Multivariate Datenanalyse

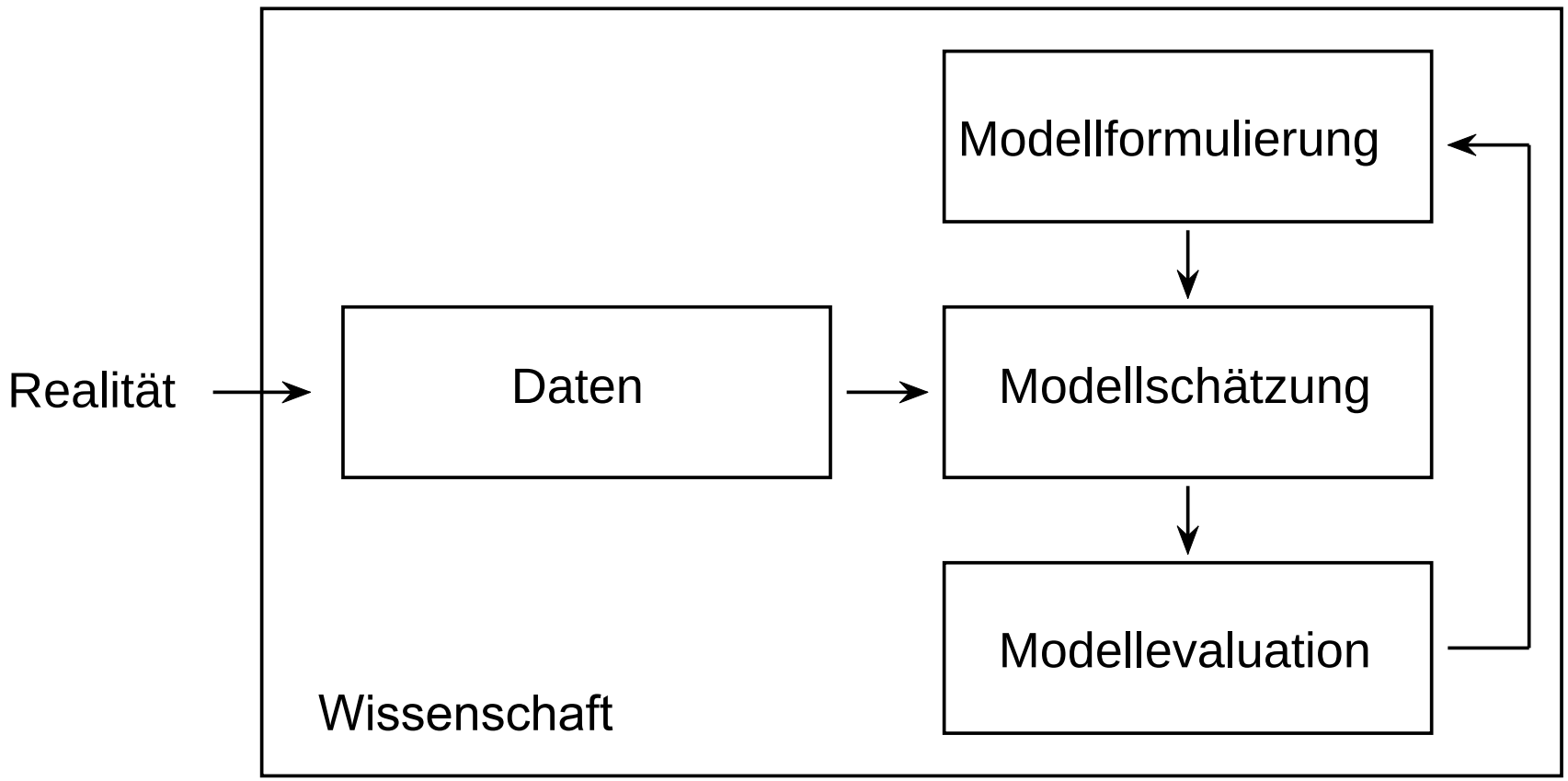
MSc Psychologie WiSe 2022/23

Prof. Dr. Dirk Ostwald

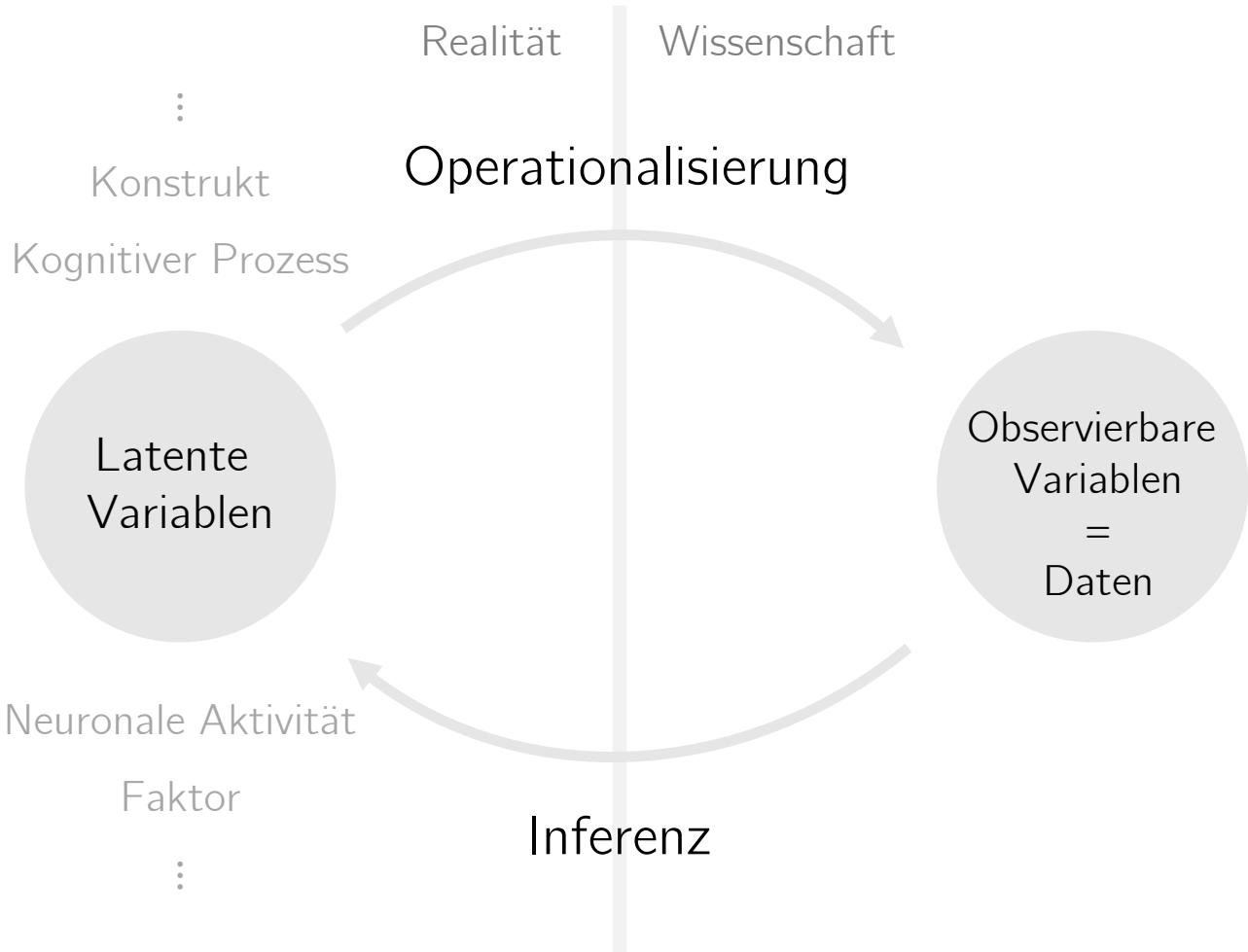
(10) Explorative Faktorenanalyse

Datum	Einheit	Thema
14.10.2022	Grundlagen	(1) Einführung
21.10.2022	Grundlagen	(2) Vektoren
28.10.2022	Grundlagen	(3) Matrizen
04.11.2022	Grundlagen	(4) Eigenanalyse
11.11.2022	Grundlagen	(5) Multivariate Wahrscheinlichkeitstheorie
18.11.2022	Grundlagen	(6) Multivariate Normalverteilungen
25.11.2022	Frequentistische Inferenz	(7) Kanonische Korrelationsanalyse
02.12.2022	Frequentistische Inferenz	(8) T^2-Tests
09.12.2022	Frequentistische Inferenz	(8) T^2-Tests
16.12.2022	Latente Lineare Modelle	(9) Hauptkomponentenanalyse
	Weihnachtspause	
13.01.2023	Latente Lineare Modelle	(10) Explorative Faktorenanalyse
20.01.2023	Latente Lineare Modelle	(11) Konfirmative Faktorenanalyse
27.01.2023	Gesamtkurs	Fragen und Antworten
Sommer 2023	Klausur	

Modellbasierte Datenwissenschaft



Psychologische Datenwissenschaft



Psychologische Datenwissenschaft

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)

- Altmodisches Verfahren mit impliziter Modellspezifikation und prinzipienfreier Schätzung
- Fokus auf der numerische Behandlung von Stichprobenkovarianzmatrizen

Konfirmative Faktorenanalyse (CFA)

