Statistik für Digital Humanities

Explorative Faktoranalyse

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

05. Juli 2021

[Letzte Aktualisierung: 04/07/2021, 18:49]

"It's a good job I'll never have to do that again" *AndyField*

Überblick

1. Latente Variablen / Faktoren
   - Was?
   - Berechnung
   - Scoring
2. Faktorenidentifizierung
   - Kommunalität
   - Principal Component Analysis vs. Faktoranalyse
   - Faktoranalyse
   - Verlässlichkeitsanalyse
3. Anwendungsbeispiel
   - Fragebögen
   - Faktoranalyse in *R*

Latente Variablen / Faktoren

- Latent : Nicht unmittelbar sichtbar
- Nicht beobachtbare Clustervariablen, die sich aus beobachtbaren Variablen zusammensetzen
- Identifizierbar mit Faktoranalyse und Principal Component Analyse (PCA)
- Cluster hoch korrelierender Variablen

Latente Variablen / Faktoren

- Latent : Nicht unmittelbar sichtbar
- Nicht beobachtbare Clustervariablen, die sich aus beobachtbaren Variablen zusammensetzen
- Identifizierbar mit Faktoranalyse und Principal Component Analyse (PCA)
- Cluster hoch korrelierender Variablen

Beispiel Burnout (Selbst nicht messbar)

- Stresslevel (messbar)
- Motivationsbereitschaft (messbar)
- Kreativität (messbar)

Latente Variablen / Faktoren

Korrelationsmatrix:

	ExoTalk	Sozialskill	Interesse	EgoTalk	Selbstsüchtig	LügnerIn
ExoTalk	1					
Sozialskill	.772	1				
Interesse	.646	.871	1			
EgoTalk	.072	-1.20	.054	1		
Selbstsüchtig	-.131	.031	-.101	.441	1	
LügnerIn	.068	0.12	.110	.361	.277	1

Exo-Talk $\rightarrow$ Sprechen über gegenüber

Ego-Talk $\rightarrow$ Sprechen über sich selbst

Latente Variablen / Faktoren

Korrelationsmatrix:

	ExoTalk	Sozialskill	Interesse	EgoTalk	Selbstsüchtig	LügnerIn
ExoTalk	1					
Sozialskill	**.772***	1				
Interesse	**.646***	**.871***	1			
EgoTalk	.072	-1.20	.054	1		
Selbstsüchtig	-.131	.031	-.101	**.441***	1	
LügnerIn	.068	0.12	.110	**.361***	**.277***	1

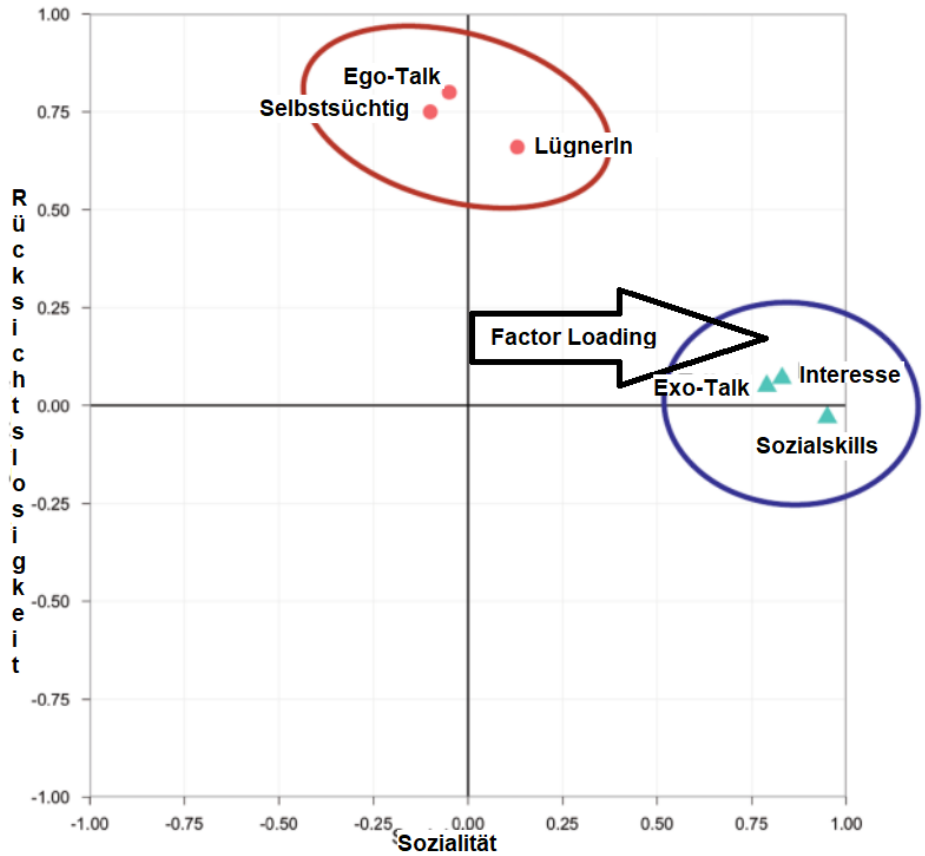
Exo-Talk $\rightarrow$ Sprechen über gegenüber

Ego-Talk $\rightarrow$ Sprechen über sich selbst

Identifizierbare Faktoren: 2 Cluster, die hoch intra- aber kaum interkorrelieren

- Sozialität:{Exo-Talk, Sozialskill, Interesse}
- Rücksichtslosigkeit:{Ego-Talk, Selbstsüchtig, LügnerIn} Lehrbuch: Consideration

Latente Variablen / Faktoren



- Faktor Loading = (Pearson)Korrelation und/oder Regressionskoeffizienten zwischen einzelner Variable und Faktor, also bspw. von Variable Exo-Talk und Faktor Sozialität
- Unabhängige Faktoren $\rightarrow$ Korrelation = Regressionskoeffizienten

Berechnung

Wir zeichnen wieder gerade Linien

Berechnung

Wir zeichnen wieder gerade Linien
$Faktor_m = b_{1m}X_1 + b_{2m}X_2 + \cdots + \epsilon$ mit b_{km} = Factor Loading von Variable k auf Faktor m

Berechnung

Wir zeichnen wieder gerade Linien
$Faktor_m = b_{1m}X_1 + b_{2m}X_2 + \cdots + \epsilon$ mit b_{km} = Factor Loading von Variable k auf Faktor m

- $Sozialität = b_{1Soz}ExoTalk + b_{2Soz}Sozialskill + b_{3Soz}Interesse + b_{4Soz}EgoTalk + b_{5Soz}Selbstsüchtig + b_{6Soz}LügnerIn + \epsilon$
- $Rücksichtslos = b_{1Ruck}ExoTalk + b_{2Ruck}Sozialskill + b_{3Ruck}Interesse + b_{4Ruck}EgoTalk + b_{5Ruck}Selbstsüchtig + b_{6Ruck}LügnerIn + \epsilon$

Berechnung

Wir zeichnen wieder gerade Linien
$Faktor_m = b_{1m}X_1 + b_{2m}X_2 + \cdots + \epsilon$ mit b_{km} = Factor Loading von Variable k auf Faktor m

