component	1	2
1	1,000	,438
2	,438	1,000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

- An assumption of the Oblimin rotation is that there is also a correlation between factors.

  So this correlation will be estimated (see table on the left side):

$r_{F1F2} = 0{,}438$

- Which solution to choose?
  - Simple structure => *always rotate*
  - VARIMAX or OBLIMIN? Choose <u>OBLIMIN if</u>
    - Simple structure better attained by OBLIMIN
    - OR, if the correlation between at least one pair of factors is ≥ |0,30|
    - Else VARIMAX (easier to work with uncorrelated/unrelated dimensions)
  - What kind of rotation would we choose for our example data?

- Ok, so now we know how to interpret the output of a factor analysis in SPSS.

- But, since we are performing a statistical techniques, aren’t there assumptions we have to check first?!

- -> Yes!

When is it allowed to perform FA?

- If
  - 1) the assumptions are met
  - 2) *N* and *J* satisfy some rules (*N* = total participants, *J* = total items)
  - 3) *R* satisfies some rules (*R* = correlation matrix of items)

- 1) **Assumptions FA**
  - Linear relation between the item pairs
  - Each item is of an ordinal measurement level with at least 5 answer categories

- 2) ***N* and *J* (*N* = total participants, *J* = total items)**
  - $N \geq 100$
  - $N \geq 5J$

- 3) **Correlation matrix *R***

- Items have to correlate sufficiently

  - (a) Bartlett's sphericity test must be signifcant ($p < 0,05$)

  - (b) KMO index must be larger than 0,6 ***This is the most important rule!***

Recap of our example data:

- Measuring depression
- 6 items: *Anxious*, *tense*, *restless*, *depressed*, *useless* & *unhappy*
- 300 respondents
- Likert items with 7 answer categories

Applied to our example data

- 1) Assumptions FA
  - Items have scores ranging from 1 upto and including 7
  - Relations are linear (or, they are not of another form): Check scatter plots!!

  $\Rightarrow$Correlation is a good measure of association

- 2) Assumptions *N* and *J*
  - N= 300 ≥ 100
  - N= 300 ≥ 5*J = 30

- 3) R
  - p < 0,05
  - KMO = 0,794

Correlation Matrix

		X1: Anxious	X2: Tense	X3: Restless	X4: Depressed	X5: Useless	X6: Unhappy
Correlation	X1: Anxious	1,000	,449	,443	,296	,314	,326
	X2: Tense	,449	1,000	,446	,312	,264	,250
	X3: Restless	,443	,446	1,000	,279	,258	,282
	X4: Depressed	,296	,312	,279	1,000	,467	,516
	X5: Useless	,314	,264	,258	,467	1,000	,497
	X6: Unhappy	,326	,250	,282	,516	,497	1,000

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,794
Bartlett's Test of Sphericity	Approx. Chi-Square	430,306
	df	15
	Sig.	,000

Very important!

- Factor analysis only works well if all assumptions are met & your sample size is big enough. Otherwise, you will get results from SPSS, but they can be biased.

- Hence, check your data for the assumptions

- If the assumptions are not met => do not apply FA

- We’ve heard so many things about reliability and validity…

How, and in what order, are all these steps applied in research practice?

- -> Step by step guide for the complete procedure of exploratory contruction of sclaes

Exploratory construction of scales: Complete procedure

1) Are the assumptions met?

2) Rotation: VARIMAX or OBLIMIN?

3) Interpretation of the factors:
   - E.g., factor 1: Which dimension is that? -> Look at the content of the items

Exploratory construction of scales: Complete procedure

- Five steps / analyses:
  - (i) **Check if the association between items is linear**

Exploratory construction of scales: Complete procedure

- (i) **Check if the association between items is linear**
- (ii) **First FA: does it make sense to perform FA + how many factors/components?**
  - KMO & Bartlett's sphericity test – assupmtions met?
  - How many components/factors do we choose? (How many dimensions?)

