Testtheorie und Testkonstruktion

MSc Klinische Psychologie und Psychotherapie

SoSe 2024

Prof. Dr. Dirk Ostwald

(7) Explorative Faktorenanalyse

Faktorielle Validität des BDI-II nach Hautzinger, Keller, and Kühner (2006)

In der Untersuchung von Beck et al. (1996) legten Scree-Tests aus explorativen Faktorenanalysen sowohl für die Patientenstichprobe als auch für die Studentenstichprobe die Extraktion zweier Faktoren nahe. Nach parallel durchgeführten Hauptachsenanalysen mit anschließender obliquer (Promax-)Rotation resultierte für die Patientenstichprobe ein Faktor, den die Autoren als "somatisch-affektiven" Faktor bezeichneteten, auf einem zweiten Faktor luden v.a. kognitive Items des BDI II, er wurde demnach als "kognitiver" Faktor bezeichnet. Die beiden obliquen Faktoren korrelierten miteinander zu $r = 0.66$. In der Studentenstichprobe repräsentierte dagegen der erste Faktor die "kognitiv-affektive", der zweite Faktor die "somatische" Dimension. Hier betrug die Korrelation der beiden Faktoren $r = 0.62$.

Beck et al. (1996) fähren dazu an, das das BDI-II zwei hoch korrelierende Faktoren repräsentiert und es daher möglich ist, dass einzelen affektive Items (wie "Traurigkeit" oder "Weinen") in Abhängigkeit der untersuchten Stichprobe einmal auf dem einen, ein anderers Mal auf dem anderen Faktor substanzieller laden."

7) Validität

a) Konstruktvalidität
- Bifaktorenmodell (mit Normstichprobe n =118)

→ durch Maximum Likelihood Methode geschätzt (χ^2 = 0,13; p =.660) : Daten passen gut

→ alle 4 Facetten (interpersonell, affektiv, Lebenswandel, Antisozial) laden substantiell of F1 bzw F2

3. Entwicklung der Belastungsskala

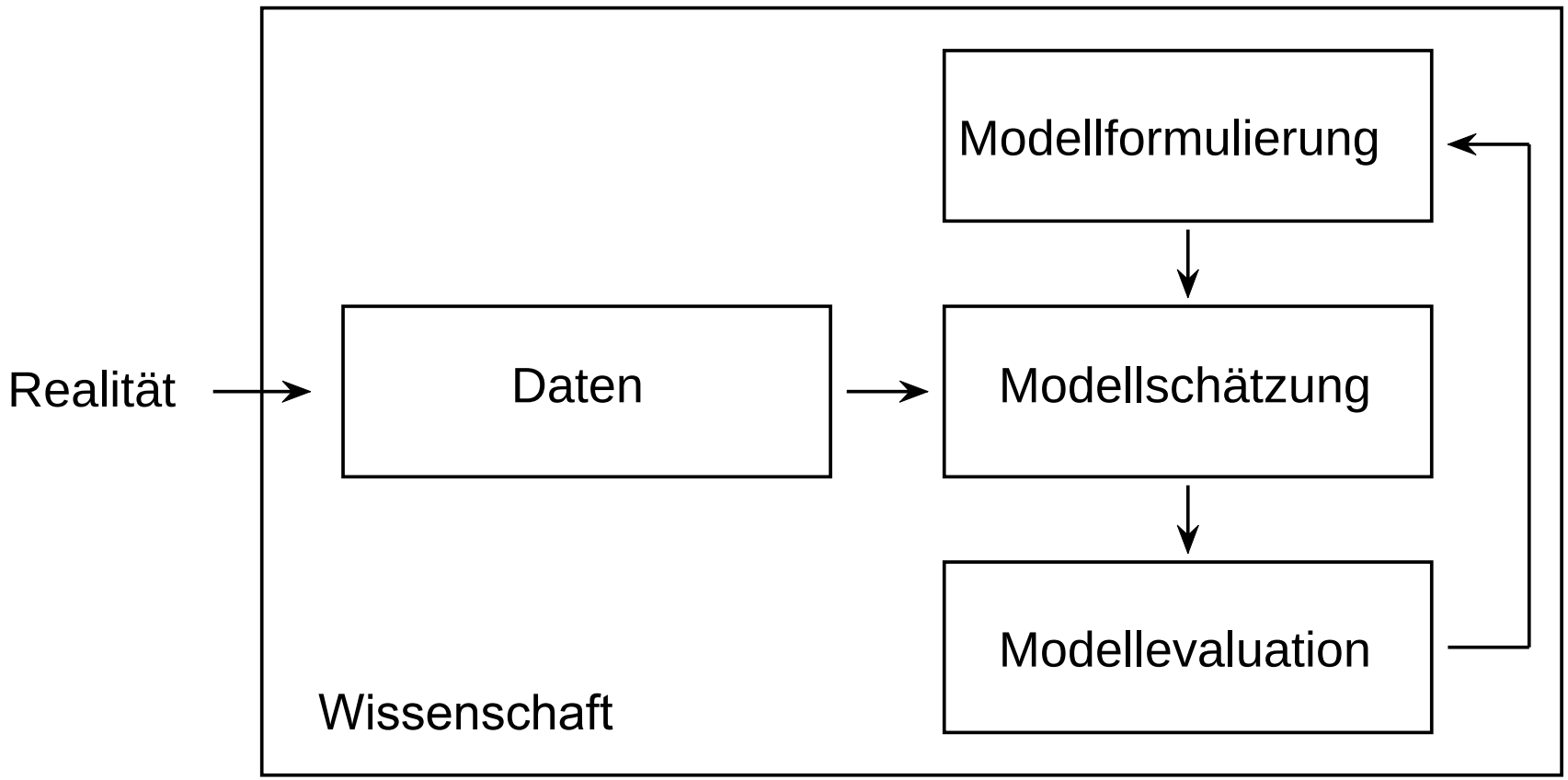
a) Identifikation diskriminierender Items → von ursprünglichen 77 nur 47 Items übernommen
b) Diskriminanzanalytische Validierung → 76,8 % der Fälle korrekt klassifiziert
c) Faktorielle Struktur → 1. Faktor erklärt 44,9 % der Varianz in den Items → Unidimensionalität kann angenommen werden

Validität - faktorielle Validität

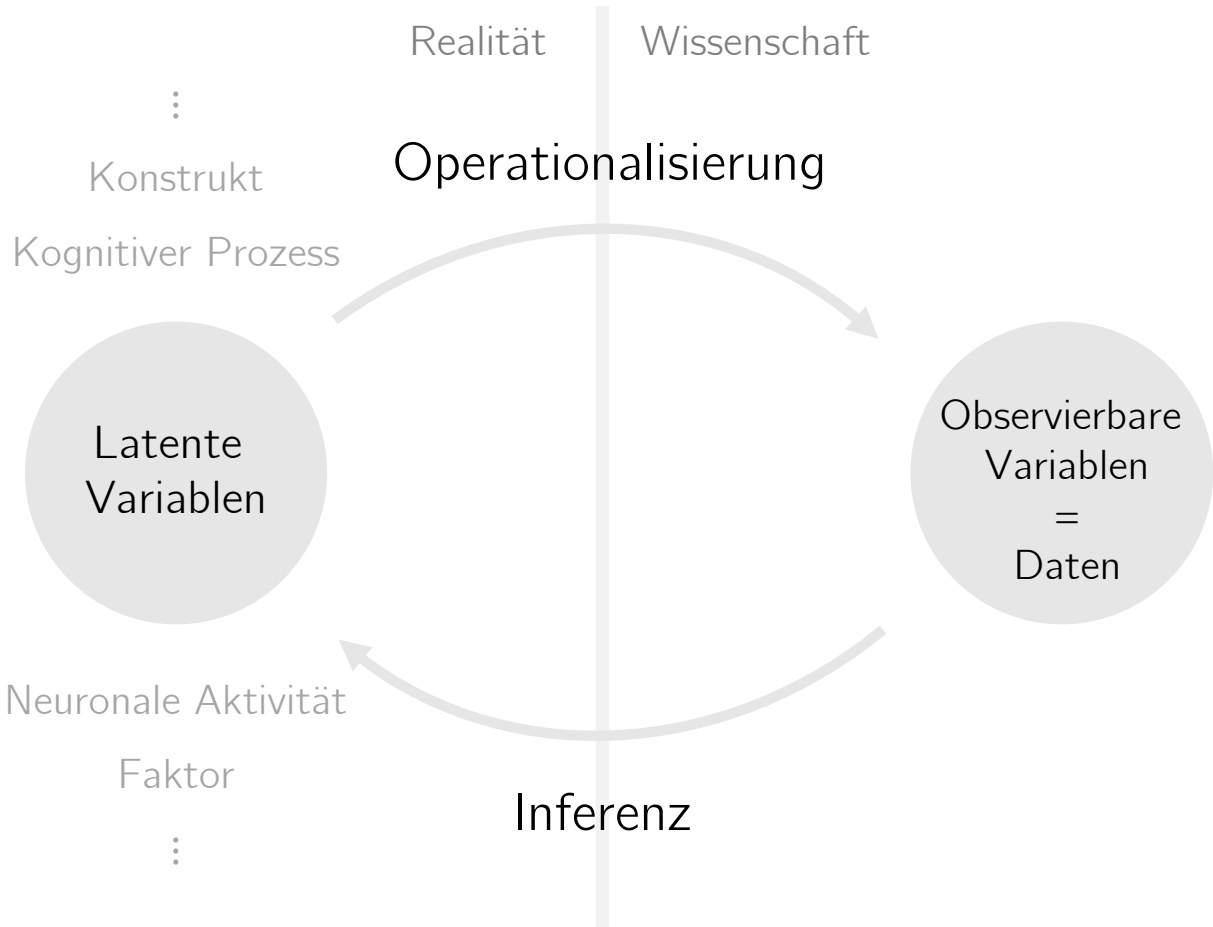
- Einbezogene Daten: 751 Personen (siehe soziodemografische Daten)
- Extraktion von 11 Faktoren und schrittweise Reduktion der Anzahl
- Verschiedene Lösungen wurden danach bewertet, inwieweit eine inhaltlich Interpretation sinn voll erschien
- Von den Faktoranalysen wurden aufgrund inhaltlicher Überlegungen eine Hauptkomponentenanalyse und eine Varimax-Rotation ausgewählt
- Diese 7 Faktoren erklären 54% der Varianz

- Nur 3 Skalen konnten in der rotierten Matrix als eigenständige Faktoren eindeutig identifziert werden
- Die anderen 8 Skalen waren zwar ebenfalls eindeutig, ABER diese Skalen fallen auf einige Faktoren mit den Items anderer Skalen zusammen
- Fazit: uneingeschränkt kann die Validität nur für 3 Skalen bestätigt werden, mit gewissen Einschränkungen für die anderen 8 Skalen

Modellbasierte Datenwissenschaft



Psychologische Datenwissenschaft



Psychologische Datenwissenschaft

Faktorenanalyse

Latente Variablenmodelle mit psychologischer Historie

- Faktorenanalyse und Strukturgleichungsmodelle
- Erklärung von Kovarianzen (vieler) beobachteter Variablen durch (wenige) latente Variablen

Klassische und aktuelle Anwendungsszenarien

- Analyse menschlicher Fähigkeiten (g-Faktor) und Persönlichkeitspsychologie (Big Five)
- Vielzahl von Phänomenen in der Soziologie, Politikwissenschaft, Biologie, Medizin, Sprachwissenschaft, ...