- $Sozialität = b_{1Soz}ExoTalk + b_{2Soz}Sozialskill + b_{3Soz}Interesse + b_{4Soz}EgoTalk + b_{5Soz}Selbstsüchtig + b_{6Soz}LügnerIn + \epsilon$
- $Rücksichtslos = b_{1Ruck}ExoTalk + b_{2Ruck}Sozialskill + b_{3Ruck}Interesse + b_{4Ruck}EgoTalk + b_{5Ruck}Selbstsüchtig + b_{6Ruck}LügnerIn + \epsilon$

- $Sozialität = 0.87ExoTalk + 0.96Sozialskill + 0.92Interesse + 0.00EgoTalk - 0.10Selbstsüchtig + 0.09LügnerIn + \epsilon$
- $Rücksichtslos = 0.01ExoTalk - 0.03Sozialskill + 0.04Interesse + 0.82EgoTalk + 0.75Selbstsüchtig + 0.70LügnerIn + \epsilon$

Faktormatrix / Komponentenmatrix (PCA)

- $Sozialität = 0.87 ExoTalk + 0.96 Sozialskill + 0.92 Interesse + 0.00 EgoTalk - 0.10 Selbstsüchtig + 0.09 LügnerIn + \epsilon$
- $Rücksichtslos = 0.01 ExoTalk - 0.03 Sozialskill + 0.04 Interesse + 0.82 EgoTalk + 0.75 Selbstsüchtig + 0.70 LügnerIn + \epsilon$

$$\text{Faktormatrix } A = \begin{pmatrix} 0.87 & 0.01 \\ 0.96 & 0.03 \\ 0.92 & 0.04 \\ 0.00 & 0.82 \\ -0.10 & 0.75 \\ 0.09 & 0.70 \end{pmatrix}$$

- Strukturmatrix verwendet Korrelationen Factor Structure Matrix
- Mustermatrix verwendet Regressionskoeffizienten Factor Pattern Matrix
- Bei Orthogonalen Faktoren austauschbar, aber sonst verschieden zu interpretieren

  Graham, J.M. &Guthrie, A.C. & Thompson, B. (2003): *Consequences of not interpreting structurecoefficients in published CFA research: A reminder*

Faktorscore der Probanden

- $Sozialität_{ute} = 0.87 ExoTalk + 0.96 Sozialskill + 0.92 Interesse + 0.00 EgoTalk - 0.10 Selbstsüchtig + 0.09 LügnerIn + \epsilon$
- $Rücksichtslos_{ute} = 0.01 ExoTalk + -0.03 Sozialskill + 0.04 Interesse + 0.82 EgoTalk + 0.75 Selbstsüchtig + 0.70 LügnerIn + \epsilon$

Triviale Lösung
$\rightarrow$

Faktorscore der Probanden

- $Sozialität_{ute} = 0.87 ExoTalk + 0.96 Sozialskill + 0.92 Interesse + 0.00 EgoTalk - 0.10 Selbstsüchtig + 0.09 LügnerIn + \epsilon$
- $Rücksichtslos_{ute} = 0.01 ExoTalk + -0.03 Sozialskill + 0.04 Interesse + 0.82 EgoTalk + 0.75 Selbstsüchtig + 0.70 LügnerIn + \epsilon$

Triviale Lösung
$\rightarrow$ Einfach Messwerte für Person einsetzen

- $Sozialität_{ute} = 0.87 * 4 + 0.96 * 9 + 0.92 * 8 + 0.00 * 6 - 0.10 * 8 + 0.09 * 6 = 19.22$
- $Rücksichtslos_{ute} = 0.01 * 4 - 0.03 * 9 + 0.04 * 8 + 0.82 * 6 + 0.75 * 8 + 0.70 * 6 = 15.21$

Faktorscore der Probanden

- $Sozialität_{ute} = 0.87 ExoTalk + 0.96 Sozialskill + 0.92 Interesse + 0.00 EgoTalk - 0.10 Selbstsüchtig + 0.09 LügnerIn + \epsilon$
- $Rücksichtslos_{ute} = 0.01 ExoTalk + -0.03 Sozialskill + 0.04 Interesse + 0.82 EgoTalk + 0.75 Selbstsüchtig + 0.70 LügnerIn + \epsilon$

Triviale Lösung
$\rightarrow$ Einfach Messwerte für Person einsetzen

- $Sozialität_{ute} = 0.87 * 4 + 0.96 * 9 + 0.92 * 8 + 0.00 * 6 - 0.10 * 8 + 0.09 * 6 = 19.22$
- $Rücksichtslos_{ute} = 0.01 * 4 - 0.03 * 9 + 0.04 * 8 + 0.82 * 6 + 0.75 * 8 + 0.70 * 6 = 15.21$

ACHTUNG: Scores von Faktoren mit verschiedenen Skalen untereinander nicht vergleichbar 

Faktorscore der Probanden

Scores von Faktoren mit verschiedenen Skalen untereinander nicht vergleichbar☹

$\rightarrow$

Faktorscore der Probanden

Scores von Faktoren mit verschiedenen Skalen untereinander nicht vergleichbar☹

→ Wir normalisieren die Faktormatrix A mit der ursprünglichen Korrelationsmatrix COR

- $\frac{A}{COR} = A * COR^{-1}$ Siehe MANOVA Vorlesung

$$\begin{pmatrix} 0.87 & 0.01 \\ 0.96 & 0.03 \\ 0.92 & 0.04 \\ 0.00 & 0.82 \\ -0.10 & 0.75 \\ 0.09 & 0.70 \end{pmatrix} * \begin{pmatrix} 4.76 & -7.46 & 3.91 & -2.15 & 2.42 & -0.49 \\ -7.46 & 18.49 & -12.42 & 5.45 & -5.54 & 1.22 \\ 3.91 & -12.42 & 10.07 & -3.65 & 3.79 & -0.96 \\ -2.35 & 5.45 & -3.65 & 2.97 & -2.16 & 0.02 \\ 2.42 & -5.54 & 3.79 & -2.16 & 2.98 & -0.56 \\ -0.49 & 1.22 & -0.96 & 0.02 & -0.56 & 1.27 \end{pmatrix} =$$

$$\begin{pmatrix} 0.343 & 0.006 \\ 0.376 & -0.020 \\ 0.362 & 0.020 \\ 0.000 & 0.473 \\ -0.037 & 0.437 \\ 0.039 & 0.405 \end{pmatrix}$$

Faktorscore der Probanden

Scores von Faktoren mit verschiedenen Skalen untereinander nicht vergleichbar☹

→ Wir normalisieren die Faktormatrix A mit der ursprünglichen Korrelationsmatrix COR