- Moderneres Verfahren mit expliziter probabilistische 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 (1921) 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 zu hierarchischen und nicht normalverteilten Szenarien (vgl. Bartholomew (2011))

Software Periode (1970 - heute)

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

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Anwendungsbeispiel

Sedimentationsdatensatz nach Rencher and Christensen (2012)

Einschätzungen von 7 Personen (P1-P7) auf einer Skala von 1 bis 9 bezüglich 5 Adjektiven durch 1 Probandin

```
fname        = file.path(getwd(),  "10_Explorative_Faktorenanalyse.csv")  # Dateiname
YT           = read.table(fname, sep = ",", header = T)                   # transponierte Datenmatrix
Y            = t(YT)                                                      # Datenmatrix
colnames(Y)  = paste("P", 1:ncol(Y), sep = "")                            # Personenlabels
```

Datenmatrix $Y \in \mathbb{R}^{m \times n}$ mit $m = 5$ und $n = 7$

	P1	P2	P3	P4	P5	P6	P7
Freundlich	1	8	9	9	1	9	9
Froh	5	7	9	9	1	7	9
Nett	1	9	9	9	1	9	9
Intelligent	5	9	8	9	9	7	7
Gerecht	1	8	8	9	9	9	7

Anwendungsbeispiel

```
# Evaluation von Stichprobenkovarianz- und Stichprobenkorrelationsmatrix
Y         = as.matrix(Y)                                    # Y \in \mathbb{R}^{m x n}
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
```