Varianten

Explorative Faktorenanalyse (EFA)

- Datengetriebenes, eher prinzipienfreies und intuitiv geleitetes Inspirationsverfahren
- Fokus auf der numerische Behandlung von Stichprobenkovarianzmatrizen

Konfirmative Faktorenanalyse (CFA)

- Moderneres Verfahren mit expliziter probabilistischer Modellspezifikation
- Fokus auf probabilistischer Modellschätzung und Modellevaluation

Strukturgleichungsmodelle (SEM)

- Generalisierte konfirmative Faktorenanalyse mit Faktoreninteraktion
- Linearer Spezialfall genereller probabilistischer Modelle

Historie

Modellfreie explorative datenzentrische Periode (1900 - 1950)

- Pearson (1901) beschreibt erste Anklänge der Faktorenanalyse
- Spearman (1904) beginnt die Einfaktorenanalyse im Rahmen der Intelligenzforschung
- Hotelling (1933) entwickelt die eng verwandte Hauptkomponentenanalyse
- Thurstone (1947) beginnt die Mehrfaktorenanalyse im Bereich der Psychometrie

Modellbasierte konfirmative inferenzzentrische Periode (1950 - heute)

- Lawley (1940) schlägt die ML-Schätzung basierend auf Fisher (1922) und Wishart (1928) vor.
- Lawley and Maxwell (1962) machen den modellbasierten Charakter der Faktorenanalyse explizit.
- Jöreskog (1970) initiiert die Generalisierung zu Strukturgleichungsmodelle (vgl. Bollen (1989))
- Weitere Generalisierungen (vgl. z.B. Mulaik (2010) oder Bartholomew, Knott, and Moustaki (2011))

Software Periode (1970 - heute)

- lisrel (kommerziell, proprietär) nach Jöreskog (1970)
- mPlus (kommerziell, proprietär) nach Muthén and Muthén (1998/2017)
- lavaan (gratis, quelloffen) nach Rosseel (2012)
- $\Rightarrow$ Faktorenanalyse jeweils als Spezialfall von Strukturgleichungsmodellen

Anwendungsbeispiel

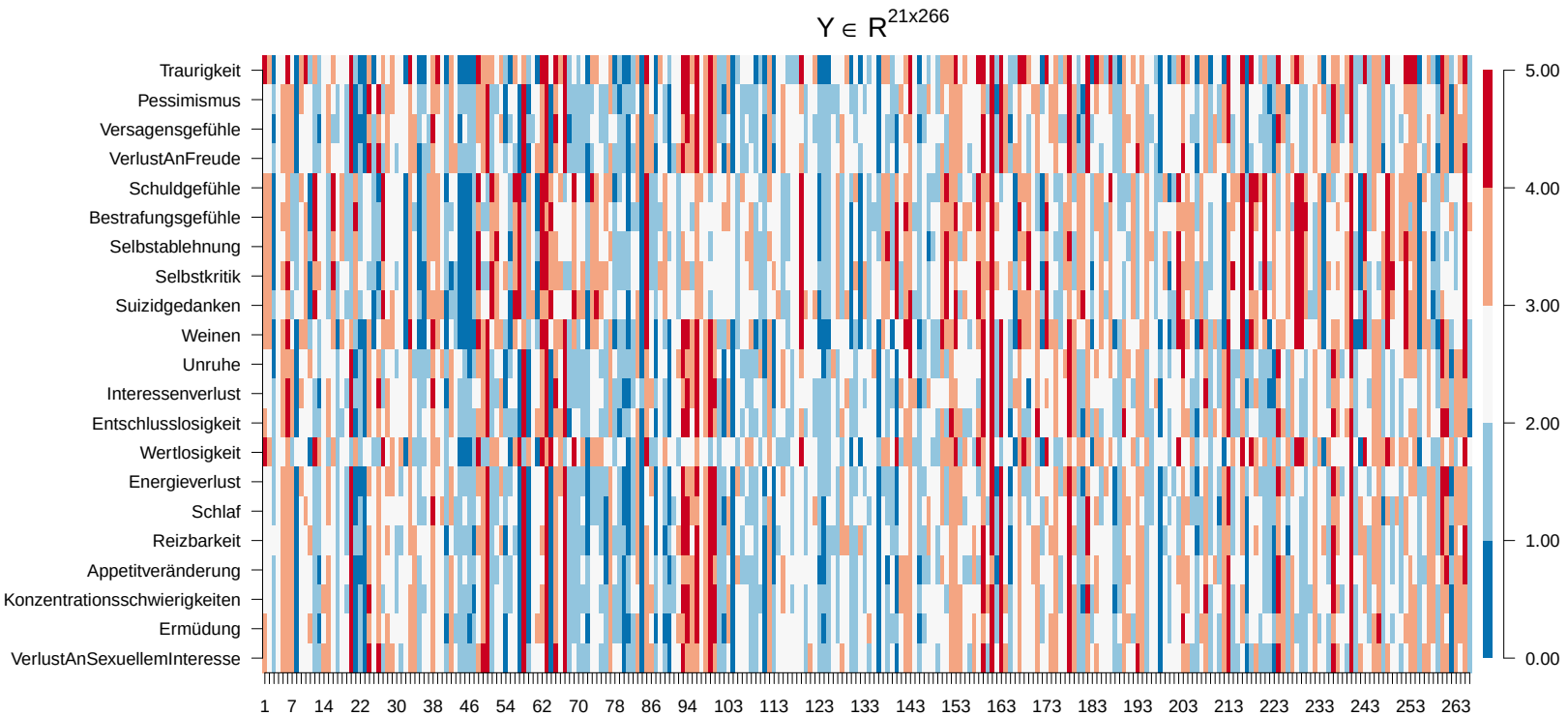
Simulierter Datensatz basierend auf Keller, Hautzinger, and Kühner (2008)

- Patient:innen 1 - 12 von n = 266 depressiven Patient:innen

	1	2	3	4	5	6	7	8	9	10
Traurigkeit	4	3	0	2	2	4	2	0	3	4
Pessimismus	2	2	1	2	3	3	3	0	3	2
Versagensgefühle	2	2	0	2	3	3	3	0	2	2
VerlustAnFreude	2	2	1	2	3	3	3	0	2	2
Schuldgefühle	3	3	0	2	2	3	1	1	3	2
Bestrafungsgefühle	3	3	0	2	3	3	1	1	2	3
Selbstablehnung	3	3	0	2	2	3	1	1	2	3
Selbstkritik	3	3	0	2	3	4	2	1	2	3
Suizidgedanken	3	3	1	2	2	3	1	2	2	3
Weinen	3	3	0	2	3	4	2	0	3	3
Unruhe	2	2	0	2	3	3	3	0	2	2
Interessenverlust	2	2	1	2	3	4	3	0	3	2
Entschlusslosigkeit	3	2	1	2	3	4	3	0	2	2
Wertlosigkeit	4	3	1	2	2	3	1	2	2	2
Energieverlust	2	2	1	2	3	3	3	0	3	2
Schlaf	2	2	1	2	3	3	3	0	2	3
Reizbarkeit	2	2	2	2	3	3	3	0	2	2
Appetitveränderung	2	2	1	2	3	3	3	0	2	2
Konzentrationsschwierigkeiten	2	2	1	2	3	3	3	0	2	2
Ermüdung	3	2	1	2	3	3	3	0	2	2
VerlustAnSexuellemInteresse	3	2	1	2	3	3	3	0	2	2

Anwendungsbeispiel

Beobachteter Datensatz (n = 266)



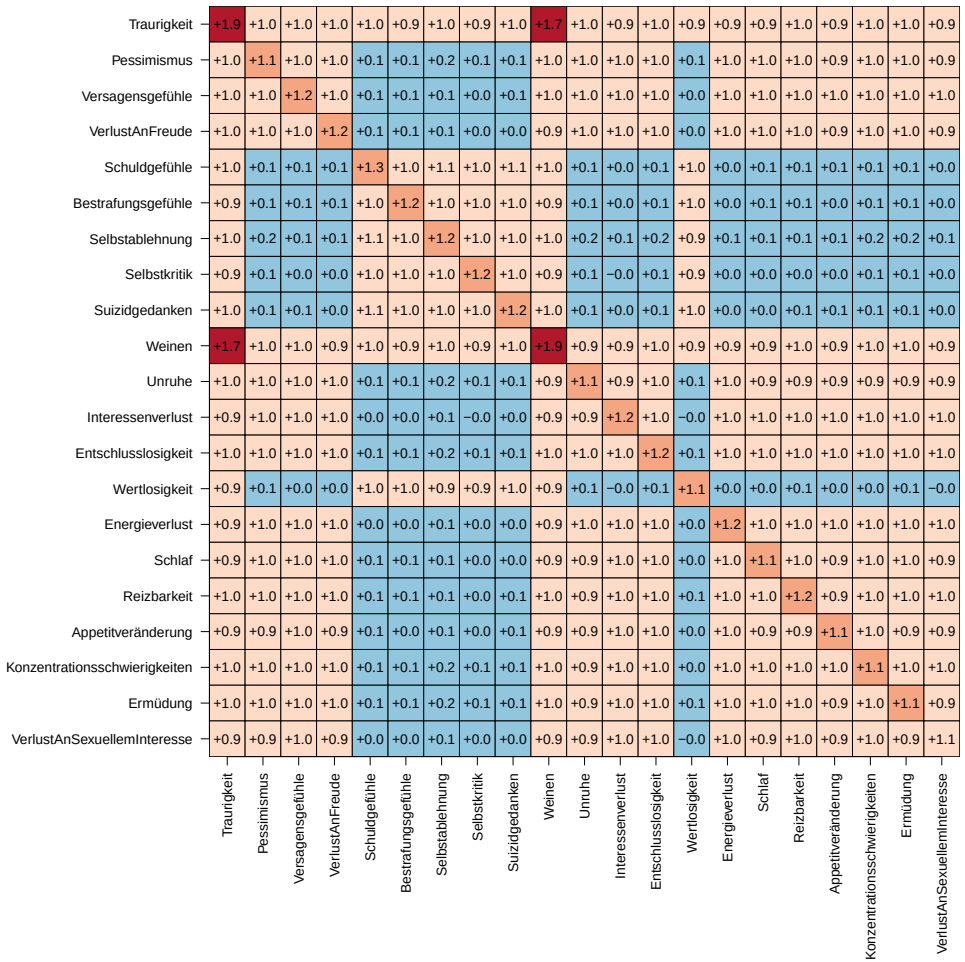
Anwendungsbeispiel

Bestimmung von Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix

```
Y       = t(read.csv("./7_Daten/7_efa.csv"))                # Dateneinlesen
n       = ncol(Y)                                           # Anzahl Datenpunkte
I_n     = diag(n)                                           # Einheitsmatrix I_n
J_n     = matrix(rep(1,n^2), nrow = n)                      # 1_{nn}
C       = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))        # Stichprobenkovarianzmatrix
D       = diag(1/sqrt(diag(C)))                             # Kov-Korr-Transformationsmatrix
R       = D %*% C %*% D                                     # Stichprobenkorrelationsmatrix
```