- $\frac{A}{COR} = A * COR^{-1}$ Siehe MANOVA Vorlesung

$$\begin{pmatrix} 0.87 & 0.01 \\ 0.96 & 0.03 \\ 0.92 & 0.04 \\ 0.00 & 0.82 \\ -0.10 & 0.75 \\ 0.09 & 0.70 \end{pmatrix} * \begin{pmatrix} 4.76 & -7.46 & 3.91 & -2.15 & 2.42 & -0.49 \\ -7.46 & 18.49 & -12.42 & 5.45 & -5.54 & 1.22 \\ 3.91 & -12.42 & 10.07 & -3.65 & 3.79 & -0.96 \\ -2.35 & 5.45 & -3.65 & 2.97 & -2.16 & 0.02 \\ 2.42 & -5.54 & 3.79 & -2.16 & 2.98 & -0.56 \\ -0.49 & 1.22 & -0.96 & 0.02 & -0.56 & 1.27 \end{pmatrix} =$$

$$\begin{pmatrix} 0.343 & 0.006 \\ 0.376 & -0.020 \\ 0.362 & 0.020 \\ 0.000 & 0.473 \\ -0.037 & 0.437 \\ 0.039 & 0.405 \end{pmatrix}$$

← Adjustierte Faktoren / Faktorscore-Koeffizienten

Faktorscore der Probanden

$$\begin{pmatrix} 0.343 & 0.006 \\ 0.376 & -0.020 \\ 0.362 & 0.020 \\ 0.000 & 0.473 \\ -0.037 & 0.437 \\ 0.039 & 0.405 \end{pmatrix} \leftarrow \text{Adjustierte Faktoren / Faktorscore-Koeffizienten}$$

- $Sozialität_{ute} =$
  $0.343 * 4 + 0.376 * 9 + 0.362 * 8 + 0.00 * 6 - 0.037 * 8 + 0.039 * 6 = 7.59$
- $Rücksichtslos_{ute} =$
  $0.006 * 4 - 0.020 * 9 + 0.020 * 8 + 0.473 * 6 + 0.437 * 8 + 0.405 * 6 = 8.768$

Faktorscore der Probanden

$$\begin{pmatrix} 0.343 & 0.006 \\ 0.376 & -0.020 \\ 0.362 & 0.020 \\ 0.000 & 0.473 \\ -0.037 & 0.437 \\ 0.039 & 0.405 \end{pmatrix} \leftarrow \text{Adjustierte Faktoren / Faktorscore-Koeffizienten}$$

- $Sozialität_{ute} =$
  $0.343 * 4 + 0.376 * 9 + 0.362 * 8 + 0.00 * 6 - 0.037 * 8 + 0.039 * 6 = 7.59$
- $Rücksichtslos_{ute} =$
  $0.006 * 4 - 0.020 * 9 + 0.020 * 8 + 0.473 * 6 + 0.437 * 8 + 0.405 * 6 = 8.768$

Interpretation: Ute erreicht etwa gleich hohe Werte bei Sozialität und Rücksichtslosigkeit

ACHTUNG: Die Scores eines Faktors können mit Variablen anderer Faktoren korrelieren

Überblick

1. Latente Variablen / Faktoren
   - Was?
   - Berechnung
   - Scoring
2. Faktorenidentifizierung
   - Kommunalität
   - Principal Component Analysis vs. Faktoranalyse
   - Faktoranalyse
   - Verlässlichkeitsanalyse
3. Anwendungsbeispiel
   - Fragebögen
   - Faktoranalyse in *R*

Methoden zur Faktorenidentifizierung

2 Szenarien:

- Datenexploration: Diese Vorlesung
- Hypothesentests: Tinsley, H.E.A. & Tinsley, D.J. (1987): *Uses of factor analysis in counseling psychology research*

Kommunalität

hohe Kommunalität ist gut für Faktoranalyse

- Geteilte Varianz: Varianz einer Variable, die sie mit anderen teilt
- Eigene Varianz: Varianz einer Variable, die sie mit niemandem teilt
- Kommunalität $= \frac{Geteilte\ \ Varianz}{Varianz\ \ Insgesamt}$
  - Kommunalität $== 1 \rightarrow$ keine eigene Varianz
  - Kommunalität $== 0 \rightarrow$ keine geteilte Varianz

Kommunalität

hohe Kommunalität ist gut für Faktoranalyse

- Geteilte Varianz: Varianz einer Variable, die sie mit anderen teilt
- Eigene Varianz: Varianz einer Variable, die sie mit niemandem teilt
- Kommunalität $= \frac{Geteilte\ \ Varianz}{Varianz\ \ Insgesamt}$
  - Kommunalität $==1 \rightarrow$ keine eigene Varianz
  - Kommunalität $==0 \rightarrow$ keine geteilte Varianz

2 Wege

- Kommunalität $= 1$ für jede Variable angenommen $\rightarrow$ Principal Component Analysis (PCA)
- Abschätzung der Kommunalität $\rightarrow$ Faktoranalyse
  - je Variable Multiple Regression mit ihr als Outcome und allen anderen als Pradiktoren $\rightarrow$ Multiples R^2 als Kommunalität
  - weitere (weniger häufige) Methoden existieren

Principal Component Analysis vs. Faktoranalyse

Genauere Abgrenzung:

- Duntemann, G.E. (1989): *Principal component analysis*
- Widaman, K.F. (2007): *Common factors versus components: Principals and principles: errors and misconceptions*

Principal Component Analysis vs. Faktoranalyse

Genauere Abgrenzung:

- Duntemann, G.E. (1989): *Principal component analysis*
- Widaman, K.F. (2007): *Common factors versus components: Principals and principles: errors and misconceptions*
- wenig Unterschiede im Ergebnis bei mehr als 30 Variablen und Kommunalität> 0.7
- erwartbare Unterschiede im Ergebnis bei weniger als 20 Variablen und Kommunalität< 0.4

Stevens, J. P. (2002): *Applied multivariate statistics for the social sciences*

Principal Component Analysis vs. Faktoranalyse

Genauere Abgrenzung:

- Duntemann, G.E. (1989): *Principal component analysis*
- Widaman, K.F. (2007): *Common factors versus components: Principals and principles: errors and misconceptions*
- wenig Unterschiede im Ergebnis bei mehr als 30 Variablen und Kommunalität>0.7
- erwartbare Unterschiede im Ergebnis bei weniger als 20 Variablen und Kommunalität<0.4

Stevens, J. P. (2002): *Applied multivariate statistics for the social sciences*

"component analysis is at best a common factor analysis and at worst an unrecognizable hodgepodge of things from which nothing can be determined" Cliff, N. (1987): *Analyzing multivariate data*

Principal Component Analysis vs. Faktoranalyse

Genauere Abgrenzung:

- Duntemann, G.E. (1989): *Principal component analysis*
- Widaman, K.F. (2007): *Common factors versus components: Principals and principles: errors and misconceptions*
- wenig Unterschiede im Ergebnis bei mehr als 30 Variablen und Kommunalität>0.7
- erwartbare Unterschiede im Ergebnis bei weniger als 20 Variablen und Kommunalität<0.4

Stevens, J. P. (2002): *Applied multivariate statistics for the social sciences*

"component analysis is at best a common factor analysis and at worst an unrecognizable hodgepodge of things from which nothing can be determined" Cliff, N. (1987): *Analyzing multivariate data*

$\rightarrow$ Unterschied laut Field vernachlässigbar, wir verwenden die Begriffe austauschbar

Faktoranalyse

Vorgehen ähnlich zu MANOVA, aber mit Korrelationsmatrix (und ohne Gruppen)