Exploratory construction of scales: Complete procedure

- (i) **Check if the association between items is linear**
- (ii) **First FA: does it make sense to perform FA + how many factors do we have?**
  - KMO & Bartlett's sphericity test – assupmtions met?
  - How many factors do we choose? (How many dimensions?)
- (iii) **Second FA: VARIMAX**
- (iv) **Third FA: OBLIMIN**

Exploratory construction of scales: Complete procedure

- (i) **Check if the association between items is linear**
- (ii) F**irst FA: does it make sense to perform FA + how many factors/components?**
  - KMO & Bartlett's sphericity test – assupmtions met?
  - How many components/factors do we choose? (How many dimensions?)
- (iii) **Second FA: VARIMAX**
- (iv) **Third FA: OBLIMIN**
- (v) **Choose a solution between (iii) & (iv) and interpret it**
  - Which solution is easier to interpret?
  - After choosing: Interpretation of the dimensions -> look at content of the items

Exploratory construction of scales: Complete procedure

- Seven steps / analyses:
  - (i) **Check if the association between items is linear**
  - (ii) F**irst FA: does it make sense to perform FA + how many factors/components?**
    - KMO & Bartlett's sphericity test – assupmtions met?
    - How many components/factors do we choose? (How many dimensions?)
  - (iii) **Second FA: VARIMAX**
  - (iv) **Third FA: OBLIMIN**
  - (v) **Choose a solution between (iii) & (iv) and interpret it**
    - Which solution is easier to interpret?
    - After choosing: Interpretation of the dimensions -> look at content of the items
  - (vi) Calculate sum scores for each scale / (sub)construct (each dimension)
  - (vii) Evaluate **reliability** of the scale, the item contributions to the reliability, and report Cronbach's alfa

We just have performed the first steps for the illustrative data (everything concerning factor analysis (FA)).

Let’s have a look at the last steps, for this particular dataset.

Exploratory construction of scales: Complete procedure

- Seven steps / analyses:
  - (i) **Check if the association between items is linear**
  - (ii) F**irst FA: does it make sense to perform FA + how many factors/components?**
    - KMO & Bartlett's sphericity test – assupmtions met?
    - How many components/factors do we choose? (How many dimensions?)
  - (iii) **Second FA: VARIMAX**
  - (iv) **Third FA: OBLIMIN**
  - (v) **Choose a solution between (iii) & (iv) and interpret it**
    - Which solution is easier to interpret?
    - After choosing: Interpretation of the dimensions -> look at content of the items
  - (vi) Calculate sum scores for each scale / (sub)construct (each dimension)
  - (vii) Evaluate **reliability** of the scale, the item contributions to the reliability, and report Cronbach's alfa

- Sum score => sum up only those items that form one scale.
- For example thus :

  Dimension 1 -> Item 4, Item 5 & Item 6 = SCALE 1 -> sumscore 1
  Dimension 2 -> Item 1, Item 2 & Item 3 = SCALE 2 -> sumscore 2

From where can we get this information (grouping of the items in scales)?

-> the output of the factor analysis (FA) gives us information on which items form together one scale.

- We have applied all steps of a factor analysis (FA) and we have chosen a VARIMAX rotation.
- To the right: the factorloadings of this VARIMAX rotation.

-> Dimension 1: = item 4, 5 & 6

-> Dimension 2: = item 1, 2 & 3

⇒First scale is sum score of X4, X5, and X6

⇒Second scale is sum score of X1, X2, and X3

Rotated Component Matrix [a]

	Component	
	1	2
X1: Anxious		,745
X2: Tense		,789
X3: Restless		,784
X4: Depressed	,779	
X5: Useless	,785	
X6: Unhappy	,813	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Advanced topics

- Some more advanced analyses are not possible in SPSS
  - **Test internal structure that is known beforehand**
    (*Confirmatory factor analysis*) -> see example script on github
  - **Test for measurement invariance**
    Compare multiple groups (e.g., companies/countries)
    (*Multiple group analysis*)
    Not possible by conducting separate factor analyses!!!