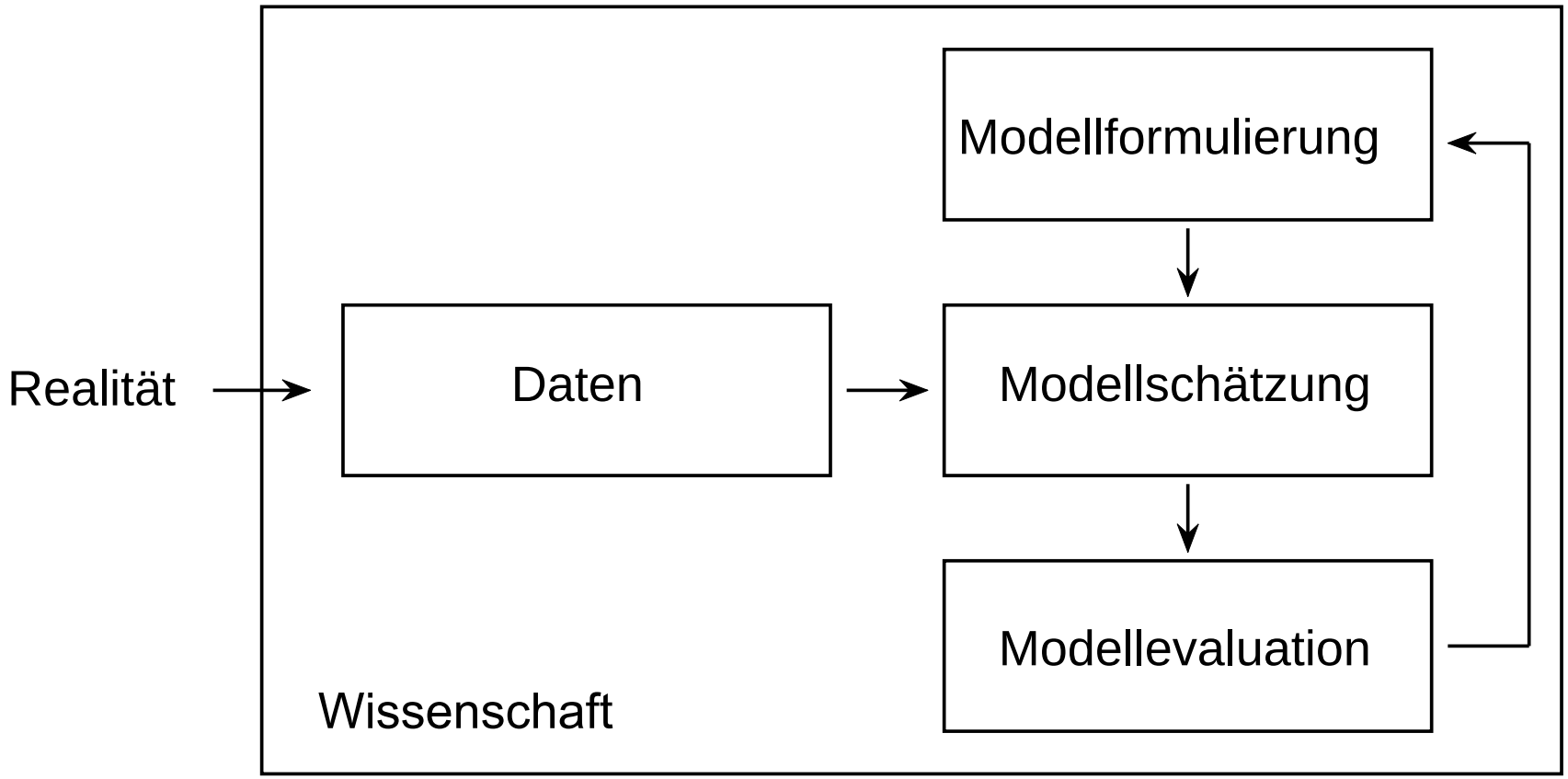
Multivariate Datenanalyse

MSc Psychologie WiSe 2022/23

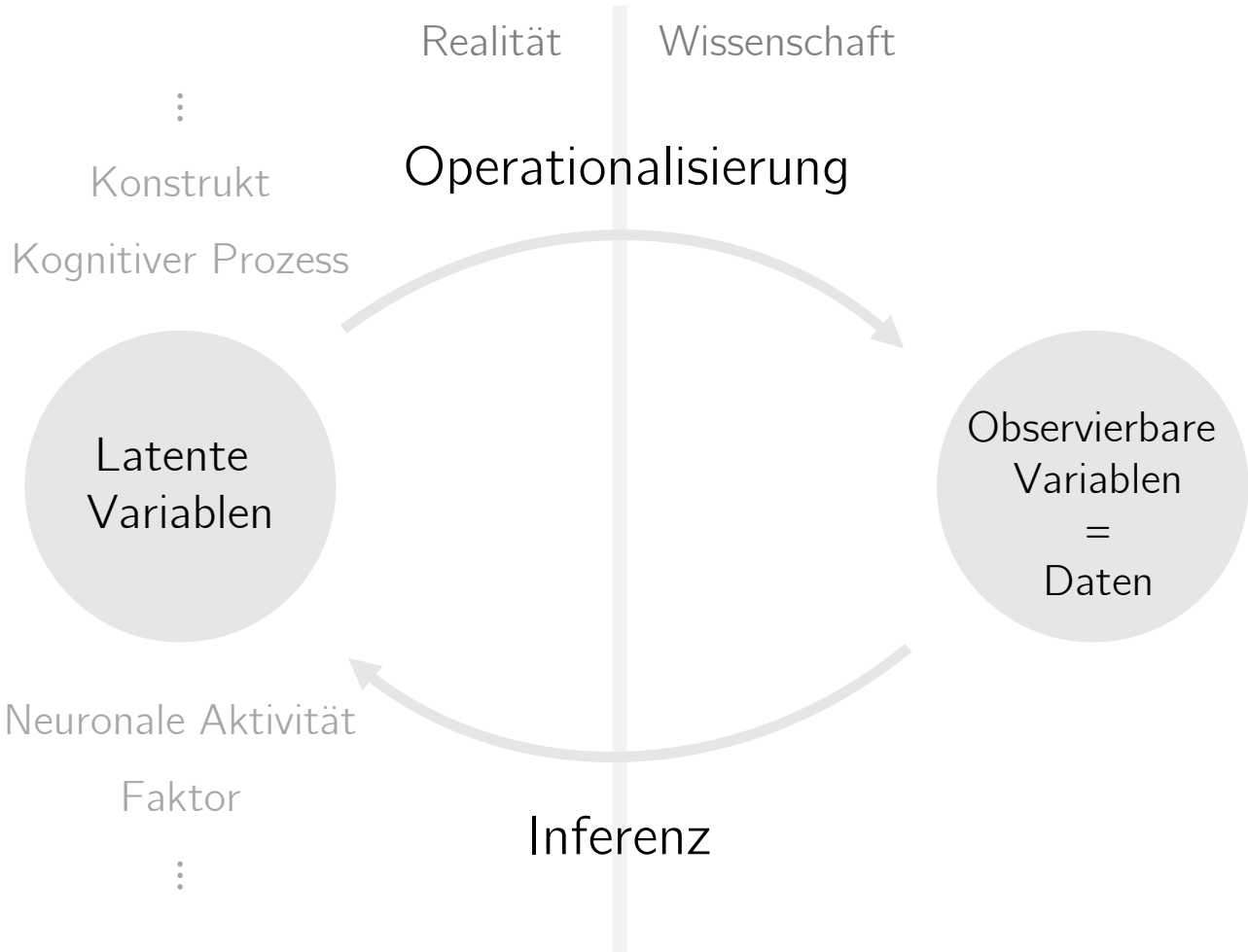
Prof. Dr. Dirk Ostwald

(11) Konfirmatorische Faktorenanalyse

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 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 (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

Anwendungsbeispiel

Intelligenzforschungsdatensatz nach Holzinger and Swineford (1939)