- Variaten der Korrelationsmatrix berechnen
- Anzahl der Variaten = Anzahl der Variablen
- Variaten = Komponenten
- Variate = Eigenvektoren der Matrix (lineare Funktion)
- Factor Loadings = (Eigen-)Werte der Eigenvektoren
  - Anmerkung zur Veranschaulichung: Eigenwerte analog zu Regressionskoeffizienten bei linearer Regression
- Höchster Eigenwert als Indikator des Einflusses der Variaten

Faktorextraktion

- Höchster Eigenwert als Indikator des Einflusses der Variaten / Faktoren / Komponenten
- $\rightarrow$ Nicht alle Faktoren werden beachtet

Auswahl der Faktoren

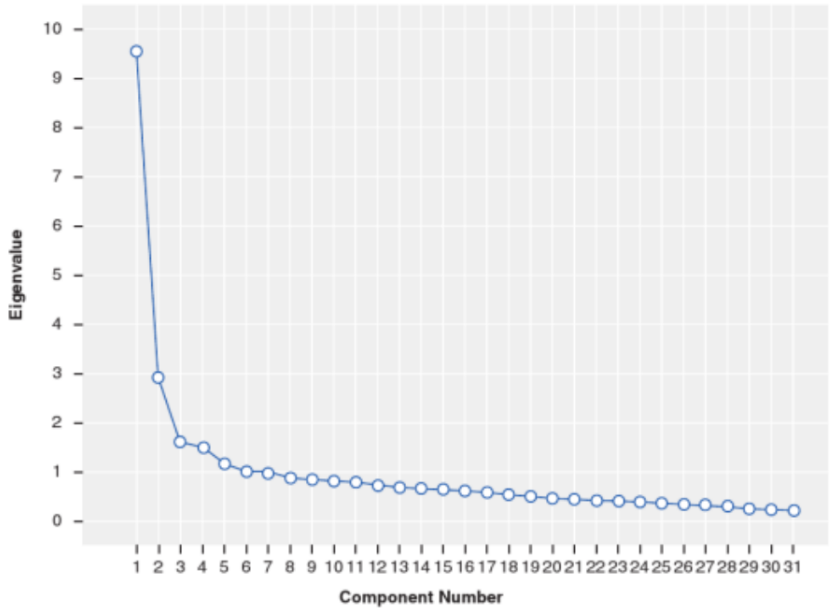
- Kaiser, H.F. (1960): *The application of electronic computers to factor analysis*
  Kaisers Kriterium $\rightarrow$ Alle mit Eigenwerten > 1
- Jolliffe, I.T. (1986): *Principal component analysis*
  Jolliffes Kriterium $\rightarrow$ Alle mit Eigenwerten > 0.7
- Visuell mit Scree Plot

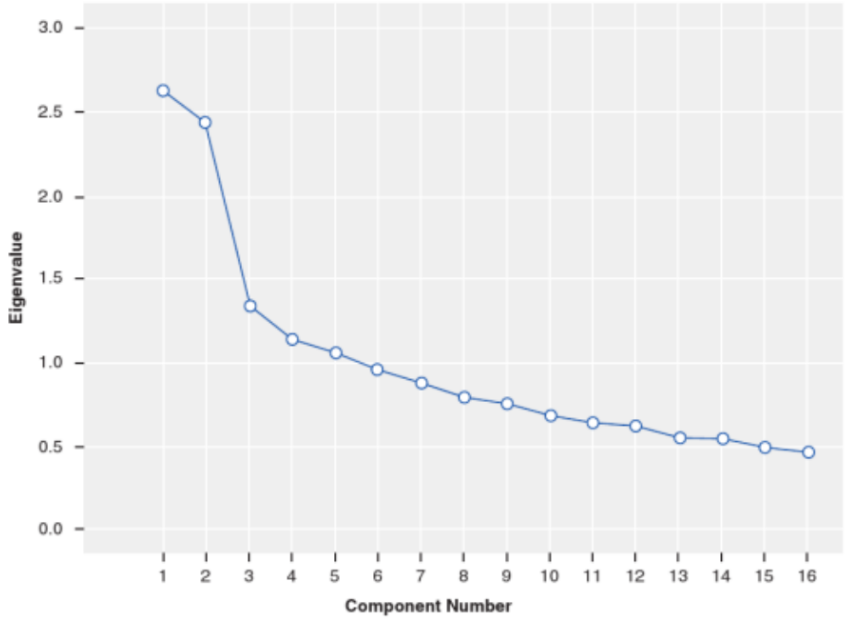
$\rightarrow$ Kaiser möglicherweise zu großzügig, aber verlässlich bei $n > 30$ und Kommunalität > 0.7 oder $n > 250$ und Kommunalität > 0.6; Scree Plot verlässlich bei Stichprobengröße > 200
Anwendungsbezogen (bspw. bei Reparatur von Multikollinearität lieber zu viele als zu wenige Faktoren)

Scree Plot

Cattel, R.B. (1966b): *The scree test for the number of factors*

- Y = Eigenwerte
- X = Komponente / Variate / Faktor
- Abschätzung der Faktoren über Wendepunkt der Kurve