Stichprobenkovarianzmatrix C

	Freundlich	Froh	Nett	Intelligent	Gerecht
Freundlich	14.62	9.857	14.86	1.690	5.98
Froh	9.86	8.571	9.90	-0.095	1.09
Nett	14.86	9.905	15.24	1.905	6.09
Intelligent	1.69	-0.095	1.90	2.238	3.60
Gerecht	5.98	1.095	6.09	3.595	8.24

Stichprobenkorrelationsmatrix R

	Freundlich	Froh	Nett	Intelligent	Gerecht
Freundlich	1.000	0.881	0.995	0.296	0.545
Froh	0.881	1.000	0.867	-0.022	0.130
Nett	0.995	0.867	1.000	0.326	0.544
Intelligent	0.296	-0.022	0.326	1.000	0.837
Gerecht	0.545	0.130	0.544	0.837	1.000

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

- Wir bezeichnen Werte von υ mit $y \in \mathbb{R}^m$, von ξ mit $x \in \mathbb{R}^k$ und von ε mit $e \in \mathbb{R}^m$.
- 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 in der Fragenbogendatenanalyse oft *Items* oder *Indikatorvariablen* genannt.
- 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

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

  die in Wahrscheinlichkeitsdichtfunktionsform geschrieben werden kann als

$$p(x, y) = p(x)p(y|x) \text{ mit } p(x) = N(x; 0_k, I_k) \text{ und } p(y|x) = N(y; Lx, \Psi). \tag{7}$$

  Wir werden diese Form im Kontext der konfirmativen Faktoranalyse genauer betrachten.
- Die Generation von n Realisierungen eines EFA Modells resultiert in einem Datensatz der Form

$$D = \left(\begin{pmatrix} x^{(1)} \\ y^{(1)} \end{pmatrix} \quad \cdots \quad \begin{pmatrix} x^{(n)} \\ y^{(n)} \end{pmatrix} \right) \in \mathbb{R}^{(k+m)\times n} \tag{8}$$

  aus konkatenierten virtuellen latenten Datenvektoren $x^{(i)} \in \mathbb{R}^k$ und beobachteten Daten $y^{(i)} \in \mathbb{R}^m$ für $i = 1, \ldots, n$. In der Anwendung sind die virtuellen Datenvektoren $x^{(i)} \in \mathbb{R}^k$ natürlich nicht vorhanden, sondern lediglich ein Datensatz an beobachteten Vektoren

$$Y = \begin{pmatrix} y^{(1)} & \cdots & y^{(n)} \end{pmatrix} \in \mathbb{R}^{m\times n}. \tag{9}$$

  Die probabilitische Inferenz über die Werte des latenten Faktorenvektors wird in der Faktorenanalyseliteratur oft als *Evaluation von Faktorscores* bezeichnet. Wir wollen sie hier vernachlässigen. Einen konzisen und modernen Einstieg in die Aspekte von Faktorscoreinferenz (Inferenz) und Parameterschätzung (Lernen) im Rahmen der Faktorenanalyse bietet (5) Faktorenanalyse aus der letztjährigen Iteration des Kurses.

Simulationsbeispiel

```
# Modellformulierung
k     = 2                                   # Dimension des latenten Zufallsvektors
m     = 5                                   # Dimension des beobachtbaren Zufallsvektors
n     = 7                                   # Beobachtungsanzahl
L     = matrix(c(1,0,                       # Faktorenladungsmatrix
                 1,0,
                 1,0,
                 0,1,
                 0,1),
                 nrow  = m,
                 byrow = TRUE)
Psi   = diag(c(2,2,4,5,2))                  # Beobachtungsrauschenkovarianmatrix
library(MASS)                               # Multivariates Normalverteilungspaket
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 Faktoren Zufallsvektors
  eps    = mvrnorm(1,rep(0,m), Psi)         # Realisierung des latenten Faktoren Zufallsvektors
  Y[,i] = L %*% x + eps                     # Realisierung des beobachtbaren Datenzufallsvektors
}
```

Simulierter Datensatz $Y \in \mathbb{R}^{m \times n}, m = 5, n = 7$

	1	2	3	4	5	6	7
y_1	0	2	-1	2	1	-2	0
y_2	0	0	1	2	2	0	2
y_3	-2	0	-1	-1	3	-3	-4
y_4	0	4	-2	-5	2	1	2
y_5	-1	0	2	-2	1	2	1

Theorem (Datenkovarianzmatrix der explorativen Faktorenanalyse)

Gegeben sei ein EFA Modell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{10}$$

Dann gilt für die marginale Kovarianzmatrix des Datenvektors

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

Beweis

Mit dem Theorem zu den Eigenschaften der Kovarianzmatrix (vgl. (5) Multivariate Wahrscheinlichkeitstheorie) gilt aufgrund der Unabhängigkeit von ξ und ε

$$\mathbb{C}(\upsilon) = L\mathbb{C}(\xi)L^T + \mathbb{C}(\varepsilon) = LI_kL^T + \Psi = LL^T + \Psi. \tag{12}$$

Bemerkungen

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

$$C = \frac{1}{n-1}\left(Y\left(I_n - \frac{1}{n}1_{nn}\right)Y^T\right). \tag{13}$$

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

$$C = \hat{L}\hat{L}^T + \hat{\Psi}. \tag{14}$$

Theorem (Varianzzerlegung der explorativen Faktorenanalyse)

Gegeben sei ein EFA Modell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_k, \Psi). \tag{15}$$