Visualisierungsaufgaben

1. Visual Perception
2. Cubes
3. Lozenges

Verbalisierungsaufgaben

4. Paragraph Comprehension
5. Sentence Completion
6. Word Meaning

Schnelligkeitsaufgaben

7. Addition
8. Counting dots
9. Straight-Curved Capitals

Anwendungsbeispiel

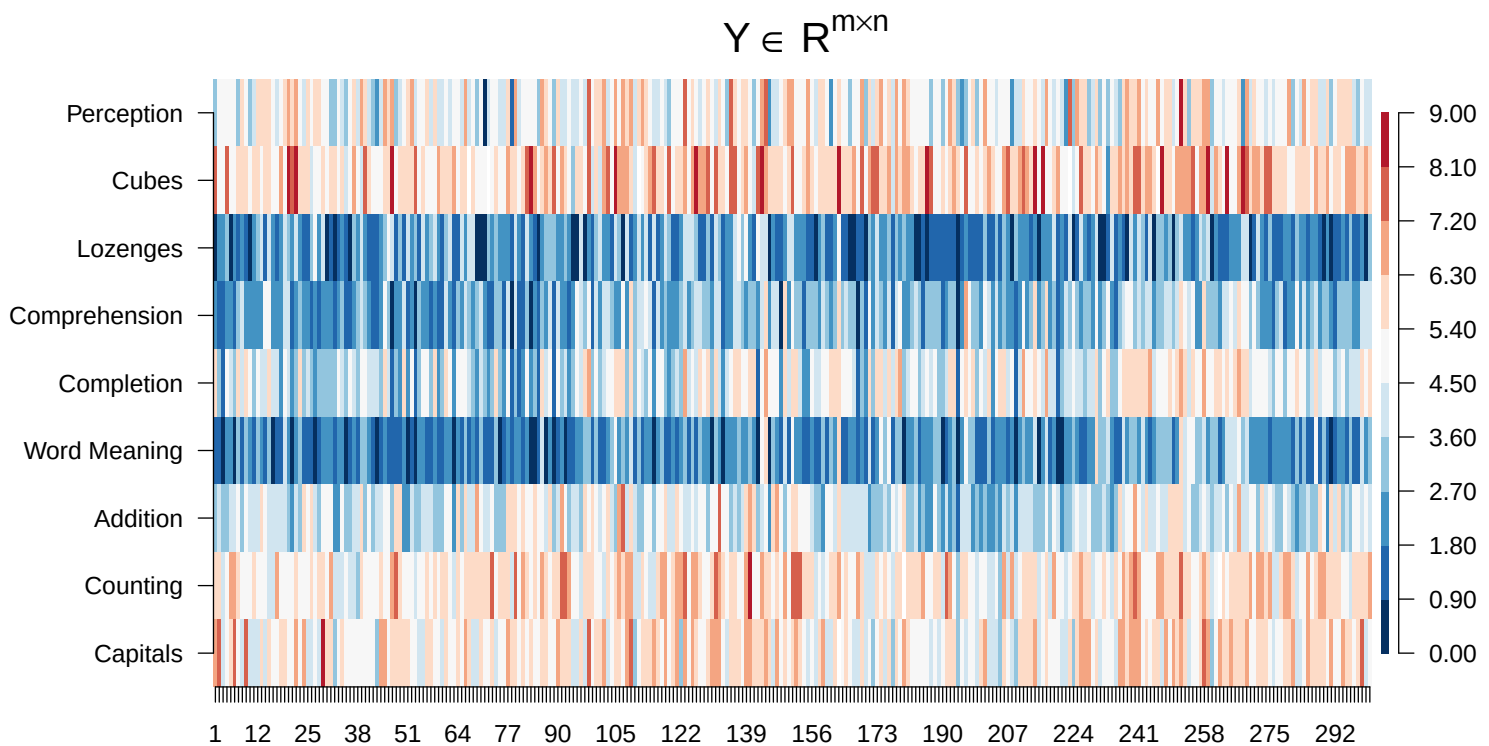
Beobachteter Datensatz (n = 301)

- 301 Proband:innen | 11 - 16 Jahre
- Probandin:innen 1 - 10

	1	2	3	4	5	6	7	8	9	10
Perception	3.33	5.33	4.50	5.33	4.83	5.33	2.83	5.67	4.50	3.50
Cubes	7.75	5.25	5.25	7.75	4.75	5.00	6.00	6.25	5.75	5.25
Lozenges	0.38	2.12	1.88	3.00	0.88	2.25	1.00	1.88	1.50	0.75
Comprehension	2.33	1.67	1.00	2.67	2.67	1.00	3.33	3.67	2.67	2.67
Completion	5.75	3.00	1.75	4.50	4.00	3.00	6.00	4.25	5.75	5.00
Word Meaning	1.29	1.29	0.43	2.43	2.57	0.86	2.86	1.29	2.71	2.57
Addition	3.39	3.78	3.26	3.00	3.70	4.35	4.70	3.39	4.52	4.13
Counting	5.75	6.25	3.90	5.30	6.30	6.65	6.20	5.15	4.65	4.55
Capitals	6.36	7.92	4.42	4.86	5.92	7.50	4.86	3.67	7.36	4.36