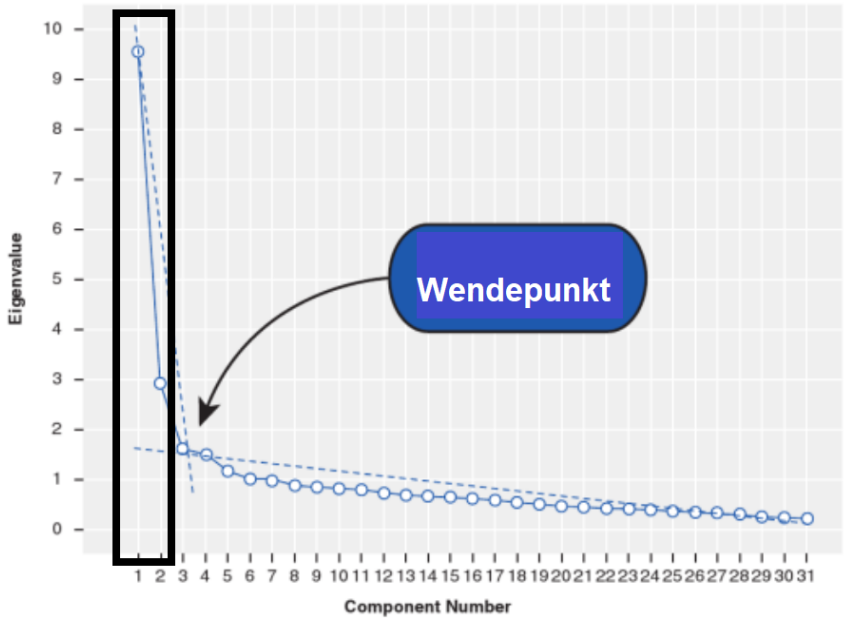


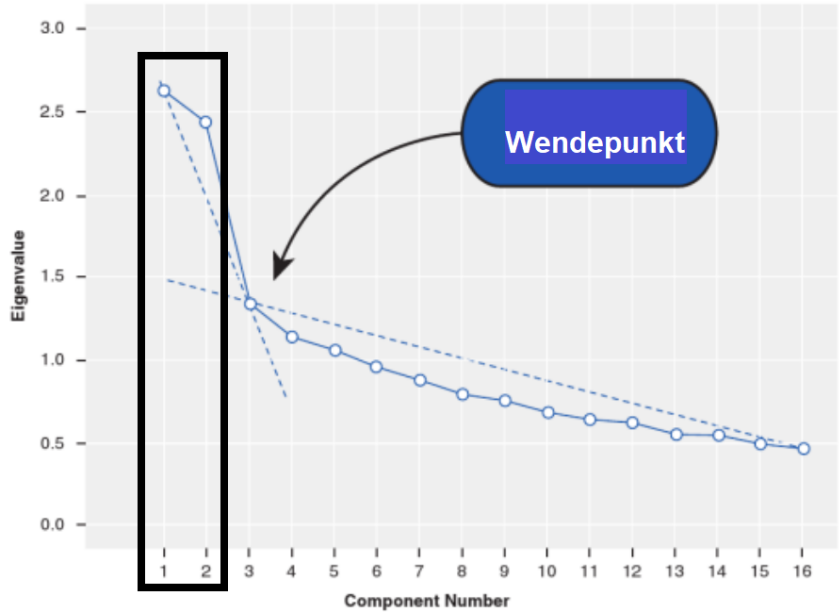


Scree Plot

Cattel, R.B. (1966b): *The scree test for the number of factors*

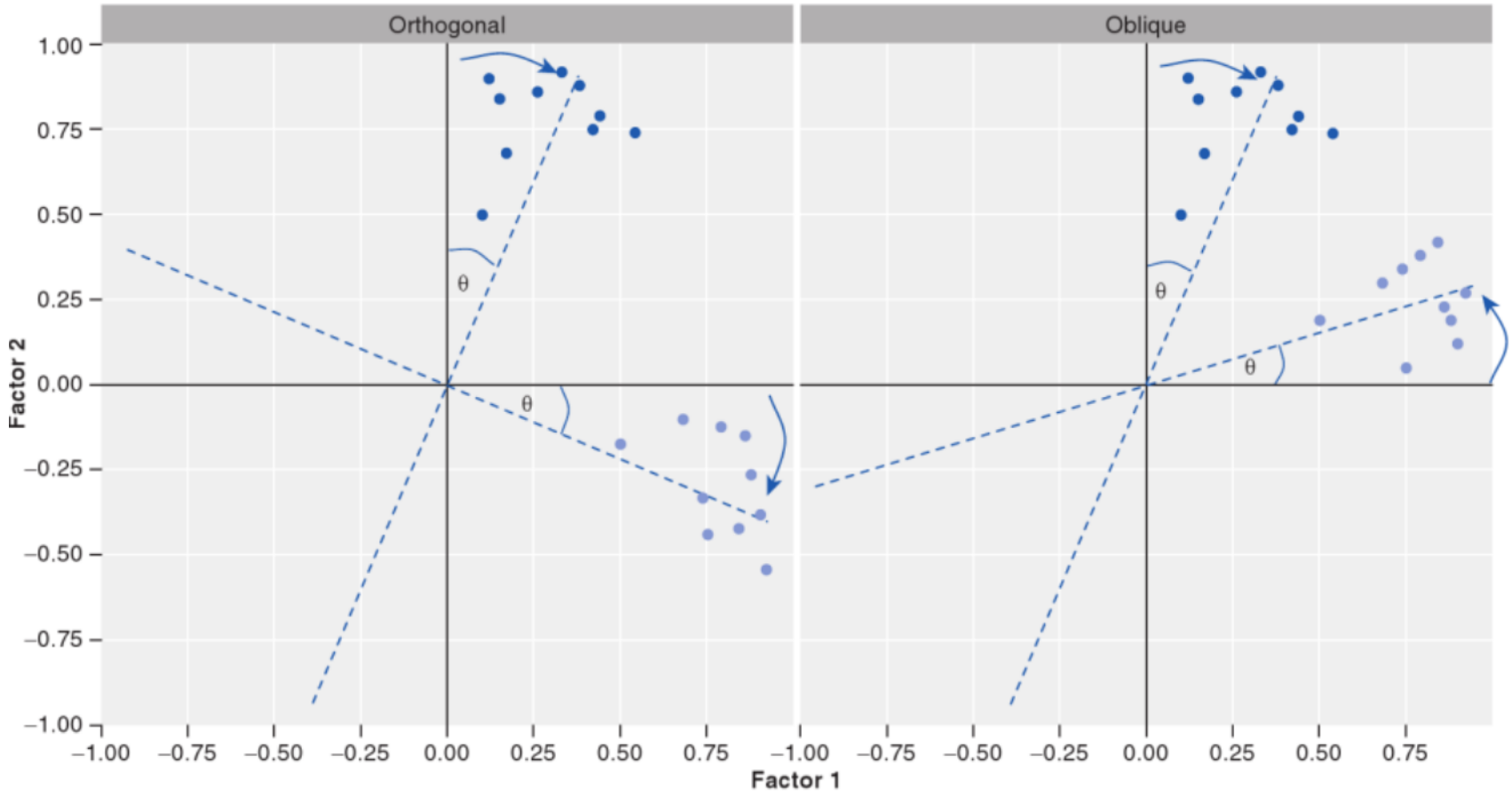
- Y = Eigenwerte
- X = Komponente / Variate / Faktor
- Abschätzung der Faktoren über Wendepunkt der Kurve





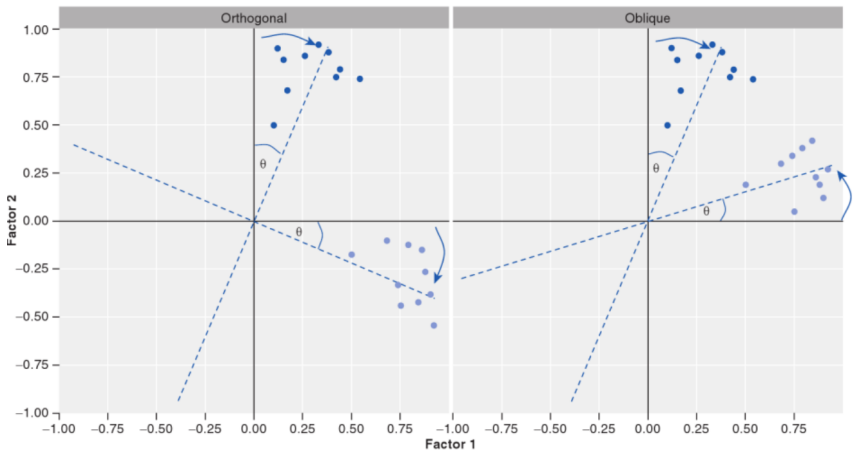
Faktorrotation

Ziel: Verbesserung der Aussagekraft durch Maximierung der starken und Minimierung der schwachen Faktorloadings