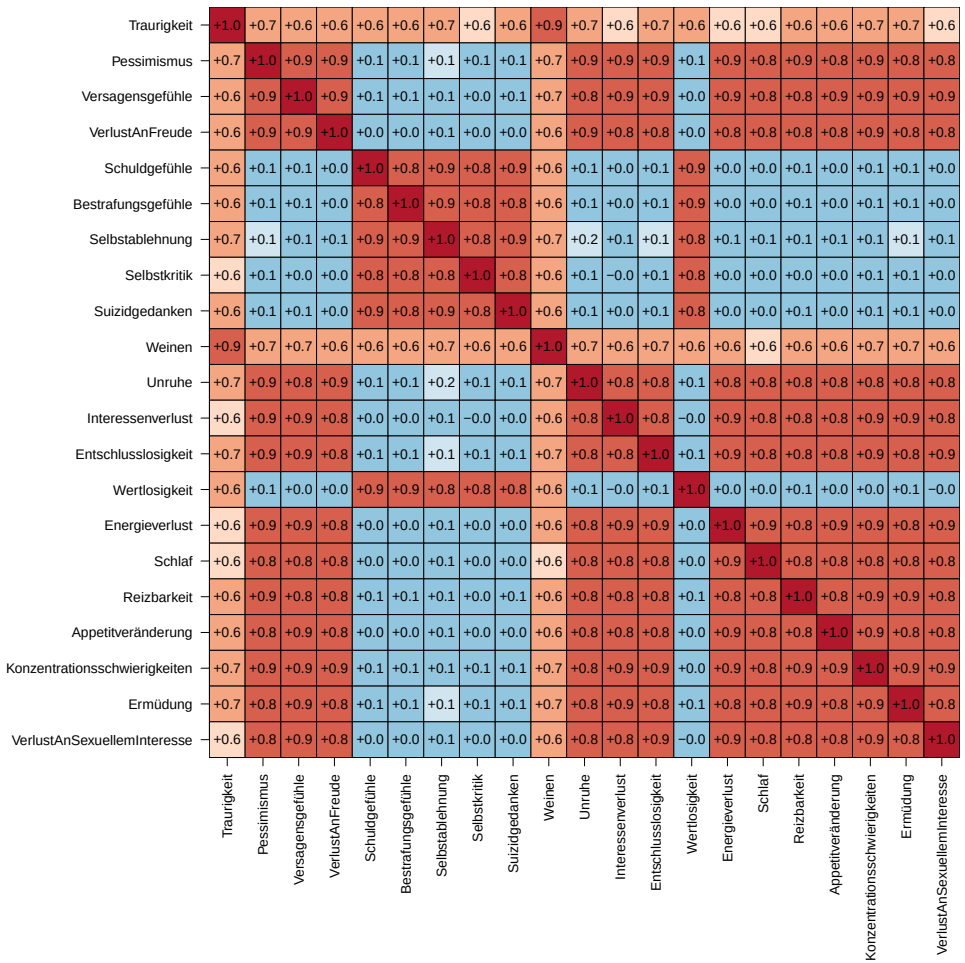
Anwendungsbeispiel

Stichprobenkovarianzmatrix



Anwendungsbeispiel

Stichprobenkorrelationsmatrix



Modellformulierung

Modellschätzung

Modellvergleich

Modellinterpretation

Selbstkontrollfragen

Definition (Modell der explorativen Faktorenanalyse)

Es sei

$$\upsilon = L\xi + \varepsilon \tag{1}$$

wobei für $m > k$

- $L = (l_{ij}) \in \mathbb{R}^{m \times k}$ eine Matrix ist,
- ξ ein k-dimensionaler latenter Zufallsvektor mit $\mathbb{E}(\xi) = 0_k$ und $\mathbb{C}(\xi) = I_k$ ist,
- ε ein m-dimensionaler latenter und von ξ unabhängiger Zufallsvektor ist mit $\mathbb{E}(\varepsilon) = 0_m$ und

$$\mathbb{C}(\varepsilon) = \text{diag}\,(\psi_1, \ldots, \psi_m) =: \Psi \text{ mit } \psi_i > 0 \text{ für } i = 1, \ldots, m \text{ und} \tag{2}$$

- υ ein m-dimensionaler beobachtbarer Zufallsvektor ist.

Dann wird (1) *Modell der explorativen Faktorenanalyse (EFA Modell) mit Parametern L und Ψ* genannt.

Bemerkungen

- Die Komponenten $\xi_i, j = 1, \ldots, k$ von ξ modellieren *(gemeinsame) Faktoren*.
- Die Komponenten $\upsilon_i, i = 1, \ldots, m$ von υ modellieren Datenkomponenten
- Die Datenkomponenten werden sind in der Fragebogendatenanalyse meist Items
- Die Matrix $L = (l_{ij}) \in \mathbb{R}^{m \times k}$ wird *Faktorladungsmatrix* genannt.
- $l_{ij} \in \mathbb{R}$ wird *Faktorladung* der $i$ten Komponente von υ auf den $j$ten Faktor genannt.

Modellformulierung

Bemerkungen (fortgeführt)

- Wir schreiben das EFA Modell im Folgenden meist in der Form

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{3}$$

  wobei die Notation $\zeta \sim (\mu, \Sigma)$ ausdrücken soll, dass $\mathbb{E}(\zeta) = \mu$ und $\mathbb{C}(\zeta) = \Sigma$.

- In generativ-hierarchischer Form kann das Modell der EFA geschrieben werden als

$$\begin{aligned} \xi &= \eta & \eta &\sim (0_k, I_k) \\ \upsilon &= L\xi + \varepsilon & \varepsilon &\sim (0_m, \Psi), \end{aligned} \tag{4}$$

  In dieser Darstellung heißt η *Zustandsrauschen* und ε *Beobachtungsrauschen*.

- In generativ-probabilistischer Form kann das EFA Modell geschrieben werden als

$$\xi \sim (0_k, I_k) \text{ und } \upsilon|\xi \sim (L\xi, \Psi) \tag{5}$$

  Dabei erschließt sich die Bedingtheit von υ gegeben ξ am besten aus generativer Sicht:

  (1) Zunächst wird ein Wert $x \in \mathbb{R}^k$ von ξ realisiert.

  (2) Dieser Wert wird in $Lx \in \mathbb{R}^m$ transformiert.

  (3) Es wird ein Wert $e \in \mathbb{R}^m$ von ε mit Erwartungswert 0_m realisiert.

  (4) Es wird ein Wert $y \in \mathbb{R}^m$ von υ durch Addition von Lx und e realisiert

  Äquivalent zu (3) und (4) wird dabei aber ein Wert y von υ mit *gegebenem* Erwartungswert Lx realisiert.

Modellformulierung

Bemerkungen (fortgeführt)

- Unter Hinzunahme der Annahme multivariat-normalverteilten Zustands- und Beobachtungsrauschens ergibt sich die generativ-probabilistische Form

$$\xi \sim N(0_k, I_k) \text{ und } \upsilon|\xi \sim N(L\xi, \Psi), \tag{6}$$

  Wir werden diese Form im Kontext der konfirmativen Faktoranalyse genauer betrachten.
- In der Theorie resultiert die Generation von n unabhängig und identisch verteilten Realisierungen eines EFA-Modells in einem Datensatz der Form

$$D = \left(\begin{pmatrix} x^{(1)} \\ y^{(1)} \end{pmatrix} \quad \cdots \quad \begin{pmatrix} x^{(n)} \\ y^{(n)} \end{pmatrix} \right) \in \mathbb{R}^{(k+m)\times n} \tag{7}$$

  aus konkatenierten Datenvektoren von latenten und beobachteten Daten.
- In der Anwendung sind die latenten Daten nicht vorhanden, es liegen nur beobachtete Daten

$$Y = \begin{pmatrix} y^{(1)} & \cdots & y^{(n)} \end{pmatrix} \in \mathbb{R}^{m\times n} \tag{8}$$

  vor.

Modellformulierung

Simulationsbeispiel

```
library(MASS)                                  # Multivariates Normalverteilungspaket
k       = 2                                    # Dimension des latenten Zufallsvektors
m       = 21                                   # Dimension des beobachtbaren Zufallsvektors
n       = 266                                  # Beobachtungsanzahl
L       = matrix(c(1,1,                        # Traurigkeit
                   1,0,                        # Pessimismus
                   1,0,                        # Versagensgefühle
                   1,0,                        # Verlust an Freude
                   0,1,                        # Schuldgefühle
                   0,1,                        # Bestrafungsgefühle
                   0,1,                        # Selbstablehnung
                   0,1,                        # Selbstkritik
                   0,1,                        # Suizidgedanken
                   1,1,                        # Weinen
                   1,0,                        # Unruhe
                   1,0,                        # Interessenverlust
                   1,0,                        # Entschlusslosigkeit
                   0,1,                        # Wertlosigkeit
                   1,0,                        # Energieverlust
                   1,0,                        # Schlaf
                   1,0,                        # Reizbarkeit
                   1,0,                        # Appetitveränderung
                   1,0,                        # Konzentrationsschwierigkeiten
                   1,0,                        # Ermüdung
                   1,0),                       # Verlust an sexuellem Interesse
                  nrow  = m,                   # Zeilenanzahl
                  byrow = TRUE)                # Matrixdimensionalität
Psi       = diag(m)                            # Beobachtungsrauschenkovarianmatrix
Y         = matrix(rep(NaN,m*n), nrow = m)     # Simulierte beobachtete Datenmatrix
for(i in 1:n){                                 # Simulationsiterationen
  x       = mvrnorm(1,rep(0,k), diag(k))       # Realisierung des latenten Faktorzufallvektors
  eps     = mvrnorm(1,rep(0,m), Psi)           # Realisierung des latenten Beobachtungsrauschenvektors
  Y[,i]   = mu + L %*% x + eps}                # Realisierung des beobachtbaren Datenzufallsvektors
Y[Y<0]    = 0                                  # Zensur
Y[Y>5]    = 5                                  # Zensur
Y         = round(Y+2, digits = 0)             # Rundung und Offset
```

Theorem (Datenkovarianzmatrix der explorativen Faktorenanalyse)

Gegeben sei ein EFA Modell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{9}$$

Dann gilt für die marginale Kovarianzmatrix des Datenvektors

$$\mathbb{C}(\upsilon) = LL^T + \Psi. \tag{10}$$

Beweis

Mit dem Theorem zu den Eigenschaften der Kovarianzmatrix gilt aufgrund der Unabhängigkeit von ξ und ε

$$\mathbb{C}(\upsilon) = L\mathbb{C}(\xi)L^T + \mathbb{C}(\varepsilon) = LI_kL^T + \Psi = LL^T + \Psi. \tag{11}$$

Bemerkungen

- Basierend auf einem Datensatz $Y \in \mathbb{R}^{m \times n}$ von n Realisierung von υ wird $\mathbb{C}(\upsilon)$ geschätzt durch

$$C = \frac{1}{n-1}\left(Y\left(I_n - \frac{1}{n}1_{nn}\right)Y^T\right). \tag{12}$$

- Wir sehen unten, dass das Ziel der EFA Schätzung die Konstruktion von C durch Schätzer $\hat{L}$ und $\hat{\Psi}$ ist,

$$C = \hat{L}\hat{L}^T + \hat{\Psi}. \tag{13}$$

Theorem (Varianzzerlegung der explorativen Faktorenanalyse)

Gegeben sei ein EFA Modell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_k, \Psi). \tag{14}$$

Dann ist für $i = 1, \ldots, m$ die Varianz der $i$ten Komponente von υ gegeben durch

$$\mathbb{C}(\upsilon_i, \upsilon_i) = \sum_{j=1}^{k} l_{ij}^2 + \psi_i. \tag{15}$$

Bemerkungen

- $\mathbb{C}(\upsilon_i, \upsilon_i)$ ist der $i$te Diagonaleintrag von $\mathbb{C}(\upsilon)$, bekanntermaßen gilt $\mathbb{V}(\upsilon_i) = \mathbb{C}(\upsilon_i, \upsilon_i)$.
- $\mathbb{C}(\upsilon_i, \upsilon_i)$ besteht aus Beiträgen der Faktorladungen und der Beobachtungsrauschenmatrix.
- Wenn l_i die $i$te Zeile von L bezeichnet, dann gilt

$$l_i l_i^T = \begin{pmatrix} l_{i1} & \cdots & l_{ik} \end{pmatrix} \begin{pmatrix} l_{i1} \\ \vdots \\ l_{ik} \end{pmatrix} = \sum_{j=1}^{k} l_{ij}^2 \tag{16}$$

- $\sum_{j=1}^{k} l_{ij}^2$ ist also das Skalarprodukt von l_i^T mit sich selbst.