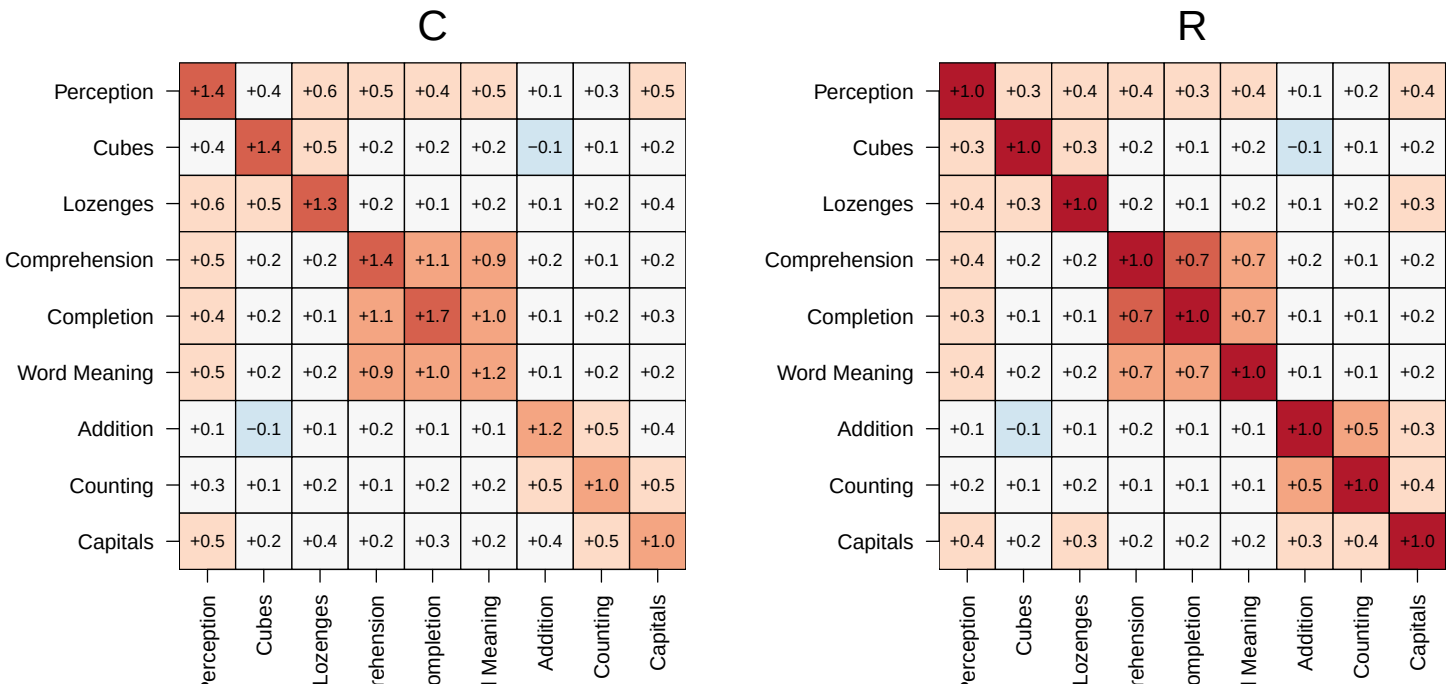
Anwendungsbeispiel

Beobachteter Datensatz (n = 301)



Anwendungsbeispiel

Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix



Überblick

Konzeption des Faktorenanalysemodells vor dem Hintergrund Frequentistischer Modellbildung

- Normalverteilungsannahmen bezüglich des Zustands- und Beobachtungsrauschens
- Relaxation der Diagonaliät des Zustandsrauschenkovarianzmatrixparameters
- Parameterrestriktionen zur Induktion identifizierbarer Modelle
- Induktion einer Log-Likelihood-Ratio-Kriterium-basierten Modellschätzung
- Induktion der Möglichkeit Frequentistischer Parameterinferenz

Definition (Modell der konfirmatorischen 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,
- $\xi \sim N(0_k, \Phi)$ ein k-dimensionaler latenter normalverteilter Zufallsvektor ist,
- $\varepsilon \sim N(0_m, \Psi)$ ein m-dimensionaler latenter normalverteilter Zufallsvektor ist mit

$$\Psi := \text{diag}\,(\psi_1, \ldots, \psi_m) \tag{2}$$

- υ ein m-dimensionaler beobachtbarer Zufallsvektor ist,

Dann wird (1) *Modell der konfirmatorischen Faktorenanalyse (CFA modell) mit Parametern* L, Φ *und* Ψ genannt.

Bemerkungen

- Die Interpretationen der Modellbestandteile entsprechenden denen des EFA Modells.
- Wir bezeichnen Werte von υ mit $y \in \mathbb{R}^m$, von ξ mit $x \in \mathbb{R}^k$ und von ε mit $e \in \mathbb{R}^m$.
- Es wird explizit die Normalverteilung von ξ und ε angenommen.
- Die Kovarianzmatrix von ξ muss nicht die Identitätsmatrix sein.