Faktorrotation

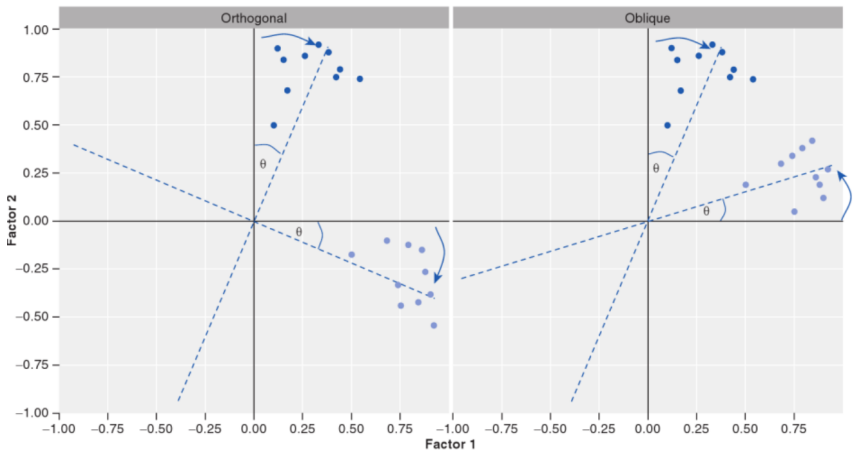
Ziel: Verbesserung der Aussagekraft durch Maximierung der starken und Minimierung der schwachen Faktorloadings



- Orthogonal: Faktoren korrelieren nicht und das wird beibehalten
  $R \rightarrow$ varimax, quartimax(, BentlerT, geominT)
- Oblique: Korrelation zwischen Faktoren erlaubt / angenommen
  $R \rightarrow$ oblimin, promax(, simplimax, BentlerQ, geominQ)

Faktorrotation

Ziel: Verbesserung der Aussagekraft durch Maximierung der starken und Minimierung der schwachen Faktorloadings



- Orthogonal: Faktoren korrelieren nicht und das wird beibehalten
  $R \rightarrow$ varimax, quartimax(, BentlerT, geominT)
- Oblique: Korrelation zwischen Faktoren erlaubt / angenommen
  $R \rightarrow$ oblimin, promax(, simplimax, BentlerQ, geominQ)

Am besten beide beachten und mit deren Unterschieden argumentieren

Faktorvariablen

Welche Variablen sollten zu einem Faktor zählen?
Theoretisch Signifikanztest möglich aber problematisch, deshalb typischerweise einfach Faktorloadings> 0.3

Faktorvariablen

Welche Variablen sollten zu einem Faktor zählen?
Theoretisch Signifikanztest möglich aber problematisch, deshalb typischerweise einfach Faktorloadings> 0.3

$\alpha = 0.01$, *two tailed*

- $n > 50 \rightarrow$ *Loading* > 0.722
- $n > 100 \rightarrow$ *Loading* > 0.512
- $n > 200 \rightarrow$ *Loading* > 0.364
- $n > 300 \rightarrow$ *Loading* > 0.298
- $n > 600 \rightarrow$ *Loading* > 0.210
- $n > 1000 \rightarrow$ *Loading* > 0.162
- weitere siehe Stevens, J. P. (2002): *Applied multivariate statistics for the social sciences*

Faktorvariablen

Welche Variablen sollten zu einem Faktor zählen?
Korrelationen zwischen Variablen beachten
Zu gering

- < 0.3 (Willkürliche Grenze)
- Bartletts Test testet auf Unterschied zu Identitätsmatrix (== geringe Korrelation)
- Signifikant bedeutet Signifikante Korrelationen

Zu hoch

- > 0.8 (Willkürliche Grenze)
- Perfekte Korrelation = Singularität
- Multikorrelation vermeiden (egal bei PCA)
- Determinante der Korrelationsmatrix $> 0.00001 \rightarrow$ Gut
- Berechnung:
  `https://mathworld.wolfram.com/Determinant.html`

Verlässlichkeitsanalyse

Faktoren können als Messgerät für eine Eigenschaft verstanden werden

- Verlässlichkeit (analog zu Vorlesung 2)
  - Erzeugt der Faktor dieselben Scores bei gleichartigen Fällen?
  - Erzeugt das Messgerät dieselben Messwerte in denselben Situationen?
- Split-Half Verlässlichkeit
  - Teile Daten anhand der Variablen zufällig in 2 Teile
  - Scores für beide Hälften sollten ähnlich sein
  - Nachteil: Zufällige Aufteilung erzeugt Schwankung

Verlässlichkeitsanalyse

Faktoren können als Messgerät für eine Eigenschaft verstanden werden

- Verlässlichkeit (analog zu Vorlesung 2)
  - Erzeugt der Faktor dieselben Scores bei gleichartigen Fällen?
  - Erzeugt das Messgerät dieselben Messwerte in denselben Situationen?
- Split-Half Verlässlichkeit
  - Teile Daten anhand der Variablen zufällig in 2 Teile
  - Scores für beide Hälften sollten ähnlich sein
  - Nachteil: Zufällige Aufteilung erzeugt Schwankung
- Cronbachs α

Cronbachs α

- Cronbachs $\alpha \approx$ Konzeptionell Durchschnitt der Korrelationskoeffizienten aller möglichen Splits
- Berechnung siehe Cronbach (1951): *Coefficient alpha and the internal structure of tests*

Cronbachs α

- Cronbachs $\alpha \approx$ Konzeptionell Durchschnitt der Korrelationskoeffizienten aller möglichen Splits
- Berechnung siehe Cronbach (1951): *Coefficient alpha and the internal structure of tests*

Interpretation

- Negative Werte zeigen gespiegelte Variablen (besonders relevant bei Fragebögen)
- > 0.8 ist gut, > 0.7 ist ok
- Steigt mit Anzahl der Variablen
- . . . Die Experten sind sich uneinig 

Überblick

1. Latente Variablen / Faktoren
   - Was?
   - Berechnung
   - Scoring
2. Faktorenidentifizierung
   - Kommunalität
   - Principal Component Analysis vs. Faktoranalyse
   - Faktoranalyse
   - Verlässlichkeitsanalyse
3. Anwendungsbeispiel
   - Fragebögen
   - Faktoranalyse in *R*

Fragebögen

- Fragebögen als Anwendungsszenario
- Einzelne Fragen als Variablen
- Fragenbündel als Faktoren, die eine gewisse Eigenschaft messen
- Do's and Dont's der Fragebogenerstellung $\rightarrow$ Siehe Moodle

Fragebögen und Stichprobengröße

- . . . Die Experten sind sich uneinig 
- 300 ist gut, 1000 super
- Abhängig von Kommunalität
  - $> 0.6 \rightarrow$ 100 ok
  - $> 0.5 \rightarrow$ 100 bis 200 ok
  - $< 0.5 \rightarrow$ 500
  - Kayser-Mayer-Olkin Maß
    - 0 (schlecht) . . . 1 (gut)
    - $< 0.5 \rightarrow$ Faktoranalyse ungeeignet
    - $0.5 \ldots 0.7 \rightarrow$ Mittelmäßig
    - $0.7 \ldots 0.8 \rightarrow$ gut
    - $0.8 \ldots 0.9 \rightarrow$ sehr gut
    - $> 0.9 \rightarrow$ Superb

Beispiel *R* Anxiety Questionnaire

SD = Strongly Disagree, D = Disagree, N = Neither, A = Agree, SA = Strongly Agree