Beweis

Mit dem Theorem zu Datenkovarianzmatrix der explorativen Faktorenanalyse gilt

$$
\begin{aligned}
\mathbb{C}(\upsilon) &= LL^T + \Psi \\
&= \begin{pmatrix} l_{11} & \cdots & l_{1k} \\ l_{21} & \cdots & l_{2k} \\ \vdots & \ddots & \vdots \\ l_{m1} & \cdots & l_{mk} \end{pmatrix} \begin{pmatrix} l_{11} & \cdots & l_{m1} \\ l_{12} & \cdots & l_{m2} \\ \vdots & \ddots & \vdots \\ l_{1k} & \cdots & l_{mk} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \psi_m \end{pmatrix} \\
&= \begin{pmatrix} \sum_{j=1}^k l_{1j} l_{1j} & \sum_{j=1}^k l_{1j} l_{2j} & \cdots & \sum_{j=1}^k l_{1j} l_{mj} \\ \sum_{j=1}^k l_{2j} l_{1j} & \sum_{j=1}^k l_{2j} l_{2j} & \cdots & \sum_{j=1}^k l_{2j} l_{mj} \\ \vdots & \cdots & \ddots & \vdots \\ \sum_{j=1}^k l_{mj} l_{1j} & \sum_{j=1}^k l_{mj} l_{2j} & \cdots & \sum_{j=1}^k l_{mj} l_{mj} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \psi_m \end{pmatrix} \qquad (17) \\
&= \begin{pmatrix} \sum_{j=1}^k l_{1j}^2 + \psi_1 & \sum_{j=1}^k l_{1j} l_{2j} & \cdots & \sum_{j=1}^k l_{1j} l_{mj} \\ \sum_{j=1}^k l_{2j} l_{1j} & \sum_{j=1}^k l_{2j}^2 + \psi_2 & \cdots & \sum_{j=1}^k l_{2j} l_{mj} \\ \vdots & \cdots & \ddots & \vdots \\ \sum_{j=1}^k l_{mj} l_{1j} & \sum_{j=1}^k l_{mj} l_{2j} & \cdots & \sum_{j=1}^k l_{mj}^2 + \psi_m \end{pmatrix}.
\end{aligned}
$$

□

Definition (Kommunalität, Spezifität, Gesamtvarianz)

Gegeben sei ein EFA Modell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_k, \Psi). \tag{18}$$

Dann werden in

$$\mathbb{C}(\upsilon_i, \upsilon_i) = \sum_{j=1}^{k} l_{ij}^2 + \psi_i \tag{19}$$

$h_i^2 := \sum_{j=1}^{k} l_{ij}^2$ die *Kommunalität* und ψ_i die *Spezifität* von υ_i genannt. Weiterhin wird

$$\mathbb{G} := \sum_{i=1}^{m} \mathbb{C}(\upsilon_i, \upsilon_i) = \sum_{i=1}^{m} \sum_{j=1}^{k} l_{ij}^2 + \sum_{i=1}^{m} \psi_i \tag{20}$$

die *Gesamtvarianz* genannt.

Bemerkungen

- Die Kommunalität ist der Varianzanteil der $i$ten Datenkomponente, die durch die Faktoren erklärt wird.
- Die Spezifität ist der Varianzanteil der $i$ten Datenkomponente, der nicht durch die Faktoren erklärt wird.
- Die Gesamtvarianz ist die Summe der Varianzen der Datenkomponenten.

Modellformulierung

Bemerkungen (fortgeführt)

- Für die $i$te Komponente des Datenvektors v gilt mit $i = 1, \ldots, m$ offenbar die Varianzzerlegung

$$\text{Varianz der } i\text{ten Komponente} \;=\; \text{Kommunalität der } i\text{ten Komponente} \;+\; \text{Spezifität der } i\text{ten Komponente} \tag{21}$$

- Für die Gesamtvarianz gilt offenbar die Varianzzerlegung

$$\text{Gesamtvarianz} \;=\; \text{Summe der Kommunalitäten} \;+\; \text{Summe der Spezifitäten} \tag{22}$$

- Die entsprechende Stichprobenvarianzzerlegung ist Grundlage der Evaluation der Modellgüte.

Definition (Orthogonale Transformation eines EFA Modells)

Gegeben sei ein EFA Modell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ mit } \varepsilon \sim (0_m, \Psi) \tag{23}$$

und $Q \in \mathbb{R}^{k \times k}$ sei eine orthogonale Matrix. Dann nennen wir

$$\tilde{\upsilon} = \tilde{L}\tilde{\xi} + \varepsilon \text{ mit } \tilde{L} := LQ \text{ und } \tilde{\xi} := Q^T\xi \tag{24}$$

eine *orthogonale Transformation des EFA Modells*

Bemerkung

- Orthogonale Transformationen von EFA Modellen sind die Grundlage der “Faktorenrotation”.

Theorem (Nichtidentifizierbarkeit und Kovarianzinvarianz der EFA)

Gegeben sei ein EFA Modell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ mit } \varepsilon \sim (0_m, \Psi) \tag{25}$$

sowie eine seiner orthogonale Transformationen

$$\tilde{\upsilon} = \tilde{L}\tilde{\xi} + \varepsilon \text{ mit } \tilde{L} := LQ \text{ und } \tilde{\xi} := Q^T\xi \text{ und } Q^TQ = QQ^T = I_k. \tag{26}$$

Dann gelten

$$\upsilon = \tilde{\upsilon} \text{ und } \mathbb{C}(\tilde{\upsilon}) = \mathbb{C}(\upsilon) \tag{27}$$

Beweis

Es gilt zum einen

$$\tilde{\upsilon} = \tilde{L}\tilde{\xi} + \varepsilon = LQQ^T\xi + \varepsilon = LI_k\xi + \varepsilon = L\xi + \varepsilon = \upsilon \tag{28}$$

Zum andereren gilt

$$\mathbb{C}(\tilde{\upsilon}) = LQ(LQ)^T + \Psi = LQQ^TL^T + \Psi = LI_kL^T + \Psi = LL^T + \Psi = \mathbb{C}(\upsilon). \tag{29}$$

Bemerkungen

- Mit

$$\upsilon = L\xi + \varepsilon = \tilde{L}\tilde{\xi} + \varepsilon \tag{30}$$

folgt unmittelbar, dass für festes υ Faktorladungsmatrix L und Faktoren ξ nicht eindeutig bestimmt sind. Diese Tatsache ist die *Nichtidentifizierbarkeit* des EFA Modells: verschiedene Faktorladungsmatrizen und Faktorwerte können die gleichen Daten erklären, aus einer gegebenen Stichprobenkovarianzmatrix kann also nicht eindeutig auf L und ξ geschlossen werden.

- Mit

$$\mathbb{C}(\upsilon) = LL^T + \Psi = \tilde{L}\tilde{L}^T + \Psi \tag{31}$$

folgt weiterhin, dass sich die Gesamtvarianz und die Kommunalitäten bei orthogonaler Transformation nicht ändern. Dies ist die *Kovarianzinvarianz* der EFA bei orthogonaler Transformation.

- In der (normalen) wissenschaftlichen Datenanalyse sind nicht-identifizierbare Modelle eigentlich nicht erwünscht, da geschätzten Parameternwerten keine Bedeutung zugewiesen werden kann. Im Rahmen der EFA wird die Nichtidentifizierbarkeit und damit freie Parameterwahl jedoch als Inspirationsquelle gesehen und hat unter dem Begriff der "Faktorenrotation" in die psychometrische Literatur Eingang gefunden.

Modellformulierung

Modellschätzung

Modellvergleich

Modellinterpretation

Selbstkontrollfragen

Modellschätzung

Hauptkomponentenschätzung

- Schätzung durch Faktorisierung der Stichprobenkovarianzmatrix ohne Berücksichtigung von Ψ

Hauptachsenschätzung

- Schätzung durch Faktorisierung der Stichprobenkovarianzmatrix mit Berücksichtigung von Ψ

Maximum-Likelihood-Schätzung

- Schätzung unter Normalverteilungsannahmen $\Rightarrow$ Konfirmatorische Faktorenanalyse

Motivation der EFA Hauptkomponentenschätzung

- Motivation der EFA Hauptkomponentenschätzung ist die Approximation der Stichprobenkovarianzmatrix als

$$C \approx \hat{L}\hat{L}^T + \hat{\Psi}. \tag{32}$$

- Die EFA Hauptkomponentenschätzung vernachlässigt dabei zunächst $\hat{\Psi}$ und nutzt die Orthonormalzerlegung

$$C = Q\Lambda Q^T = Q\Lambda^{1/2}\Lambda^{1/2}Q^T = \left(Q\Lambda^{1/2}\right)\left(Q\Lambda^{1/2}\right)^T. \tag{33}$$

- Die EFA Hauptkomponentenschätzung vernachlässigt dann die $k+1, \ldots, m$ Spalten von Q und Λ und setzt

$$\hat{L}\hat{L}^T = Q_k\Lambda_k^{1/2}\left(Q_k\Lambda_k^{1/2}\right)^T \text{ mit } \hat{L} \in \mathbb{R}^{m\times k}. \tag{34}$$

- Für die Diagonalelemente c_{ii}, $\hat{h}_i^2$ und $\hat{\psi}_i$ von $C, \hat{L}\hat{L}^T$ und $\hat{\Psi}$, respektive, folgt schließlich, dass

$$c_{ii} = \sum_{j=1}^{k} \hat{l}_{ij}^2 + \hat{\psi}_i \Leftrightarrow \hat{\psi}_i = c_{ii} - \sum_{j=1}^{k} \hat{l}_{ij}^2. \tag{35}$$

Definition (Hauptkomponentenschätzer $k$ter Ordnung von L und Ψ)

Gegeben sei ein Datensatz $Y \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen eines EFA Modells

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{36}$$

$C \in \mathbb{R}^{m \times m}$ sei die Stichprobenkovarianzmatrix von Y und

$$C = Q\Lambda Q^T \tag{37}$$

sei ihre Orthonormalzerlegung mit spaltenweise der Größe nach sortierten Eigenwerten und zugehörigen Eigenvektoren. Dann sind die *Hauptkomponentenschätzer $k$ter Ordnung von L und Ψ* definiert als

$$\hat{L} := Q_k \Lambda_k^{1/2} \text{ und } \hat{\Psi} := \text{diag}\left(\hat{\psi}_1, \ldots, \hat{\psi}_m\right) \tag{38}$$

wobei Λ_k und Q_k die ersten k Spalten von $\Lambda \in \mathbb{R}^{m \times m}$ und $Q \in \mathbb{R}^{m \times m}$ und für $i = 1, \ldots, m$

$$\hat{\psi}_i := c_{ii} - \sum_{j=1}^{k} \hat{l}_{ij}^2 \tag{39}$$

mit den Diagonaleinträgen c_{ii} von C sind.

Bemerkungen

- Alternativ kann die analoge Schätzung aufgrund der Stichprobenkorrelationsmatrix vorgenommen werden.
- Die Selektion der ersten k Spalten von C und Λ impliziert ein Faktorenanalysemodell mit k Faktoren.

Definition (Varianz-, Kommunalitäts- und Spezifitätsschätzer)

Für einen Datensatz $Y \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen eines EFA Modells

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{40}$$

und ein $k < m$ seien $\hat{L} = (\hat{l}_{ij})_{1 \leq i \leq m, 1 \leq j \leq k} \in \mathbb{R}^{m \times k}$ und $\hat{\Psi} = \text{diag}(\hat{\psi}_1, ..., \hat{\psi}_m) \in \mathbb{R}^{m \times m}$ die durch die Approximation

$$C \approx \hat{L}\hat{L}^T + \Psi \tag{41}$$

der Stichprobenkovarianzmatrix $C = (c_{ij})_{1 \leq i,j \leq m} \in \mathbb{R}^{m \times m}$ von Y gewonnenen Hauptkomponentenschätzer $k$ter Ordnung. Dann ergeben sich sich neben den Stichprobenkovarianzmatrix-impliziten Schätzern

- c_{ii} als Schätzer der Varianz von υ_i und
- $G := \sum_{i=1}^{m} c_{ii}$ als Schätzer der Gesamtvarianz von υ

weiterhin

- $\hat{h}_i^2 := \sum_{j=1}^{k} \hat{l}_{ij}^2$ als Schätzer der Kommunalität h_i^2 von υ_i und
- $\hat{\psi}_i$ als Schätzer der Spezifität ψ_i von υ_i.