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

$$\mathbb{C}(\upsilon_i, \upsilon_i) = \sum_{j=1}^{k} l_{ij}^2 + \psi_i. \tag{16}$$

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

$$l_i l_i^T = \begin{pmatrix} l_{i1} & \cdots & l_{ik} \end{pmatrix} \begin{pmatrix} l_{i1} \\ \vdots \\ l_{ik} \end{pmatrix} = \sum_{j=1}^{k} l_{ij}^2 \tag{17}$$

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

Beweis

Mit dem Theorem zu Datenkovarianzmatrix der explorativen Faktorenanalyse gilt

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

$$= \begin{pmatrix} l_{11} & \cdots & l_{1k} \\ l_{21} & \cdots & l_{2k} \\ \vdots & \ddots & \vdots \\ l_{m1} & \cdots & l_{mk} \end{pmatrix} \begin{pmatrix} l_{11} & \cdots & l_{m1} \\ l_{12} & \cdots & l_{m2} \\ \vdots & \ddots & \vdots \\ l_{1k} & \cdots & l_{mk} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \psi_m \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{k} l_{1j} l_{1j} & \sum_{j=1}^{k} l_{1j} l_{2j} & \cdots & \sum_{j=1}^{k} l_{1j} l_{mj} \\ \sum_{j=1}^{k} l_{2j} l_{1j} & \sum_{j=1}^{k} l_{2j} l_{2j} & \cdots & \sum_{j=1}^{k} l_{2j} l_{mj} \\ \vdots & \cdots & \ddots & \vdots \\ \sum_{j=1}^{k} l_{mj} l_{1j} & \sum_{j=1}^{k} l_{mj} l_{2j} & \cdots & \sum_{j=1}^{k} l_{mj} l_{mj} \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \psi_m \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{k} l_{1j}^2 + \psi_1 & \sum_{j=1}^{k} l_{1j} l_{2j} & \cdots & \sum_{j=1}^{k} l_{1j} l_{mj} \\ \sum_{j=1}^{k} l_{2j} l_{1j} & \sum_{j=1}^{k} l_{2j}^2 + \psi_2 & \cdots & \sum_{j=1}^{k} l_{2j} l_{mj} \\ \vdots & \cdots & \ddots & \vdots \\ \sum_{j=1}^{k} l_{mj} l_{1j} & \sum_{j=1}^{k} l_{mj} l_{2j} & \cdots & \sum_{j=1}^{k} l_{mj}^2 + \psi_m \end{pmatrix}.$$

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.

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.

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

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

- $C(\upsilon_i, \upsilon_i) := c_{ii}$ 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

```
# EFA mit Hauptkomponentenschätzung für k = 2
fname     = file.path(getwd(),"10_Explorative_Faktorenanalyse.csv") # Dateiname
YT        = read.table(fname, sep = ",", header = T)                # 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
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 $k = 2$

```
> L_hat =

>       [,1]   [,2]
> [1,] -3.81  0.160
> [2,] -2.55  1.353
> [3,] -3.89  0.115
> [4,] -0.53 -1.224
> [5,] -1.66 -2.327

> Psi_hat =

>      [,1]  [,2]  [,3] [,4]   [,5]
> [1,] 0.04 0.000 0.000 0.00 0.0000
> [2,] 0.00 0.263 0.000 0.00 0.0000
> [3,] 0.00 0.000 0.117 0.00 0.0000
> [4,] 0.00 0.000 0.000 0.46 0.0000
> [5,] 0.00 0.000 0.000 0.00 0.0838
```

Varianz-, Kommunalitäts, und Spezifitätsschätzer bei $k = 2$