Theorem (WDF der gemeinsamen Verteilung von Faktoren und Daten)

Es sei

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

Dann hat die gemeinsame Verteilung von ξ und υ die Wahrscheinlichkeitsdichteform

$$p(x, y) = N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 0_k \\ 0_m \end{pmatrix}, \begin{pmatrix} \Phi & \Phi L^T \\ L\Phi & L\Phi L^T + \Psi \end{pmatrix}\right). \tag{4}$$

Bemerkung

Die gemeinsame Verteilung von ξ und υ des konfimatorischen Faktorenanalysemodells ist die Grundlage

(1) der Form der marginalen Datenkovarianzmatrix,
(2) der marginalen Log-Likelihood-Funktion der konfirmatorischen Faktorenanalyse,
(3) der daraus resultierenden Diskrepanzfunktion der CFA Parameterschätzung,
(4) der Evaluation von Faktorenscores.

Beweis

Wir haben in Einheit (10) Konfirmatorische Faktorenanalyse gesehen, dass die Wahrscheinlichkeitsdichtefunktion des konfirmatorischen Faktorenanalysemodells geschrieben werden kann als

$$p(x, y) = p(x)p(y|x) \text{ mit } p(x) = N(x; 0_k, \Phi) \text{ und } p(y|x) = N(y; Lx, \Psi). \tag{5}$$

Mit dem Theorem zu Gemeinsamen Normalverteilungen (vgl. Einheit (6) Multivariate Normalverteilungen) ergibt sich dann für den Erwartungswertparameter $\mu_{\xi,\upsilon} \in \mathbb{R}^{k+m}$ der gemeinsamen Verteilung

$$\mu_{\xi,\upsilon} = \begin{pmatrix} 0_k \\ L0_k \end{pmatrix} = \begin{pmatrix} 0_k \\ 0_m \end{pmatrix} \tag{6}$$

und für den Kovarianzmatrixparameter $\Sigma_{\xi,\upsilon} \in \mathbb{R}^{(k+m)\times(k+m)}$ ergibt sich direkt

$$\Sigma_{\xi,\upsilon} = \begin{pmatrix} \Phi & \Phi L^T \\ L\Phi & L\Phi L^T + \Psi \end{pmatrix} \tag{7}$$

Damit folgt das Theorem dann aber schon direkt.

□

Theorem (WDF und Eigenschaften der marginalen Datenverteilung)

Es sei

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi) \tag{8}$$

Dann hat die marginale Verteilung der Daten υ die Wahrscheinlichkeitsdichteform

$$p(y) = N\left(y; 0_m, L\Phi L^T + \Psi\right) \tag{9}$$

sowie den Erwartungswert und die Kovarianzmatrix

$$\mathbb{E}(\upsilon) = 0_m \text{ und } \mathbb{C}(\upsilon) = L\Phi L^T + \Psi. \tag{10}$$

Beweis

Das Theorem ergibt sich direkt mit dem Theorem zu marginalen Normalverteilungen (vgl. Einheit (6) Multivariate Normalverteilungen) und den Ergebnissen zu Erwartungswert und Kovarianzmatrix multivariater Normalverteilungen.

Bemerkungen

- Für $\Phi := I_k$ ergibt sich mit $LL^T + \Psi$ die marginale Datenkovarianzmatrix der EFA.
- Wie die EFA fokussiert die CFA Parameterschätzung auf die Approximation der Stichprobenkovarianzmatrix,

$$C \approx \hat{L}\hat{\Phi}\hat{L}^T + \hat{\Psi}. \tag{11}$$

Anwendungsbeispiel

Für das Anwendungsbeispiel nach Holzinger and Swineford (1939) gilt $m = 9$ und $C \in \mathbb{R}^{9\times 9}$.

Die drei Aufgabentypen legen jeweils einen gemeinsamen Faktor für jedes Aufgabentripel, also $k = 3$, nahe.

Ein (naives) konfirmatorisches Faktorenanalysemodell für diesen Datensatz hat also die Form

$$\upsilon = L\xi + \varepsilon \Leftrightarrow \begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \upsilon_4 \\ \upsilon_5 \\ \upsilon_6 \\ \upsilon_7 \\ \upsilon_8 \\ \upsilon_9 \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \\ l_{41} & l_{42} & l_{43} \\ l_{51} & l_{52} & l_{53} \\ l_{61} & l_{62} & l_{63} \\ l_{71} & l_{72} & l_{73} \\ l_{81} & l_{82} & l_{83} \\ l_{91} & l_{92} & l_{93} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix} \tag{12}$$

mit

$$\xi \sim N(0_3, \Phi) \text{ und } \varepsilon \sim N(0_9, \Psi) \tag{13}$$

und

$$\Phi := \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \text{ und } \Psi := \text{diag}\left(\psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6, \psi_7, \psi_8, \psi_9\right). \tag{14}$$

Modellidentifizierbarkeit und Modellrestriktionen

Die Frage nach der Identifizierbarkeit eines Modells stellt ist die Frage, ob basierend auf beobachtbaren Daten wahre, aber unbekannten, Parameterwerte eines Modells prinzipiell eindeutig geschätzt werden können. Wenn unterschiedliche wahre, aber unbekannte, Parameterwerte in den gleichen Datenverteilungen resultieren, dann sind sie und das Modell nicht identifizierbar.