Bemerkungen

- Über die Güte der Schätzer machen wir hier keine Aussagen.

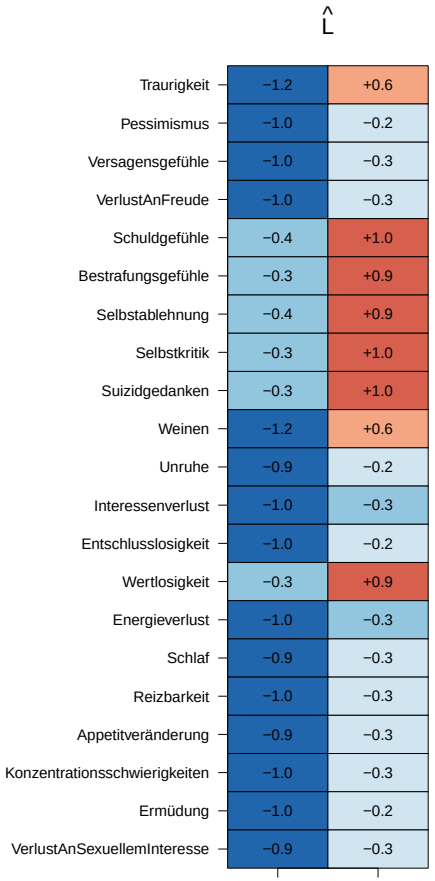
Anwendungsbeispiel

Hauptkomponentenschätzer der Faktorenanalysemodellparameter bei $k = 2$

```
# EFA mit Hauptkomponentenschätzung für k = 2
Y         = as.matrix(t(read.csv("./7_Daten/7_efa.csv") ))          # Y   \in \mathbb{R}^{m x n}
m         = nrow(Y)                                                 # Datendimension
n         = ncol(Y)                                                 # Datenpunktanzahl
k         = 2                                                       # Faktoranzahl
I_n       = diag(n)                                                 # Einheitsmatrix I_n
J_n       = matrix(rep(1,n^2), nrow = n)                            # 1_{nn}
C         = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))              # Stichprobenkovarianzmatrix
EA        = eigen(C)                                                # Eigenanalyse von R
lambda_k  = EA$values[1:k]                                          # k größte Eigenwerte von R
Q_k       = EA$vectors[,1:k]                                        # k zugehörige Eigenvektoren von R
L_hat     = Q_k %*% diag(sqrt(lambda_k))                            # Faktorladungsmatrixschätzer
Psi_hat   = diag(diag(C) - diag(L_hat %*% t(L_hat)))                # Beobachtungsrauschenkovarianzmatrixschätzer
V_i_hat   = diag(C)                                                 # Varianzschätzer
h2_i_hat  = rowSums(L_hat^2)                                        # Kommunalitätsschätzer
psi_i_hat = diag(Psi_hat)                                           # Spezifitätsschätzer
```

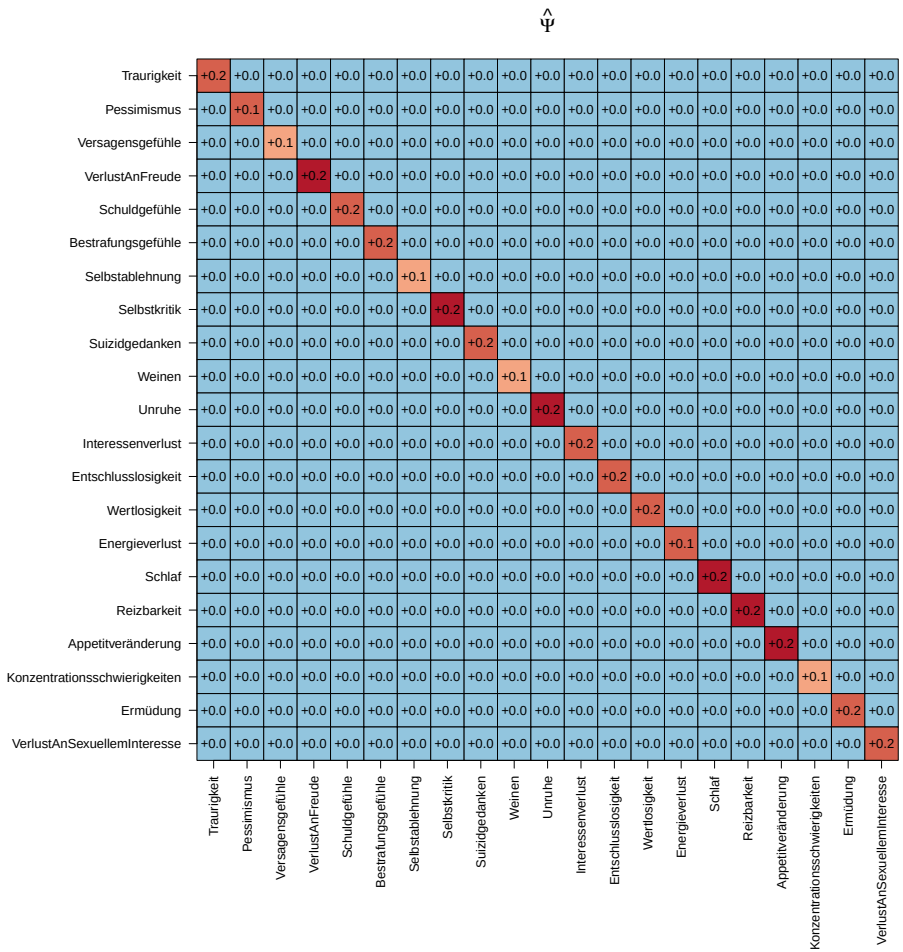
Anwendungsbeispiel

Hauptkomponentenschätzer der Faktorenanalysemodellparameter bei $k = 2$



Anwendungsbeispiel

Hauptkomponentenschätzer der Faktorenanalysemodellparameter bei $k = 2$



Modellformulierung

Modellschätzung

Modellvergleich

Modellinterpretation

Selbstkontrollfragen

Überblick Modellevaluation | Modellvergleich

Modellvergleich = Wahl der Anzahl k von Faktoren

Grundlegendes Ziel ist die Erklärung von möglichst viel Datenvarianz mit möglichst wenigen Faktoren.

Quantitative Grundlage dafür ist die Zerlegung der *Gesamtstichprobenvarianz* G anhand von

$$G = F + R \tag{42}$$

in eine *Faktorenbasierte Stichprobenvarianz* F und eine *Beobachtungsrauschenbasierte Stichprobenvarianz* R.

- Man wählt die Anzahl k der Faktoren so, dass k möglichst klein, aber F/R möglichst groß ist.
- Traditionell gibt es zu diesem Zweck eine Reihe von Heuristiken.

Wir zeigen zunächst die Validität obiger Varianzzerlegung und diskutieren dann Möglichkeiten zur Wahl von k.

Definition (EFA Stichprobenvarianzzerlegung)

$Y \in \mathbb{R}^{m \times n}$ sei ein Datensatz von n unabhängigen Beobachtungen eines EFA Modells

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{43}$$

$C \in \mathbb{R}^{m \times m}$ sei die Stichprobenkovarianzmatrix von Y und $\hat{L} \in \mathbb{R}^{m \times k}$ und $\hat{\Psi} \in \mathbb{R}^{m \times m}$ seien die durch Orthonormalzerlegung von C und Betrachtung der $k < m$ größten Eigenwerte $\lambda_1, \ldots, \lambda_k$ und zugehörigen Eigenvektoren gewonnenen Hauptkomponentenschätzer $k$ter Ordnung, so dass

$$C \approx \hat{L}\hat{L}^T + \hat{\Psi}. \tag{44}$$

Dann wird

- die Summe der Diagonalemente von C als *Gesamtstichprobenvarianz*,
- die Summe der Diagonalelemente von $\hat{L}\hat{L}^T$ als *Faktorbasierte Stichprobenvarianz* und
- die Summe der Diagonalelemente von $\hat{\Psi}$ als *Beobachtungsrauschenbasierte Stichprobenvarianz*.

bezeichnet

Theorem (EFA Stichprobenvarianzzerlegung)

Für einen Datensatz $Y \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen eines EFA Modells

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{45}$$

seien

- $C = (c_{ij})_{1 \leq i,j \leq m} \in \mathbb{R}^{m \times m}$ die Stichprobenkovarianzmatrix,
- $\hat{L} = (\hat{l}_{ij})_{1 \leq i \leq m, 1 \leq j \leq k} \in \mathbb{R}^{m \times k}$ der Hauptkomponentenschätzer $k$ter Ordnung von L,
- $\hat{\Psi} = \text{diag}(\hat{\psi}_i, ..., \hat{\psi}_m) \in \mathbb{R}^{m \times m}$ der Hauptkomponentenschätzer $k$ter Ordnung von Ψ,

sowie

- $G := \sum_{i=1}^{m} c_{ii}$ die Gesamtstichprobenvarianz,
- $F := \sum_{i=1}^{m} \sum_{j=1}^{k} \hat{l}_{ij}^2$ die Faktorbasierte Stichprobenvarianz,
- $R := \sum_{i=1}^{m} \hat{\psi}_i$ die Beobachtungsrauschenbasierte Stichprobenvarianz.

Dann gilt

$$G = F + R. \tag{46}$$

Außerdem gilt mit den Eigenwerten $\lambda_1, ..., \lambda_k$ von C, dass

$$F = \sum_{j=1}^{k} \lambda_j, \text{ wobei } \lambda_j = \sum_{i=1}^{m} \hat{l}_{ij}^2 \tag{47}$$

für $j = 1, ..., k$ der Anteil des $j$ten Faktors an F ist.

Beweis

Wir erinnern zunächst daran, dass die Diagonalelemente von $\hat{L}\hat{L}^T$ durch

$$\sum_{j=1}^{k} \hat{l}_{ij}^2 \tag{48}$$

gegeben sind, wovon man sich durch Betrachtung der Einträge von $\hat{L}\hat{L}^T$ überzeugt:

$$\begin{aligned} \hat{L}\hat{L}^T &= \begin{pmatrix} \hat{l}_{11} & \cdots & \hat{l}_{1k} \\ \hat{l}_{21} & \cdots & \hat{l}_{2k} \\ \vdots & \ddots & \vdots \\ \hat{l}_{m1} & \cdots & \hat{l}_{mk} \end{pmatrix} \begin{pmatrix} \hat{l}_{11} & \cdots & \hat{l}_{m1} \\ \hat{l}_{12} & \cdots & \hat{l}_{m2} \\ \vdots & \ddots & \vdots \\ \hat{l}_{1k} & \cdots & \hat{l}_{mk} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^{k} l_{1j}^2 & \sum_{j=1}^{k} l_{1j}l_{2j} & \cdots & \sum_{j=1}^{k} l_{1j}l_{mj} \\ \sum_{j=1}^{k} l_{2j}l_{1j} & \sum_{j=1}^{k} l_{2j}^2 & \cdots & \sum_{j=1}^{k} l_{2j}l_{mj} \\ \vdots & \cdots & \ddots & \vdots \\ \sum_{j=1}^{k} l_{mj}l_{1j} & \sum_{j=1}^{k} l_{mj}l_{2j} & \cdots & \sum_{j=1}^{k} l_{mj}^2 \end{pmatrix} \end{aligned} \tag{49}$$

Beweis (fortgeführt)