```
> V_i_hat   = 14.6 8.57 15.2 2.24 8.24
> h2_i_hat  = 14.6 8.31 15.1 1.78 8.15
> psi_i_hat = 0.04 0.263 0.117 0.46 0.0838
```

Anwendungsbeispiel

Das geschätzte Faktorenanalysemodell ist also

$$\upsilon = \hat{L}\xi + \varepsilon \Leftrightarrow \begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \upsilon_4 \\ \upsilon_5 \end{pmatrix} = \begin{pmatrix} -3.81 & 0.16 \\ -2.54 & 1.35 \\ -3.89 & 0.11 \\ -0.53 & -1.22 \\ -1.66 & -2.32 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix} \tag{42}$$

mit

$$\xi \sim (0_2, I_2) \text{ und } \varepsilon \sim (0_5, \hat{\Psi}), \tag{43}$$

und

$$\hat{\Psi} = \begin{pmatrix} 0.04 & 0 & 0 & 0 & 0 \\ 0 & 0.26 & 0 & 0 & 0 \\ 0 & 0 & 0.12 & 0 & 0 \\ 0 & 0 & 0 & 0.46 & 0 \\ 0 & 0 & 0 & 0 & 0.08 \end{pmatrix} \tag{44}$$

Modellformulierung

Modellschätzung

Modellevaluation

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

$$G = F + R \tag{45}$$

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

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{46}$$

$C \in \mathbb{R}^{m \times m}$ 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

$$C \approx \hat{L}\hat{L}^T + \hat{\Psi}. \tag{47}$$

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

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim (0_k, I_k) \text{ und } \varepsilon \sim (0_m, \Psi). \tag{48}$$

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,
- $\hat{\Psi} = \text{diag}(\hat{\psi}_i, \ldots, \hat{\psi}_m) \in \mathbb{R}^{m \times m}$ 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

$$G = F + R. \tag{49}$$

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

$$F = \sum_{j=1}^{k} \lambda_j, \text{ wobei } \lambda_j = \sum_{i=1}^{m} \hat{l}_{ij}^2 \tag{50}$$

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

Beweis

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

$$\sum_{j=1}^{k} \hat{l}_{ij}^2 \tag{51}$$

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

$$\begin{aligned}
\hat{L}\hat{L}^T &= \begin{pmatrix} \hat{l}_{11} & \cdots & \hat{l}_{1k} \\ \hat{l}_{21} & \cdots & \hat{l}_{2k} \\ \vdots & \ddots & \vdots \\ \hat{l}_{m1} & \cdots & \hat{l}_{mk} \end{pmatrix} \begin{pmatrix} \hat{l}_{11} & \cdots & \hat{l}_{m1} \\ \hat{l}_{12} & \cdots & \hat{l}_{m2} \\ \vdots & \ddots & \vdots \\ \hat{l}_{1k} & \cdots & \hat{l}_{mk} \end{pmatrix} \\
&= \begin{pmatrix} r\sum_{j=1}^{k} l_{1j}^2 & \sum_{j=1}^{k} l_{1j}l_{2j} & \cdots & \sum_{j=1}^{k} l_{1j}l_{mj} \\ \sum_{j=1}^{k} l_{2j}l_{1j} & \sum_{j=1}^{k} l_{2j}^2 & \cdots & \sum_{j=1}^{k} l_{2j}l_{mj} \\ \vdots & \cdots & \ddots & \vdots \\ \sum_{j=1}^{k} l_{mj}l_{1j} & \sum_{j=1}^{k} l_{mj}l_{2j} & \cdots & \sum_{j=1}^{k} l_{mj}^2 \end{pmatrix}
\end{aligned} \tag{52}$$

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

$$\hat{\psi}_i := c_{ii} - \sum_{j=1}^{k} \hat{l}_{ij}^2 \text{ für } i = 1, \ldots, m \tag{53}$$

konstruiert wird. Um als nächstes

$$F = \sum_{j=1}^{k} \lambda_j \tag{54}$$

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

$$q_j^T q_j = \sum_{i=1}^{m} q_{ij}^2 = 1. \tag{55}$$

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

$$F = \sum_{i=1}^{m}\sum_{j=1}^{k} \hat{l}_{ij}^2 = \sum_{j=1}^{k}\sum_{i=1}^{m} \hat{l}_{ij}^2 = \sum_{j=1}^{k}\sum_{i=1}^{m} \left(\sqrt{\lambda}_j q_{ij}\right)^2 = \sum_{j=1}^{k} \lambda_j \sum_{i=1}^{m} q_{ij}^2 = \sum_{j=1}^{k} \lambda_j \tag{56}$$

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

$$\sum_{i=1}^{m} \hat{l}_{ij}^2 = \lambda_j \text{ für } j = 1, \ldots, k. \tag{57}$$

Anwendungsbeispiel

```
# EFA mit Hauptkomponentenschätzung für k = 2
fname     = file.path(getwd(),"10_Explorative_Faktorenanalyse.csv") # Dateiname
YT        = read.table(fname, sep = ",", header = T)                # 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   = 48.9
> F   = 47.9
> R   = 0.963
> F+B = 48.9
```

Darstellung der Faktorenbasierten Stichprobenvarianz

```
> Faktorbasierte Stichprobenvarianz F        = 47.9
> Summe der Eigenwerte lambda_1,...,lambda_k = 47.9
```

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 λ_j, λ_j / G und $\sum_{j=1}^{k} \lambda_j / G$ zu visualisieren.

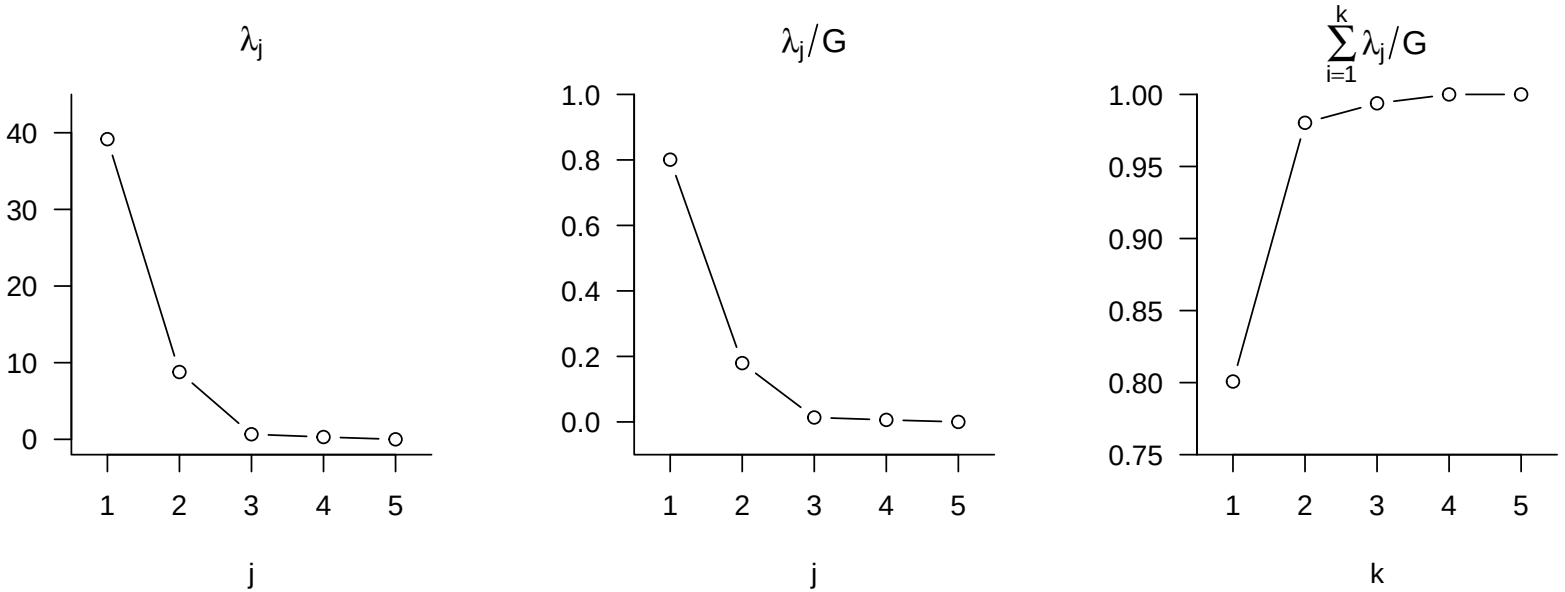
Basierend darauf kann man dann k so wählen, dass k möglichst klein und $\sum_{j=1}^{k} \lambda_j / G$ möglichst groß ist.

Die Visualisierung der λ_j wird in diesem Kontext *Scree-Plot* genannt.

Anwendungsbeispiel