Für die konfirmatorische Faktorenanalyse sind allgemeine hinreichende und notwendige Bedingungen für die Modellidentifizierbarkeit nicht vollumfänglich bekannt, die Modellidentifizierbarkeit im Bereich der Faktorenanalyse und dem eng verwandten Feld der Strukturgleichungsmodelle bleibt also in aktives Forschungsfeld. In der Anwendungspraxis sind deshalb simulationsbasierte Parameterrecoverystudien sicherlich eine gute Forschungspraxis.

Im Folgenden definieren wir zunächst die Identifizierbarkeit eines Faktorenanalysemodells und führen dann mit der *Ordnungsbedingung der konfirmatorischen Faktorenanalyse* eine häufig verwendete Heuristik zur Modellidentifizierbarkeit ein, die die Intuition Intuitiv formalisiert, dass ein Modell im Sinne der datenanalytischen Datenreduktion nicht mehr Parameter haben sollte als Datenstatistiken betrachtet werden. Dazu zählen wir dabei zunächst die Anzahl der unikalen Parameter und der Statistiken eines konfirmatorischen Faktorenanalysemodells.

Definition (Identifizierbares Faktorenanalysemodell)

Es sei

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

ein konfirmatorisches Faktorenanalysemodell. Weiterhin sei der Parametervektor dieses Modells definiert als die spaltenweise Konkatenierung der Parametermatrizen L, Φ, Ψ, also

$$\theta := \text{vec}(L, \Phi, \Psi), \tag{16}$$

so dass die marginale Datenkovarianz geschrieben werden kann als

$$\Sigma_\theta := L\Phi L^T + \Psi. \tag{17}$$

Dann heißen das Faktorenanalysemodell und der Parametervektor θ *identifizierbar*, wenn für alle $\theta_1, \theta_2 \in \Theta$ gilt, dass

$$\Sigma_{\theta_1} = \Sigma_{\theta_2} \Leftrightarrow \theta_1 = \theta_2. \tag{18}$$

Wenn für $\theta_1 \neq \theta_2$ gilt, dass $\Sigma_{\theta_1} = \Sigma_{\theta_2}$, so heißen das Modell und θ *nicht identifizierbar*.

Bemerkungen

- Bei nicht identifizierbaren Modellen ergeben unterschiedliche Parameterwerte die gleiche Datenverteilung.
- Bei nicht identifizierbaren Modellen kann aus Daten nicht eindeutig auf Parameterwerte geschlossen werden.
- Es gibt bis dato keine allgemein gültigen notwendigen und hinreichenden Bedingungen für CFA Identifizierbarkeit.
- Die Ordnungsbedingung ist eine notwendige Bedingung für die CFA Identifizierbarkeit.

Theorem (Anzahl unikaler skalarer Parameter und Statistiken der CFA)

Es sei

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi) \tag{19}$$

ein konfirmatorisches Faktorenanalysemodell. Dann gilt für die Anzahl an unikalen skalaren Parametern des Modells

$$n_\theta = mk + \frac{k(k+1)}{2} + m \tag{20}$$

Weiterhin sei $Y = (y_1, ..., y_n)$ ein Datensatz von n unabhängigen Beobachtungen eines konfirmatorischen Faktorenanalysemodells und C sei die Stichprobenkovarianzmatrix dieses Datensatzes. Dann gilt für die Anzahl an unikalen skalaren Statistiken des Modells

$$n_c = \frac{m(m+1)}{2}. \tag{21}$$

Bemerkungen

- Weil Kovarianzmatrizen symmetrisch sind, sind nur die Einträge über und inklusive ihrer Hauptdiagonale unikal.
- n_θ ist die Dimension des unikalen Parametervektors θ, es gilt also $\theta \in \mathbb{R}^{n_\theta}$.
- n_c ist die Anzahl der unikalen skalaren Einträge von C.

Beweis

Die Anzahl der Einträge der Faktorladungsmatrix $L \in \mathbb{R}^{m \times k}$ ist mk.