Die Identität von G und $F + R$ folgt dann direkt aus der Identität der Diagonalelemente von C, $\hat{L}\hat{L}^T$ und $\hat{\Psi}$, die im Rahmen der Hauptkomponentenschätzung mithilfe von

$$\hat{\psi}_i := c_{ii} - \sum_{j=1}^{k} \hat{l}_{ij}^2 \text{ für } i = 1, ..., m \tag{50}$$

konstruiert wird. Um als nächstes

$$F = \sum_{j=1}^{k} \lambda_j \tag{51}$$

zu zeigen halten zunächst fest, dass mit der Definition des Hauptkomponentenschätzer $\hat{L}$ die Summe der quadrierten Einträge in der $j$ten Spalte von $\hat{L}$ gleich der Summe der quadrierten Einträge in der $j$ten Spalte von $Q_k \Lambda_k^{1/2}$ ist. Dies mag man sich zum Beispiel für $m = 5$ und $k = 2$ verdeutlichen:

$$\hat{L} = Q_k \Lambda_k^{1/2} \Leftrightarrow \begin{pmatrix} \hat{l}_{11} & \hat{l}_{12} \\ \hat{l}_{21} & \hat{l}_{22} \\ \hat{l}_{31} & \hat{l}_{32} \\ \hat{l}_{41} & \hat{l}_{42} \\ \hat{l}_{51} & \hat{l}_{52} \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \\ q_{31} & q_{32} \\ q_{41} & q_{42} \\ q_{51} & q_{52} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} q_{11} & \sqrt{\lambda_2} q_{12} \\ \sqrt{\lambda_1} q_{21} & \sqrt{\lambda_2} q_{22} \\ \sqrt{\lambda_1} q_{31} & \sqrt{\lambda_2} q_{32} \\ \sqrt{\lambda_1} q_{41} & \sqrt{\lambda_2} q_{42} \\ \sqrt{\lambda_1} q_{51} & \sqrt{\lambda_2} q_{52} \end{pmatrix}$$

Beweis (fortgeführt)

Weiterhin halten wir fest, dass, wenn q_j für $j = 1, .., k$ die $j$te Spalte von Q_k bezeichnet aufgrund der Orthonormalität von Q folgt, dass

$$q_j^T q_j = \sum_{i=1}^{m} q_{ij}^2 = 1. \tag{52}$$

Dann ergibt sich für die Summe der Diagonalelemente von $\hat{L}\hat{L}^T$ aber

$$F = \sum_{i=1}^{m}\sum_{j=1}^{k} \hat{l}_{ij}^2 = \sum_{j=1}^{k}\sum_{i=1}^{m} \hat{l}_{ij}^2 = \sum_{j=1}^{k}\sum_{i=1}^{m} \left(\sqrt{\lambda}_j q_{ij}\right)^2 = \sum_{j=1}^{k} \lambda_j \sum_{i=1}^{m} q_{ij}^2 = \sum_{j=1}^{k} \lambda_j \tag{53}$$

Die Tatsache, dass der $j$te Eigenwert λ_j von C dabei der Anteil der durch den $j$ten Faktor erklärten Gesamtstichprobenvarianz ist ergibt sich dabei durch die Einsicht, dass der Beitrag des $j$ten Faktors in der $j$ten Spalte von $\hat{L}$ enkodiert ist und obige Gleichungskette impliziert, dass

$$\sum_{i=1}^{m} \hat{l}_{ij}^2 = \lambda_j \text{ für } j = 1, \ldots, k. \tag{54}$$

Anwendungsbeispiel

```
# EFA mit Hauptkomponentenschätzung für k = 2
YT        = read.csv("./7_Daten/7_efa.csv")                    # Y^T \in \mathbb{R}^{n x m}
Y         = as.matrix(t(YT))                                   # Y   \in \mathbb{R}^{m x n}
m         = nrow(Y)                                            # Datendimension
n         = ncol(Y)                                            # Datenpunktanzahl
k         = 2                                                  # Faktoranzahl
I_n       = diag(n)                                            # Einheitsmatrix I_n
J_n       = matrix(rep(1,n^2), nrow = n)                       # 1_{nn}
C         = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))         # Stichprobenkovarianzmatrix
EA        = eigen(C)                                           # Eigenanalyse von R
lambda_k  = EA$values[1:k]                                     # k größte Eigenwerte von R
Q_k       = EA$vectors[,1:k]                                   # k zugehörige Eigenvektoren von R
L_hat     = Q_k %*% diag(sqrt(lambda_k))                       # Faktorladungsmatrixschätzer
Psi_hat   = diag(diag(C) - diag(L_hat %*% t(L_hat)))           # Beobachtungsrauschenkovarianzmatrixschätzer
GG        = sum(diag(C))                                       # Gesamtstichprobenvarianz
FF        = sum(diag(L_hat %*% t(L_hat)))                      # Faktorenbasierte Stichprobenvarianz
RR        = sum(diag(Psi_hat))                                 # Beobachtungsrauschenbasierter Stichprobenvarianz
FF_lambda = sum(lambda_k)                                      # Summe der Eigenwerte \lambda_1,...,\lambda_k
```

Anwendungsbeispiel

Darstellung der Gesamtstichprobenvarianz

```
G   = 25.84432
F   = 22.50997
R   = 3.334346
F+B = 25.84432
```

Darstellung der Faktorenbasierten Stichprobenvarianz

```
Faktorbasierte Stichprobenvarianz F        = 22.50997
Summe der Eigenwerte lambda_1,...,lambda_k = 22.50997
```

Wahl der Anzahl k von Faktoren

Man wählt k so, dass ein vorgegebener Anteil der Gesamtstichprobenvarianz durch das Modell erklärt wird.

Dazu mag es Sinn machen, sich obige folgende Einsichten noch einmal zu vergegenwärtigen:

- Der durch den $j$ten Faktor erklärte Anteil an G ist λ_j.
- Der durch den $j$ten Faktor erklärte relative Anteil an G ist λ_j/G.
- Der durch die $j = 1, \ldots, k$ Faktoren erklärte relative Anteil an G ist $\sum_{j=1}^{k} \lambda_j/G$.

Es macht also Sinn, sich $\lambda_j, \lambda_j/G$ und $\sum_{j=1}^{k} \lambda_j/G$ zu visualisieren.

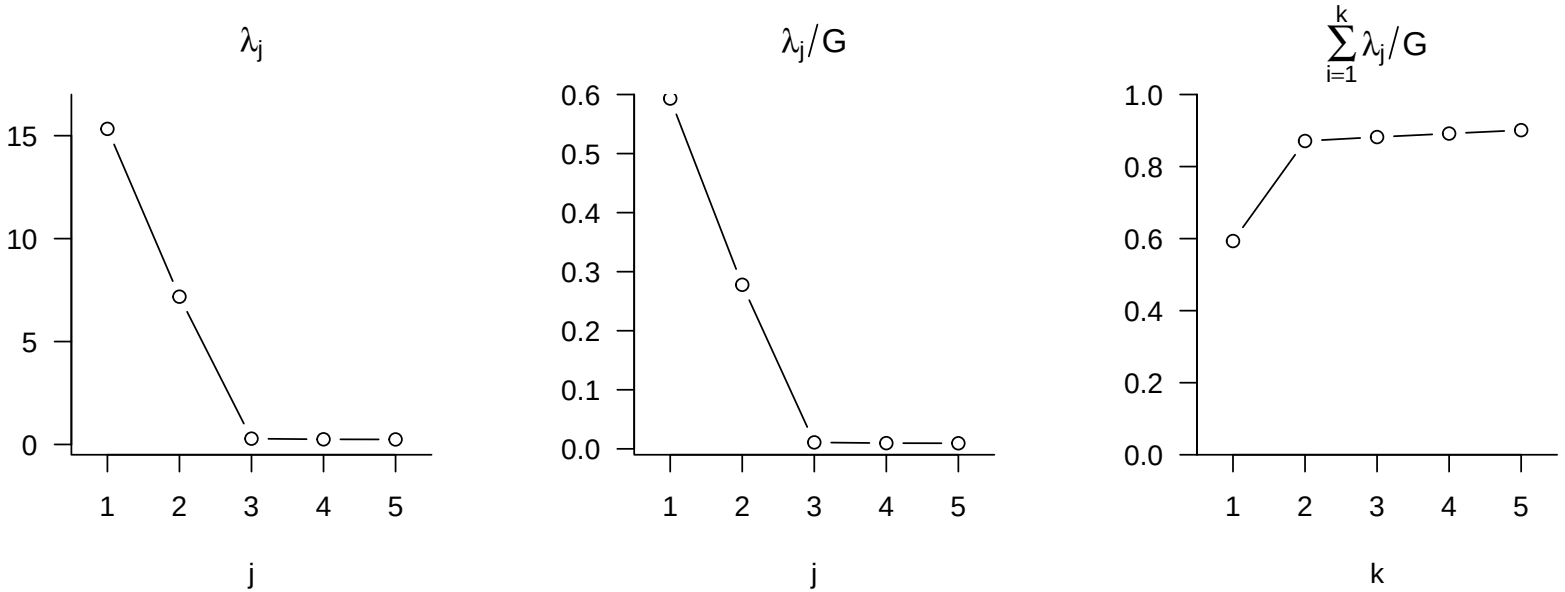
Basierend darauf kann man dann k so wählen, dass k möglichst klein und $\sum_{j=1}^{k} \lambda_j/G$ möglichst groß ist.

Die Visualisierung der λ_j wird in diesem Kontext *Scree-Plot* genannt (*Scree, engl. Geröllhalde, Schutthalde, Talus*)

Anwendungsbeispiel

```
# EFA mit Hauptkomponentenschätzung für k = 5
YT        = read.csv("./7_Daten/7_efa.csv")                    # Y^T \in \mathbb{R}^{n x m}
Y         = as.matrix(t(YT))                                   # Y   \in \mathbb{R}^{m x n}
m         = nrow(Y)                                            # Datendimension
n         = ncol(Y)                                            # Datenpunktanzahl
k         = 5                                                  # Faktoranzahl
I_n       = diag(n)                                            # Einheitsmatrix I_n
J_n       = matrix(rep(1,n^2), nrow = n)                       # 1_{nn}
C         = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))         # Stichprobenkovarianzmatrix
EA        = eigen(C)                                           # Eigenanalyse von R
lambda_k  = EA$values[1:k]                                     # k größte Eigenwerte von R
Q_k       = EA$vectors[,1:k]                                   # k zugehörige Eigenvektoren von R
L_hat     = Q_k %*% diag(sqrt(lambda_k))                       # Faktorladungsmatrixschätzer
Psi_hat   = diag(diag(C) - diag(L_hat %*% t(L_hat)))           # Beobachtungsrauschenkovarianzmatrixschätzer
G         = sum(diag(C))                                       # Gesamtstichprobenvarianz
```

Anwendungsbeispiel



- Mit $k = 1$ können 56% der Gesamtstichprobenvarianz erklärt werden.
- Mit $k = 2$ können 87% der Gesamtstichprobenvarianz erklärt werden.
- Mit $k = 3$ können 88% der Gesamtstichprobenvarianz erklärt werden.

$$\Rightarrow k = 2 \text{ scheint eine sinnvolle Wahl}$$

Modellformulierung

Modellschätzung

Modellvergleich

Modellinterpretation

Selbstkontrollfragen

Überblick Modellevaluation | Modellinterpretation = Rotationsverfahren

Per Datenkomponente sind Faktorladungen gewünscht, die möglichst leicht eine eindeutige Faktorzuordnung erlauben. Statt bpsw. einer geschätzten Faktorladungsmatrix $\hat{L}$ wäre also eine geschätzte Faktorladungsmatrix wie $\hat{L}^*$ gewünscht.

$$\hat{L} = \begin{pmatrix} -3.81 & 0.16 \\ -2.54 & 1.35 \\ -3.89 & 0.11 \\ -0.53 & -1.22 \\ -1.66 & -2.32 \end{pmatrix} \quad \Rightarrow \quad \hat{L}^* = \begin{pmatrix} 1.00 & 0.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 0.00 & 1.00 \end{pmatrix} \tag{55}$$