		SD	D	N	A	SA
1	Statistics make me cry	O	O	O	O	O
2	My friends will think I'm stupid for not being able to cope with R	O	O	O	O	O
3	Standard deviations excite me	O	O	O	O	O
4	I dream that Pearson is attacking me with correlation coefficients	O	O	O	O	O
5	I don't understand statistics	O	O	O	O	O
6	I have little experience of computers	O	O	O	O	O
7	All computers hate me	O	O	O	O	O
8	I have never been good at mathematics	O	O	O	O	O
9	My friends are better at statistics than me	O	O	O	O	O
10	Computers are useful only for playing games	O	O	O	O	O
11	I did badly at mathematics at school	O	O	O	O	O
12	People try to tell you that R makes statistics easier to understand but it doesn't	O	O	O	O	O
13	I worry that I will cause irreparable damage because of my incompetence with computers	O	O	O	O	O
14	Computers have minds of their own and deliberately go wrong whenever I use them	O	O	O	O	O
15	Computers are out to get me	O	O	O	O	O
16	I weep openly at the mention of central tendency	O	O	O	O	O
17	I slip into a coma whenever I see an equation	O	O	O	O	O
18	R always crashes when I try to use it	O	O	O	O	O
19	Everybody looks at me when I use R	O	O	O	O	O
20	I can't sleep for thoughts of eigenvectors	O	O	O	O	O
21	I wake up under my duvet thinking that I am trapped under a normal distribution	O	O	O	O	O
22	My friends are better at R than I am	O	O	O	O	O
23	If I am good at statistics people will think I am a nerd	O	O	O	O	O

Faktoranalyse in *R*

Daten:

- 23 Fragen mit 5 Punkte Likert Skala
- 2571 Antworten (also offensichtlich fiktive Daten)

Q01	Q02	...	Q22	Q23
4	5	...	2	4

Siehe raq.dat im Moodle

```
library(corpcor)
library(GPArotation)
library(psych)

raqData<-read.delim("raq.dat", header = TRUE)
raqMatrix<-cor(raqData)   # Korrelationsmatrix
round(raqMatrix, 2)
```

Korrelationsmatrix

```
      Q01   Q02   Q03   Q04   Q05   Q06   Q07  ...
Q01  1.00 -0.10 -0.34  0.44  0.40  0.22  0.31  ...
Q02 -0.10  1.00  0.32 -0.11 -0.12 -0.07 -0.16  ...
Q03 -0.34  0.32  1.00 -0.38 -0.31 -0.23 -0.38  ...
...
Q21  0.33 -0.20 -0.42  0.41  0.33  0.27  0.48  ...
Q22 -0.10  0.23  0.20 -0.10 -0.13 -0.17 -0.17  ...
Q23  0.00  0.10  0.15 -0.03 -0.04 -0.07 -0.07  ...
```

Faktoranalyse in *R*

Korrelationsmatrix

```
      Q01   Q02   Q03   Q04   Q05   Q06   Q07  ...
Q01  1.00 -0.10 -0.34  0.44  0.40  0.22  0.31  ...
Q02 -0.10  1.00  0.32 -0.11 -0.12 -0.07 -0.16  ...
Q03 -0.34  0.32  1.00 -0.38 -0.31 -0.23 -0.38  ...
...
Q21  0.33 -0.20 -0.42  0.41  0.33  0.27  0.48  ...
Q22 -0.10  0.23  0.20 -0.10 -0.13 -0.17 -0.17  ...
Q23  0.00  0.10  0.15 -0.03 -0.04 -0.07 -0.07  ...
```

Fehlermeldung *NaNs produced* $\rightarrow$ non positive definite matrix.

- Sackgasse, Daten sind schlecht 
- Singularität in den Daten, zu wenig Antworten, ...
- Eventuell Variablen reduzieren oder mehr Antworten sammeln

Bartletts Test

```
cortest.bartlett(raqData) # Von Daten
cortest.bartlett(raqMatrix, n = 2571) # Von Kor.Matrix

$chisq
[1] 19334.49
$p.value
[1] 0       #<-- Wahrscheinlichkeit < 0.01 -> PCA angemessen
$df         #Korrelationen sind hoch genug
[1] 253
```

Faktoranalyse in *R*

Kayser-Mayer-Olkin Test

- Nicht in R enthalten
- Function by G. Jay Kerns `http://tolstoy.newcastle.edu.au/R/e2/help/07/08/22816.html`

```
kmo(raqData)

$overall
[1] 0.9302245
$report
[1] "The KMO test yields a degree of common variance marvelous."

$individual
      Q01       Q02       Q03    ...
0.9297610 0.8747754 0.9510378    ...
```

- Entfernung von Variablen mit Individuellen KMO < 0.5 sinnvoll
- Wiederholung nach Entfernung

Faktoranalyse in *R*

Determinante

```
det(raqMatrix)
```

```
[1] 0.0005271037 # > 0.00001, also gut
```

Faktoranalyse (unrotiert, Jede Variable ist ein Faktor)

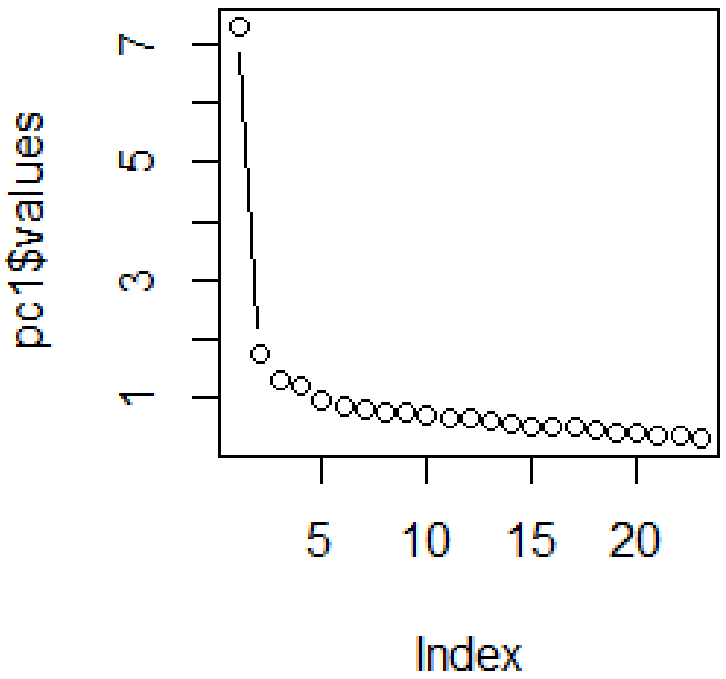
```
pc1 <-  principal(raqData, nfactors = length(raqData), rotate = "none")
pc1

Call: principal(r = raqData, nfactors = length(raqData), rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1   PC2   PC3   PC4   PC5   PC6   ...
Q01  0.59  0.18 -0.22  0.12 -0.40 -0.11   ...
Q02 -0.30  0.55  0.15  0.01 -0.03 -0.38   ...
...
     PC21  PC22  PC23 h2       u2 com
Q01 -0.21  0.05  0.01  1 -1.1e-15 6.0
Q02 -0.02  0.03  0.02  1 -3.8e-15 6.1
# h2 = Kommunalitäten (alle 1 weil jede Variable ein Faktor ist)
# u2 = Uniqueness = 1 - Kommunalität


                       PC1  PC2 ...
SS loadings           7.29 1.74 ... # Varianz erklärt durch Faktor
Proportion Var        0.32 0.08 ... # Anteilig (7.29 / 23)
Cumulative Var        0.32 0.39 ...
```