```
# EFA mit Hauptkomponentenschätzung für k = 5
fname     = file.path(getwd(),"10_Explorative_Faktorenanalyse.csv") # Dateiname
YT        = read.table(fname, sep = ",", header = T)               # 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 80% der Gesamtstichprobenvarianz erklärt werden.
- Mit $k = 2$ können 98% der Gesamtstichprobenvarianz erklärt werden.
- Mit $k = 3$ können 99% der Gesamtstichprobenvarianz erklärt werden.

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

Überblick Modellevaluation | Modellinterpretation

Modellinterpretation = Rotationsverfahren

Per Datenkomponente sind Faktorladungen gewünscht, die möglichst leicht eine eindeutige Faktorzuordnung erlauben

Statt der geschätzten Faktorladungsmatrix $\hat{L}$ wäre also eine Faktorladungsmatrix wie $\hat{L}^*$ gewünscht.

$$\hat{L} = \begin{pmatrix} -3.81 & 0.16 \\ -2.54 & 1.35 \\ -3.89 & 0.11 \\ -0.53 & -1.22 \\ -1.66 & -2.32 \end{pmatrix} \quad \Rightarrow \quad \hat{L}^* = \begin{pmatrix} 1.00 & 0.00 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \\ 0.00 & 1.00 \end{pmatrix} \tag{58}$$

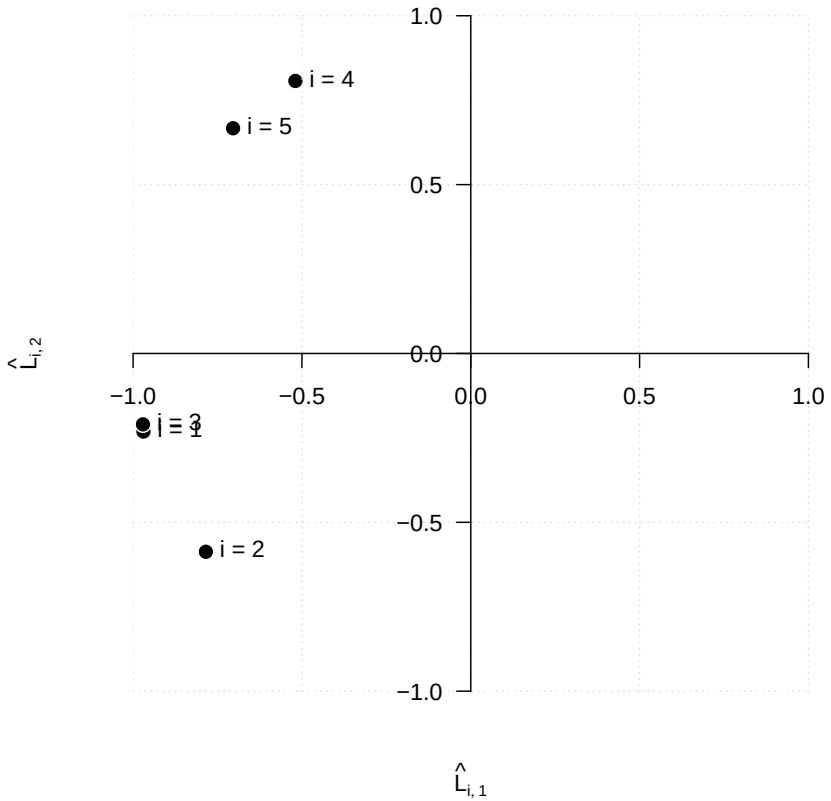
Eindeutige Zuordnungen von Datenkomponenten zu Faktoren induzieren dann Cluster von Datenkomponenten, "die auf jeweils einen Faktor laden". Eine entsprechend modifizierte Faktorladungsmatrix $\hat{L}^*$ nennt man auch "Einfachstruktur" und man hofft dann, durch Inspektion der Cluster zu eine inhaltlichen 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 Datenvektorkomponten



$i = 1$ Freundlich, $i = 2$ Froh, $i = 3$ Nett, $i = 4$ Intelligent, $i = 5$ Gerecht