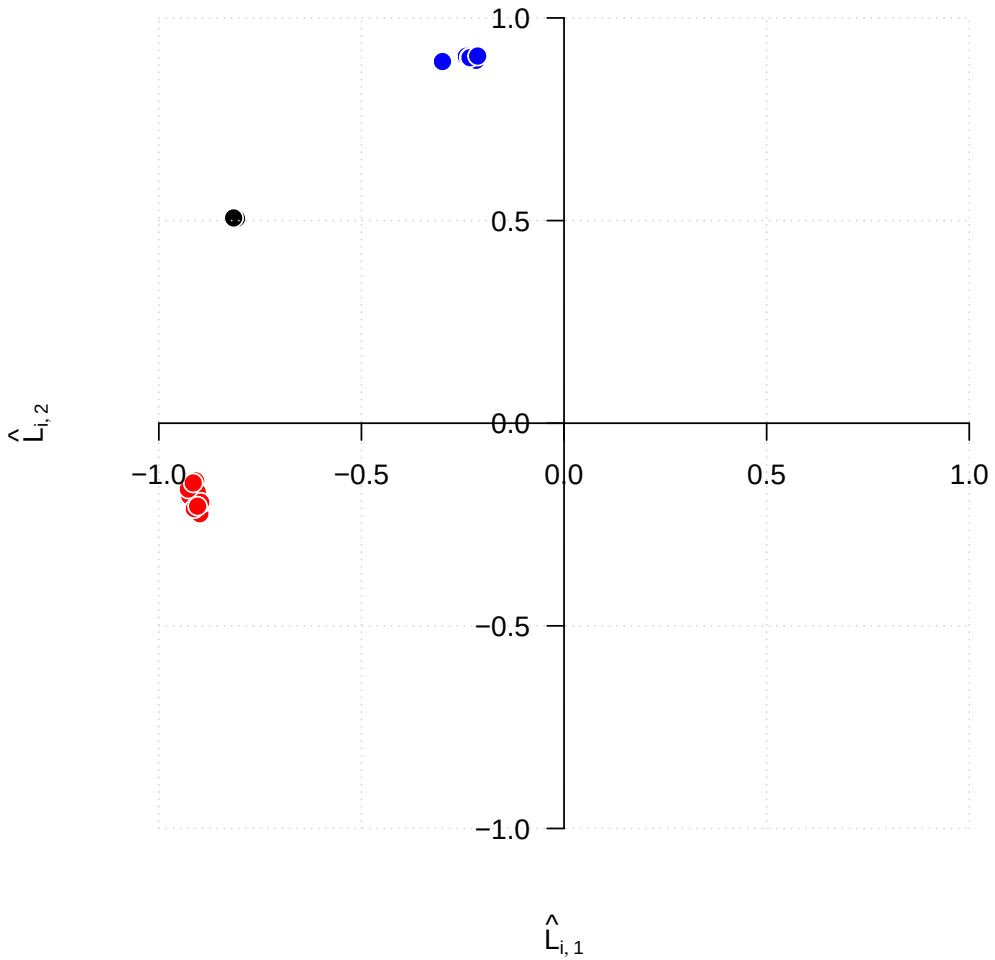
Eindeutige Zuordnungen von Datenkomponenten zu Faktoren induzieren dann Cluster von Datenkomponenten, "die auf jeweils einen Faktor laden". Eine entsprechend modifizierte Faktorladungsmatrix $\hat{L}^*$ nennt man auch "Einfachstruktur" und man hofft dann, durch Inspektion der Cluster zu eine inhaltlichen Interpretation der Faktoren inspiriert zu werden. Für eine Standardisierung der Einträge von $\hat{L}$ gehen wir dabei zunächst zu Hauptkomponentenschätzung auf Grundlage der Stichprobenkorrelationsmatrix über.

Da weiterhin die Faktorladungen perse sowieso nur bis auf die Multiplikation mit einer orthogonalen Matrix bestimmt sind, kann man die Multiplikation der so geschätzten Faktorladungsmatrix mit verschiedenen orthogonalen Matrizen ausprobieren ohne die erklärte Gesamtstichprobenvarianz des geschätzten Modells zu verändern.

Geometrisch entspricht die Multiplikation mit einer orthogonalen Matrix einer Vektorkoordinatentransformation, also der Wahl einer alternativen Orthogonalbasis zur Bestimmung der Faktorladungskoordinaten. Wir beschränken uns in der Diskussion auf Faktorrotationen bei $k := 2$ und die sogenannte "Varimaxrotation".

Anwendungsbeispiel

Visualisierung der geschätzten Faktorladungen jeder Datenvektorkomponenten



Anwendungsbeispiel

Visualisierung der geschätzten Faktorladungen jeder Datenvektorkomponenten

Zeilen von $\hat{L}_2$ mit •

```
 [1] "Pessimismus"                   "Versagensgefühle"
 [3] "VerlustAnFreude"               "Unruhe"
 [5] "Interessenverlust"             "Entschlusslosigkeit"
 [7] "Energieverlust"                "Schlaf"
 [9] "Reizbarkeit"                   "Appetitveränderung"
[11] "Konzentrationsschwierigkeiten" "Ermüdung"
[13] "VerlustAnSexuellemInteresse"
```

Zeilen von $\hat{L}_2$ mit •

```
[1] "Schuldgefühle"      "Bestrafungsgefühle" "Selbstablehnung"
[4] "Selbstkritik"       "Suizidgedanken"     "Wertlosigkeit"
```

Zeilen von $\hat{L}_2$ mit •

```
[1] "Traurigkeit" "Weinen"
```

Theorem (Drehmatrizen in $\mathbb{R}^{2\times 2}$)

Es sei

$$M_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \in \mathbb{R}^{2\times 2} \text{ für } 0 \leq \theta \leq 2\pi \tag{56}$$

eine sogenannte *Drehmatrix*. Dann gelten

(1) M_θ ist eine orthogonale Matrix

(2) Die Spalten von M_θ bilden eine Orthonormalbasis von $\mathbb{R}^2$

(3) Multiplikation mit M_θ^T transfomiert die Koordinaten eines Vektors $v \in \mathbb{R}^2$ hinsichtlich der kanonischen Orthonormalbasis $B_v := \{e_1, e_2\}$ in Koordinaten desselben Vektors hinsichtlich der Basis

$$B_w := \left\{ \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} \right\} \tag{57}$$

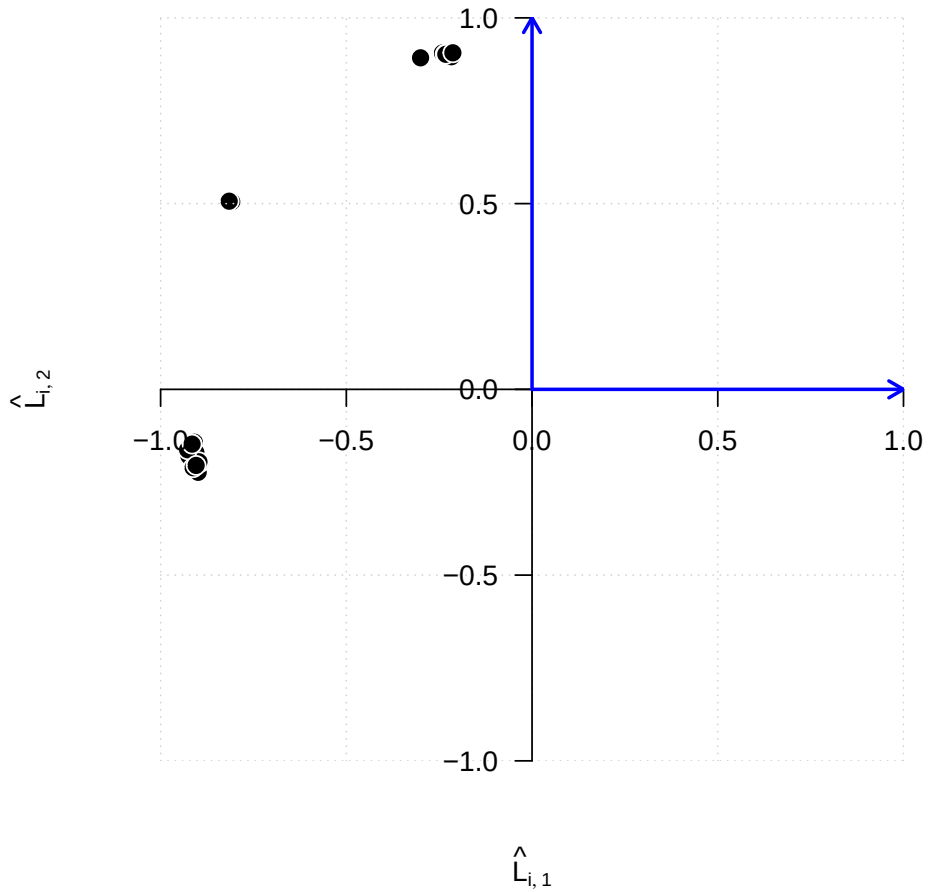
Bemerkungen

- Wir verzichten auf einen Beweis.
- Das Theorem stellt eine unendliche, durch θ parameterisierte Menge von Orthonormalbasen von $\mathbb{R}^2$ bereit und ermöglicht weiterhin, die Faktorladungskoordinaten jeder Datenkomponente bezüglich jeder dieser Orthonormalbasen zu evaluieren. Wir bezeichnen die die auf diese Weise für $\theta \in [0, 2\pi]$ gewonnene geschätzte Faktorladungskoordinatenmatrix mit

$$\hat{L}_\theta := \left(M_\theta^T \hat{L}^T\right)^T \tag{58}$$

Anwendungsbeispiel

Drehmatrixbasisvektoren für $\theta := 0$ und Faktorladungskoordinatenmatrix



Anwendungsbeispiel

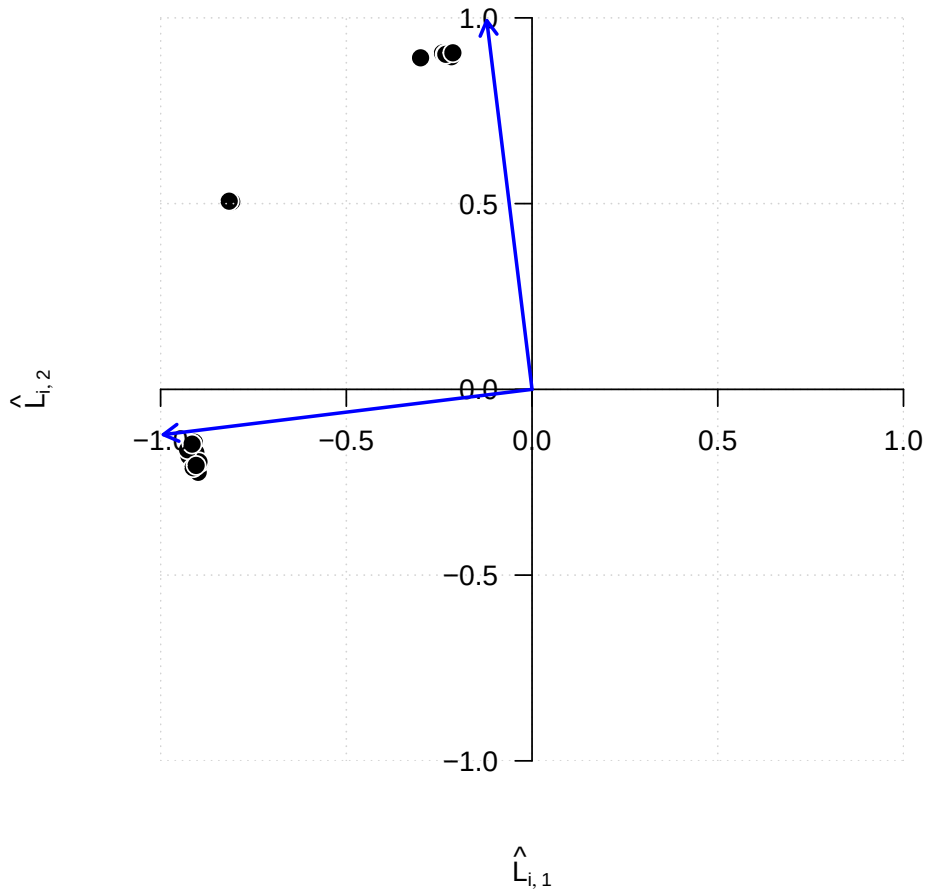
$\hat{L}_\theta := \left(M_\theta^T \hat{L}^T\right)$ für $\theta := 0$

```
theta       = 0
M           = matrix(c(cos(theta), -sin(theta),
                       sin(theta),  cos(theta)),
                       nrow = 2, byrow = TRUE)
L_hat_theta = t((M) %*% t(L_hat))
```

```
             [,1]       [,2]
 [1,] -0.8085081  0.5051137
 [2,] -0.9208545 -0.1539652
 [3,] -0.9236270 -0.1798691
 [4,] -0.9015289 -0.1944968
 [5,] -0.2412607  0.9054020
 [6,] -0.2339233  0.9022031
 [7,] -0.3000994  0.8924383
 [8,] -0.2171111  0.8953713
 [9,] -0.2321825  0.9019178
[10,] -0.8154280  0.5064901
[11,] -0.9090954 -0.1424025
[12,] -0.8988565 -0.2233184
[13,] -0.9148247 -0.1474827
[14,] -0.2132255  0.9059484
[15,] -0.9125346 -0.2108940
[16,] -0.8934833 -0.1965261
[17,] -0.9052280 -0.1697388
[18,] -0.8968051 -0.1949494
[19,] -0.9273076 -0.1624947
[20,] -0.9152565 -0.1474276
[21,] -0.9044185 -0.2047372
```