Scree Plot

```
plot(pc1$values, type = "b")
```



Interpretation schwierig, aber 4 Faktoren scheinen sinnvoll

Faktoranalyse in *R*

Faktoranalyse (unrotiert, 4 Faktoren)

```
pc2 <-  principal(raqData, nfactors = 4, rotate = "none")
pc2

Call: principal(r = raqData, nfactors = 4, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1   PC2   PC3   PC4   h2   u2 com
Q01  0.59  0.18 -0.22  0.12 0.43 0.57 1.6
Q02 -0.30  0.55  0.15  0.01 0.41 0.59 1.7
...
                       PC1  PC2  PC3  PC4
SS loadings           7.29 1.74 1.32 1.23
Proportion Var        0.32 0.08 0.06 0.05
Cumulative Var        0.32 0.39 0.45 0.50
```

Abgesehen von den Kommunalitäten und der Faktoranzahl hat sich nichts geändert

Faktoranalyse in *R*

Faktoranalyse (Rotiert, 4 Faktoren)

```
pc3 <-  principal(raqData, nfactors = 4, rotate = "varimax")
print.psych(pc3, cut = 0.3, sort = FALSE)  # Filter und Sortieren nach Loading

Standardized loadings (pattern matrix) based upon correlation matrix
    item   RC3   RC1   RC4   RC2   h2   u2 com
Q06    6  0.80                    0.65 0.35 1.0
Q18   18  0.68  0.33              0.60 0.40 1.5
...
Q20   20        0.68              0.48 0.52 1.1
Q21   21        0.66              0.55 0.45 1.5
...
Q11   11              0.75        0.69 0.31 1.5
Q09    9                    0.65 0.48 0.52 1.3
...
                       RC3  RC1  RC4  RC2
SS loadings           3.73 3.34 2.55 1.95
Proportion Var        0.16 0.15 0.11 0.08
Cumulative Var        0.16 0.31 0.42 0.50
```

Kommunalitäten haben sich nicht geändert, aber Loadings sind eindeutiger

- RC* sind die Faktoren

Identifizierte Faktoren

SD = Strongly Disagree, D = Disagree, N = Neither, A = Agree, SA = Strongly Agree

		SD	D	N	A	SA
1	Statistics make me cry	O	O	O	O	O
2	My friends will think I'm stupid for not being able to cope with R	O	O	O	O	O
3	Standard deviations excite me	O	O	O	O	O
4	I dream that Pearson is attacking me with correlation coefficients	O	O	O	O	O
5	I don't understand statistics	O	O	O	O	O
6	I have little experience of computers	O	O	O	O	O
7	All computers hate me	O	O	O	O	O
8	I have never been good at mathematics	O	O	O	O	O
9	My friends are better at statistics than me	O	O	O	O	O
10	Computers are useful only for playing games	O	O	O	O	O
11	I did badly at mathematics at school	O	O	O	O	O
12	People try to tell you that R makes statistics easier to understand but it doesn't	O	O	O	O	O
13	I worry that I will cause irreparable damage because of my incompetence with computers	O	O	O	O	O
14	Computers have minds of their own and deliberately go wrong whenever I use them	O	O	O	O	O
15	Computers are out to get me	O	O	O	O	O
16	I weep openly at the mention of central tendency	O	O	O	O	O
17	I slip into a coma whenever I see an equation	O	O	O	O	O
18	R always crashes when I try to use it	O	O	O	O	O
19	Everybody looks at me when I use R	O	O	O	O	O
20	I can't sleep for thoughts of eigenvectors	O	O	O	O	O
21	I wake up under my duvet thinking that I am trapped under a normal distribution	O	O	O	O	O
22	My friends are better at R than I am	O	O	O	O	O
23	If I am good at statistics people will think I am a nerd	O	O	O	O	O

- Faktor 1: Q6, Q18, Q13, Q7, Q14, Q10, Q15 *Angst vor Computern*
- Faktor 2: Q20, Q21, Q3, Q12, Q4, Q16, Q1, Q5 *Angst vor Statistik*
- Faktor 3: Q8, Q17, Q11 *Angst vor Mathematik*
- Faktor 4: Q9, Q22, Q2, Q19 *Angst vor bösem Feedback*

Faktoranalyse in *R*

Faktorscores

```
pc5 <- principal(raqData, nfactors = 4, rotate = "oblimin", scores = TRUE)
pc5$scores
head(pc5$scores, 10) # Nur die ersten 10 anzeigen
```

```
              TC1        TC4         TC3        TC2
 [1,]  0.37296709  1.8808424  0.95979596  0.3910711
 [2,]  0.63334164  0.2374679  0.29090777 -0.3504080
 [3,]  0.39712768 -0.1056263 -0.09333769  0.9249353
 [4,] -0.78741595  0.2956628 -0.77703307  0.2605666
 [5,]  0.04425942  0.6815179  0.59786611 -0.6912687
 [6,] -1.70018648  0.2091685  0.02784164  0.6653081
 [7,]  0.66139239  0.4224096  1.52552021 -0.9805434
 [8,]  0.59491329  0.4060248  1.06465956 -1.0932598
 [9,] -2.34971189 -3.6134797 -1.42999472 -0.5443773
[10,]  0.93504597  0.2285419  0.96735727 -1.5712753
```

Beachte die geänderte Rotation

Score von jedem Proband auf jeden Faktor

- RC* sind die Faktoren
- Variablen, die nur auf 1 Faktor laden, sind sichere Kandidaten

DH - Beispiele

`https://handbuch.tib.eu/w/DH-Handbuch/Tools`

- Strukturen erkennen im hochdimensionalen Raum: Die Principal Component Analysis
- Stilometrie

Luhmann, J. & Burghardt, M. & Tiepmar, J. (2020): *Subrosa: Determining Movie Similiarities based on Subtitles (currently in review)*

- Nicht explizit PCA oder Faktoranalyse, aber ähnliches "Mindset"
- Vergleich verschiedener Arten von Vektoren auf Basis von Filmuntertiteln

Zusammenfassung

- Latente Variablen / Faktoren sind hochkorrelierende Variablencluster
  - Korrelation zwischen 0.3 und 0.8
  - Bartletts Test und Kayser-Mayer-Olkin Maß
- Faktorloading = Einfluss von Variable auf Faktor
- Kommunalität = $\frac{Geteilte\ \ Varianz}{Varianz\ \ Insgesamt}$
- Screeplot zeigt empfehlenswerte Faktorenanzahl
- Rotation optimiert Loadings
- Faktorscores pro ProbandIn berechenbar
- Flexibles Werkzeug mit hohem Willkürfaktor

"It's a good job I'll never have to do that again" *JochenTiepmar*