$\Rightarrow$ Cluster 1: Freundlich, Froh, Nett $\Rightarrow$ Cluster 2: Intelligent, Gerecht

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

Es sei

$$M_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \in \mathbb{R}^{2\times 2} \text{ für } 0 \leq \theta \leq 2\pi \tag{59}$$

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

$$B_w := \left\{ \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} \right\} \tag{60}$$

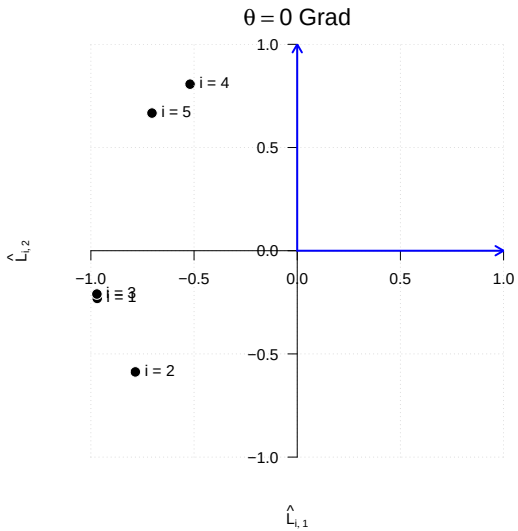
Bemerkungen

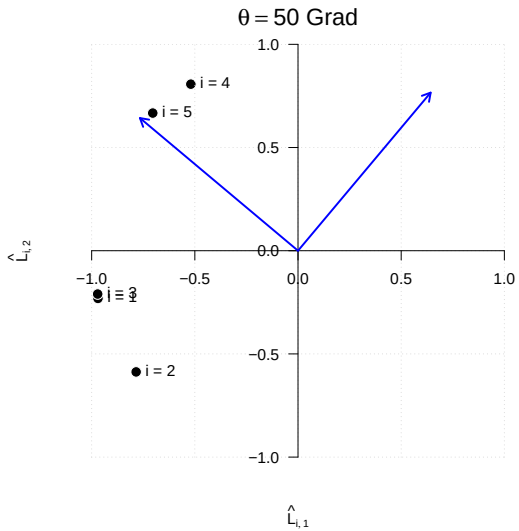
- Wir verzichten auf einen Beweis.
- Das Theorem stellt eine unendliche, durch θ parameterisierte Menge von Orthonormalbasen von $\mathbb{R}^2$ bereit und ermöglicht weiterhin, die Faktorladungskoordinaten jeder Datenkomponente bezüglich jeder dieser Orthonormalbasen zu evaluieren. Wir bezeichnen die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

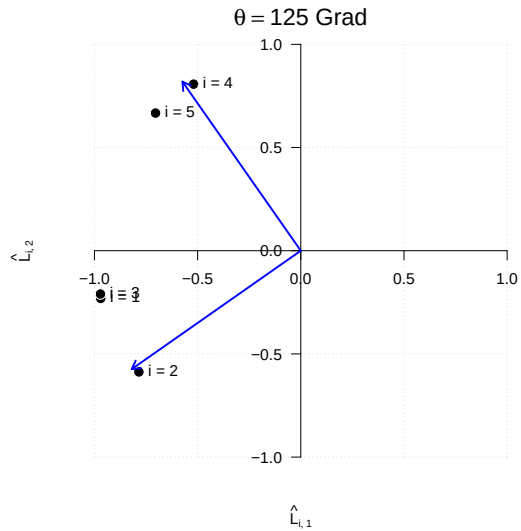
$$\hat{L}_\theta := \left(M_\theta^T \hat{L}^T \right)^T \tag{61}$$

Anwendungsbeispiel

Drehmatrixbasisvektoren und Faktorladungskoordinatenmatrizen







$$\hat{L}_\theta = \begin{pmatrix} -0.97 & -0.23 \\ -0.79 & -0.59 \\ -0.97 & -0.21 \\ -0.52 & 0.80 \\ -0.70 & 0.66 \end{pmatrix}$$

$$\hat{L}_\theta = \begin{pmatrix} -0.80 & 0.59 \\ -0.95 & 0.22 \\ -0.78 & 0.61 \\ +0.28 & 0.92 \\ +0.06 & 0.96 \end{pmatrix}$$

$$\hat{L}_\theta = \begin{pmatrix} 0.37 & 0.93 \\ -0.03 & 0.98 \\ 0.38 & 0.92 \\ 0.96 & -0.04 \\ 0.95 & 0.19 \end{pmatrix}$$

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

$$f : \mathbb{R}^{m \times 2} \to \mathbb{R}_{\geq 0}, \hat{L} \mapsto f(\hat{L}) := \sum_{j=1}^{2} \sum_{i=1}^{m} \left(\hat{l}_{ij}^2 - \bar{l}_j^2 \right)^2 \text{ mit } \bar{l}_j^2 := \frac{1}{m} \sum_{i=1}^{m} \hat{l}_{ij}^2. \tag{62}$$

Weiterhin sei

$$\hat{L}_\theta := \left(M_\theta^T \hat{L}^T \right)^T \tag{63}$$

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

$$\hat{L}_\theta^* := \underset{0 \leq \theta \leq 2\pi}{\arg\max} f(\hat{L}_\theta) \tag{64}$$