Anwendungsbeispiel

Drehmatrixbasisvektoren für $:= 93 und Faktorladungskoordinatenmatrix



Anwendungsbeispiel

$\hat{L}_\theta := \left(M_\theta^T \hat{L}^T\right)$ für $\theta := 103$

```
theta       = 97
M           = matrix(c(cos(theta), -sin(theta),
                       sin(theta),  cos(theta)),
                       nrow = 2, byrow = TRUE)
L_hat_theta = t((M) %*% t(L_hat))
```

```
       [,1]  [,2]
 [1,]  0.56 -0.77
 [2,]  0.91 -0.21
 [3,]  0.92 -0.18
 [4,]  0.91 -0.16
 [5,] -0.12 -0.93
 [6,] -0.13 -0.92
 [7,] -0.06 -0.94
 [8,] -0.14 -0.91
 [9,] -0.13 -0.92
[10,]  0.56 -0.78
[11,]  0.90 -0.21
[12,]  0.92 -0.13
[13,]  0.90 -0.21
[14,] -0.15 -0.92
[15,]  0.92 -0.15
[16,]  0.90 -0.16
[17,]  0.90 -0.19
[18,]  0.90 -0.16
[19,]  0.92 -0.20
[20,]  0.90 -0.21
[21,]  0.91 -0.15
```

Definition (Varimaxfaktorladungsmatrix)

Für $\hat{L} := (\hat{l}_{ij})_{1 \leq i \leq m, 1 \leq j \leq 2} \in \mathbb{R}^{m \times 2}$ sei die *Varimaxfunktion* definiert als

$$f : \mathbb{R}^{m \times 2} \to \mathbb{R}_{\geq 0}, \hat{L} \mapsto f(\hat{L}) := \sum_{j=1}^{2} \sum_{i=1}^{m} \left(\hat{l}_{ij}^2 - \bar{l}_j^2 \right)^2 \text{ mit } \bar{l}_j^2 := \frac{1}{m} \sum_{i=1}^{m} \hat{l}_{ij}^2. \tag{59}$$

Weiterhin sei

$$\hat{L}_\theta := \left(M_\theta^T \hat{L}^T \right)^T \tag{60}$$

die Matrix der Vektorkoordinaten von $\hat{L}$ bezüglich der Orthonormalbasis der Spalten von M_θ. Dann heißt

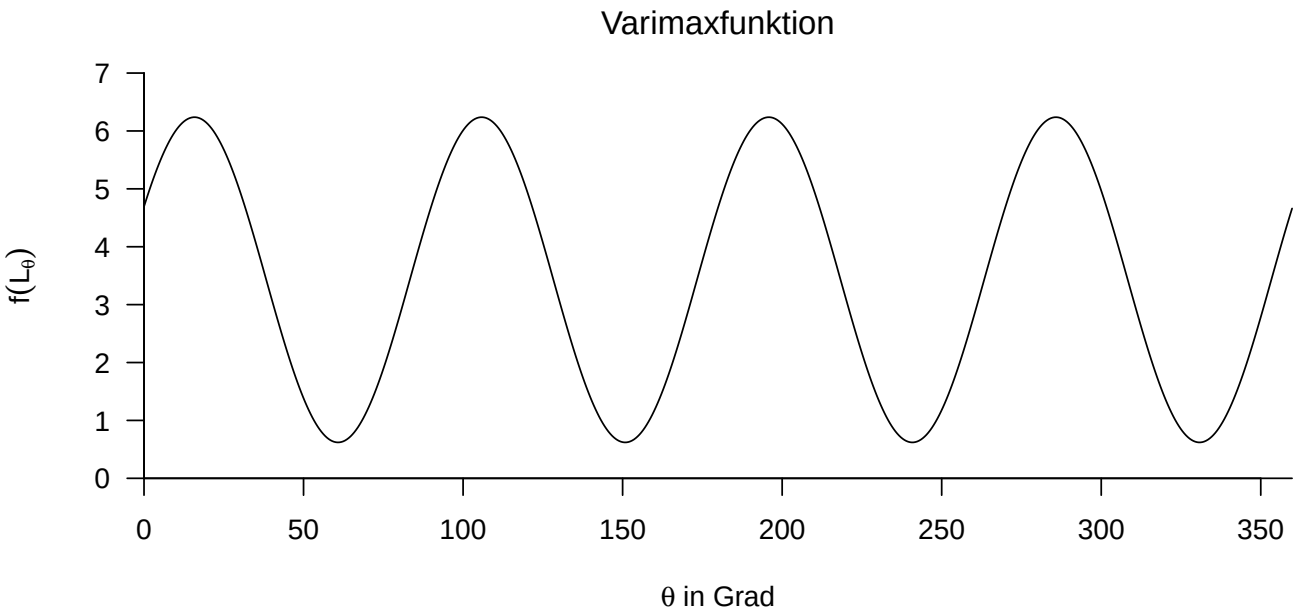
$$\hat{L}_\theta^* := \underset{0 \leq \theta \leq 2\pi}{\arg\max} f(\hat{L}_\theta) \tag{61}$$

die *Varimaxfaktorladungsmatrix*.

Bemerkungen

- Intuitiv ist $f(M)$ die Summe der Stichprobenvarianzen der Spalten von L.
- Wenn die Faktorladungen einer Spalte alle gleich sind, ist der $j$te Beitrag zu $f(L) = 0$.
- Wenn einige Faktorladungen in einer Spalte groß sind und andere klein sind, ist $j$te Beitrag zu $f(M)$ groß.
- Die f favorisiert also Faktorladungsmatrizen mit vielen sehr großen und vielen sehr kleinen Werten.
- $\hat{L}_\theta^*$ optimiert dieses Kriterium unter allen Matrizen die $M_\theta \hat{L}$ gebildet werden können.

Anwendungsbeispiel



Modellformulierung

Modellschätzung

Modellvergleich

Modellinterpretation

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition des Modells der explorativen Faktorenanalyse (EFA) wieder.
2. Erläutern Sie das Modell der EFA.
3. Geben Sie das Theorem zur Datenkovarianzmatrix der EFA wieder.
4. Geben Sie das Theorem zur Varianzzerlegung der EFA wieder.
5. Erläutern Sie das Theorem zur Varianzzerlegung der EFA.
6. Definieren Sie die Kommunalität und die Spezifität einer Datenkomponente im EFA Modell.
7. Definieren Sie den Begriff der Gesamtvarianz im EFA Modell.
8. Warum gilt im EFA Modell "Gesamtvarianz = Summe der Kommunalitäten + Summe der Spezifitäten"?
9. Definieren Sie den Begriff der orthogonalen Transformation eines EFA Modells.
10. Geben Sie das Theorem zur Nichtidentifizierbarkeit und Kovarianzinvarianz des EFA Modell wieder.
11. Erläutern Sie die Nichtidentifizierbarkeit eines EFA Modells.
12. Erläutern Sie die Kovarianzinvarianz eines EFA Modells.
13. Definieren Sie die Hauptkomponentenschätzer $k$ter Ordnung der EFA Modellparameter L und Ψ.
14. Definieren Sie die Varianz-, Kommunalitäts- und Spezifitätsschätzer der EFA.
15. Definieren Sie die Gesamtstichprobenvarianz, die Faktorbasierte Stichprobenvarianz und die Beobachtungsrauschenbasierte Stichprobenvarianz der EFA.
16. Warum gilt für die EFA "Gesamtstichprobenvarianz = Faktorbasierte Stichprobenvarianz + Beobachtungsrauschenbasierte Stichprobenvarianz"?
17. Warum ist der $j$te Eigenwert λ_j der Stichprobenkovarianzmatrix der Anteil der durch den $j$ten Faktor erklärten Gesamtstichprobenvarianz?
18. Erläutern Sie das Ziel von Rotationsverfahren im Kontext der EFA.
19. Geben Sie das Theorem zu Drehmatrizen in $\mathbb{R}^2$ wieder.
20. Definieren Sie die Varimaxfaktorladungsmatrix.

Referenzen I

–>

Bartholomew, David J., M. Knott, and Irini Moustaki. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. 3rd ed. Wiley Series in Probability and Statistics. Chichester, West Sussex: Wiley.

Beck, Aaron T., Robert A. Steer, Roberta Ball, and William F. Ranieri. 1996. "Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients." *Journal of Personality Assessment* 67 (3): 588–97. https://doi.org/10.1207/s15327752jpa6703_13.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. Wiley New York.

Fisher, R. A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (594-604): 309–68. https://doi.org/10.1098/rsta.1922.0009.

Hautzinger, M, F. Keller, and C. Kühner. 2006. *BDI-II Beck Depressions-Inventar*. Pearson.

Hotelling, Harold. 1933. "Analysis of Complex Variables into Principal Components." *Journal of Educational Psychology* 24: 417-441 and 498-520.

Jöreskog, K. G. 1970. "A General Method for Analysis of Covariance Structures." *Biometrika* 57 (2): 239. https://doi.org/10.2307/2334833.

Keller, Ferdinand, Martin Hautzinger, and Christine Kühner. 2008. "Zur faktoriellen Struktur des deutschsprachigen BDI-II." *Zeitschrift für Klinische Psychologie und Psychotherapie* 37 (4): 245–54. https://doi.org/10.1026/1616-3443.37.4.245.

Lawley, Derrick. 1940. "The Estimation of Factor Loadings by the Method of Maximum Likelihood." *Proceedings of the Royal Society of Edinburgh. Section B: Biological Sciences*.

Referenzen II

Lawley, Derrick, and Arthur Maxwell. 1962. "Factor Analysis as a Statistical Method." *The Statistician* 12 (3): 209. https://doi.org/10.2307/2986915.

Mulaik, Stanley A. 2010. *Foundations of Factor Analysis*. CRC Press, Taylor & Francis Group.

Muthén, L. K., and B. O. Muthén. 1998/2017. *Mplus User's Guide. Eighth Edition*.

Pearson, Karl. 1901. "On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72. https://doi.org/10.1080/14786440109462720.

Rosseel, Yves. 2012. "**Lavaan** : An *R* Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2). https://doi.org/10.18637/jss.v048.i02.

Spearman, C. 1904. ""General Intelligence," Objectively Determined and Measured." *The American Journal of Psychology* 15 (2): 201. https://doi.org/10.2307/1412107.

Thurstone, L L. 1947. "Multiple Factor Analysis." *University of Chicago Press*.

Wishart, J. 1928. "The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population." *Biometrika* 20A (1/2): 32. https://doi.org/10.2307/2331939.