Die Anzahl der Einträge einer symmetrischen Matrix $S \in \mathbb{R}^{k \times k}$ ist k^2. Aufgrund der Symmetrie von S sind dabei allerdings nur die k Einträge der Hauptdiagonale und die Einträge oberhalb (oder unterhalb) der Hauptdiagonalen unikal. Die Anzahl der Einträge oberhalb (oder unterhalb) der Hauptdiagonalen ist die Hälfte aller $k^2 - k$ nicht-diagonalen Einträge von S, also $(k^2 - k)/2$. Zusammen mit den Einträgen auf der Hauptdiagonalen ergibt sich damit für die Anzahl unikaler Einträge einer symmetrischen Matrix

$$\frac{k^2 - k}{2} + k = \frac{k^2 - k}{2} + \frac{2k}{2} = \frac{k^2 + k}{2} = \frac{k(k+1)}{2}. \tag{22}$$

Als Kovarianzmatrix ist $\Phi \in \mathbb{R}^{k \times k}$ symmetrisch und hat damit $\frac{k(k+1)}{2}$ unikale Einträge. Die Anzahl der von Null verschiedenden Einträge der Beobachtungsrauschematrix $\Psi \in \mathbb{R}^{m \times m}$ ist m.

Für die Anzahl an unikalen skalaren Parametern des Faktorenanalysemodells ergibt sich also zusammenfassend

$$n_\theta = mk + \frac{k(k+1)}{2} + m. \tag{23}$$

Die Stichprobenkovarianzmatrix $C \in \mathbb{R}^{m \times m}$ eines Datensatzes ist symmetrisch. Mit obigen Überlegungen zu den unikalen Einträgen einer symmetrischen Matrix ergibt sich also direkt

$$n_c = \frac{m(m+1)}{2}. \tag{24}$$

□

Definition (Ordnungsbedingung)

Gegeben sei ein konfimatorisches Faktorenanalysemodell

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

Dann sagt man, dass das Modell der *Ordnungsbedingung* genügt, wenn für die Anzahl n_θ der unikalen skalaren Parameter und für die Anzahl n_c der unikalen skalaren Statistiken gilt, dass

$$n_\theta \leq n_c. \tag{26}$$

Bemerkung

- Die Ordnungsbedingung besagt, dass die Anzahl der unbekannten und damit zu schätzenden Parameter des Modells kleiner oder gleich der unikalen Einträge der Stichprobenkovarianzmatrix ist. Softwarelösungen wie lavaan implementieren die Ordnungsbedingung meist per default.

Anwendungsbeispiel

Für das Anwendungsbeispiel nach Holzinger and Swineford (1939) gilt $m = 9$, dass $n_c = 9(9+1)/2 = 45$. Das restringierte Modell nach Rosseel (2012) hat die Form

$$v = L\xi + \varepsilon \Leftrightarrow \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 0 & 0 \\ l_{31} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & l_{52} & 0 \\ 0 & l_{62} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & l_{83} \\ 0 & 0 & l_{93} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix} \tag{27}$$

mit

$$\xi \sim N(0_3, \Phi) \text{ und } \varepsilon \sim N(0_9, \Psi) \tag{28}$$

und

$$\Phi := \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \text{ und } \Psi := \text{diag}\left(\psi_1, \psi_2, \psi_3, \psi_4, \psi_5, \psi_6, \psi_7, \psi_8, \psi_9\right). \tag{29}$$

Für die Anzahl der als unbekannt voraussgesetzten Parameter in diesem Modell ergibt sich also

$$n_\theta = 6 + 6 + 9 = 21 < 45 = n_c \tag{30}$$

und die Ordnungsbedingung ist erfüllt.

Anwendungsbeispiel

Spezifikation des restringierten Modells nach Rosseel (2012) für den Datensatz nach Holzinger and Swineford (1939)

```
library(lavaan)                                  # Lavaan SEM Paket
data(HolzingerSwineford1939)                     # Datensatz
YT           = HolzingerSwineford1939[,7:15]     # transponierte Datenmatrix
colnames(YT) = paste("y", 1:9, sep = "")         # vorlesungskonsistente Variablennamen
rownames(YT) = paste("i =", 1:nrow(YT))          # vorlesungskonsistente Variablennamen
cfa_mod      = 'x_1 =~  y1 + y2 + y3             # Faktor x_1 mit Nicht-null Ladungen für y_1,y_2,y_3
                x_2 =~  y4 + y5 + y6             # Faktor x_2 mit Nicht-null Ladungen für y_4,y_5,y_6
                x_3 =~  y7 + y8 + y9'            # Faktor x_3 mit Nicht-null Ladungen für y_7,y_8,y_9
cfa_mod      = cfa(cfa_mod, data = YT)           # CFA Modellformulierung
theta        = lavInspect(cfa_mod, what = "free") # Parameterinspektion
L            = theta$lambda                      # Faktorladungsmatrix
Phi          = theta$psi                         # Beobachtungsrauschenmatrix
Psi          = theta$theta                       # Faktorrauschenmatrix
```