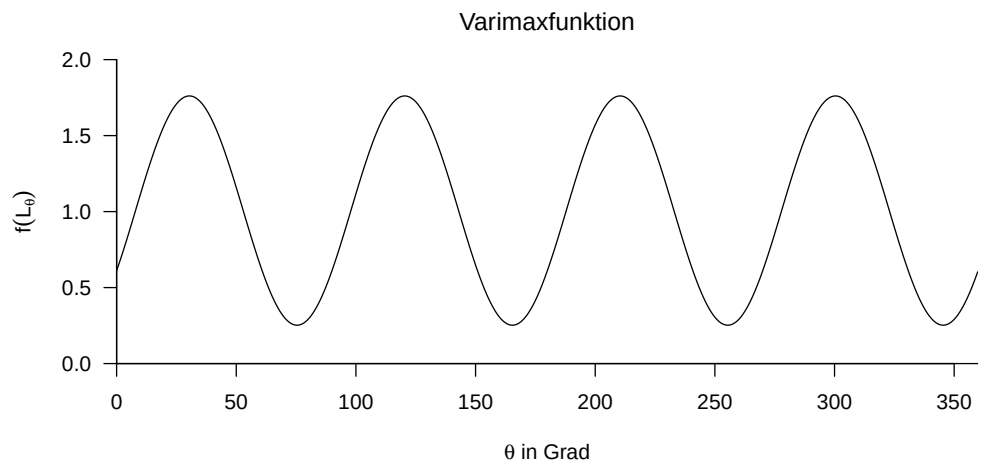
die *Varimaxfaktorladungsmatrix*.

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

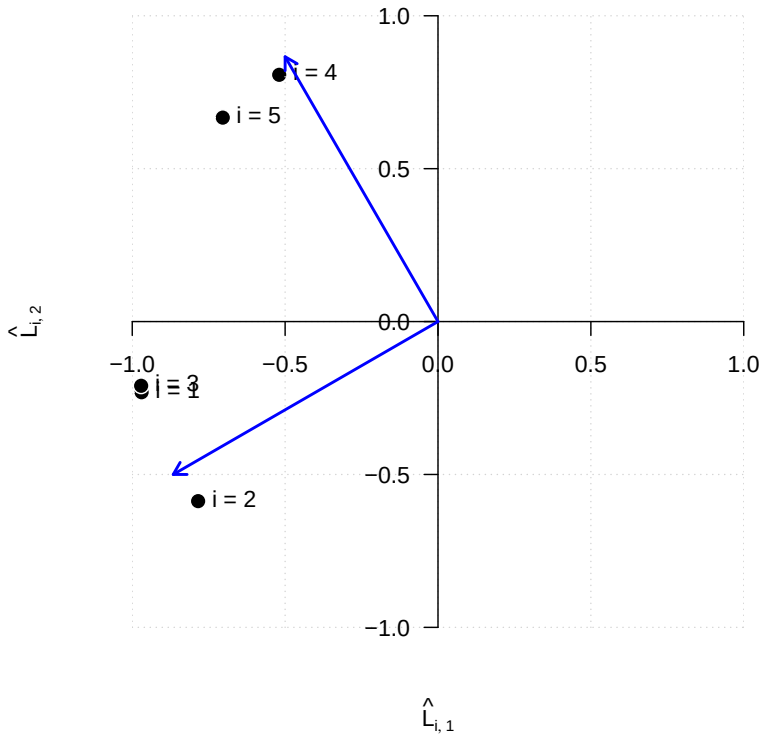
```
f = function(L){
  # Diese Funktion evaluiert die Varimaxfunktion.
  # Input
  #     L      : m x 2 Faktorladungsmatrix
  # Output
  #     f      : 1 x 1 Funktionswert f(L)
  # ------------------------------------------------------------------------
  return((nrow(L)-1)*(var(L[,1]^2) + var(L[,2]^2)))}                    # Varimaxfunktion
theta           = seq(0,2*pi,0.01)                                    # \theta Raum
fL_hat_theta    = rep(NaN, length(theta))                             # Funktionswertarray
for(i in 1:length(theta)){
  M_theta         = matrix(c(cos(theta[i]),-sin(theta[i]) ,           # Basisvektormatrix
                             sin(theta[i]), cos(theta[i])), nrow = 2, byrow = T)
  fL_hat_theta[i] = f(t(t(M_theta) %*% t(L_hat)))}                      # Varimaxfunktionsauswertung
```



$\Rightarrow$ Maxima für $\theta = 30, \theta = 120, \theta = 210, \theta = 300$.

Anwendungsbeispiel

Varimaxlösung



$$\hat{L}_\theta^* = \begin{pmatrix} 0.28 & 0.96 \\ -0.12 & 0.97 \\ 0.30 & 0.95 \\ 0.96 & 0.05 \\ 0.93 & 0.28 \end{pmatrix} \tag{65}$$

Freundlich y_1, Froh y_2 und Nett y_3 laden auf Faktor x_2, Intelligent y_4 und Gerecht y_5 laden auf Faktor x_1

$\Rightarrow$ Faktor x_1 mag die Rationalität einer Person, Faktor x_2 die Liebenswürdigkeit einer Person repräsentieren

Modellformulierung

Modellschätzung

Modellevaluation

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

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

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

Fisher, R. A. 1921. "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.

Hotelling, Harold. 1933. "Analysis of Complext 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.

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

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

Muthén, L. K., and B. O. Muthén. 1998. *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.

Rencher, Alvin C., and William F. Christensen. 2012. *Methods of Multivariate Analysis*. Third Edition. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley.

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.