Anwendungsbeispiel

Einträge ungleich 0 repräsentieren als unbekannt angenommene und damit zu schätzende Parameter

```
> L =

>    x_1 x_2 x_3
> y1   0   0   0
> y2   1   0   0
> y3   2   0   0
> y4   0   0   0
> y5   0   3   0
> y6   0   4   0
> y7   0   0   0
> y8   0   0   5
> y9   0   0   6

> Psi =

>    y1 y2 y3 y4 y5 y6 y7 y8 y9
> y1  7
> y2  0  8
> y3  0  0  9
> y4  0  0  0 10
> y5  0  0  0  0 11
> y6  0  0  0  0  0 12
> y7  0  0  0  0  0  0 13
> y8  0  0  0  0  0  0  0 14
> y9  0  0  0  0  0  0  0  0 15

> Phi =

>     x_1 x_2 x_3
> x_1 16
> x_2 19  17
> x_3 20  21  18
```

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Überblick

Traditionell werden die Parameter der konfirmatorischen Faktorenanalyse durch Minimierung der sogenannten *Diskrepanzfunktion*, geschätzt, vgl. Lawley (1940), K. G. Jöreskog (1967) und K. G. Jöreskog (1969). Die funktionale Form der Diskrepanzfunktion ist dabei durch ein Log-Likelihood-Kriterium bei Betrachtung der Frequentistischen Verteilung der Stichprobenkovarianz motiviert. Diese ist nach Wishart (1928) benannt.

Neuere Sichtweisen im Kontext von Strukturgleichungsmodellen motivieren die funktionale Form der Diskrepanzfunktion allerdings direkt durch ein Log-Likelihood-Kriterium bei Betrachtung der multivariaten Datennormalverteilung, vgl. z.B. Bollen (1989) und Rosseel (2021). Wir wollen hier diesen neueren Weg nachzeichnen und somit auch auf eine Einführung der Wishart-Verteilung verzichten. Zu diesem Zweck nehmen wir durchgängig die *Zentrierung* des betrachteten Datensatzes $Y \in \mathbb{R}^{m \times n}$, also $\bar{y} = 0_m$, sowie die Identifizierbarkeit des Modells an.

Für einen gänzlich alternativen modernen Zugang zur Schätzung des konfirmatorischen Faktorenanalysemodells mithilfe des Expectation-Maximization Algorithmus im Rahmen der variationalen Inferenz, siehe z.B. Rubin and Thayer (1982), Roweis and Ghahramani (1999) und (5) Faktorenanalyse aus der letztjährigen Iteration des Kurses. Die genauen Bezüge zwischen den traditionellen und modernen Schätzverfahren für die konfirmatorische Faktorenanalyse sind dabei eine offene Forschungsfrage.

Überblick

Zur Diskussion der Modellschätzung der konfirmatorischen Faktorenanalyse gehen wir hier also im Folgenden wie folgt vor.

(1) Wir erinnern zunächst an die Definition von Log-Likelihood-Funktion und Maximum-Likelihood-Schätzern.
(2) Wir evaluieren als nächstes die Log-Likelihood-Funktion der konfirmatorischen Faktorenanalyse.
(3) Wir definieren dann die funktionale Form der Diskrepanzfunktion.
(4) Wir zeigen schließlich, dass Minimumstellen der Diskrepanzfunktion Maximum-Likelihood-Schätzer sind.

Die Minimierung der Diskrepanzfunktion wird heutzutage mithilfe von Standardverfahren der nichtlinearen Optimierung durchgeführt (vgl. z.B. Rosseel (2012), Rosseel (2021)), welche wir hier nicht vertiefen wollen. Eine Einführung gibt z.B. (6) Optimierung aus der letztjährigen Iteration des Kurses. Die Motivation der funktionalen Form der Diskrepanzfunktion selbst ergibt sich im Kontext der Modellevaluation.

Definition (Log-Likelihood-Funktion und Maximum-Likelihood-Schätzer)

Gegeben sei ein Datensatz $Y := (y_1, ..., y_n) \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen eines Zufallsvektors υ mit Ergebnisraum $\mathcal{Y}$ und parameterabhängiger WDF p_θ mit $\theta \in \Theta$. Dann ist die *Log-Likelihood-Funktion von Y unter p_θ* definiert als

$$\ell_Y : \Theta \to \mathbb{R}, \theta \mapsto \ell_Y(\theta) := \ln \prod_{i=1}^{n} p_\theta(y_i) \tag{31}$$

und ein *Maximum-Likelihood-Schätzer von θ* ist definiert als

$$\hat{\theta}_{\text{ML}} : \mathcal{Y}^n \to \Theta, Y \mapsto \hat{\theta}_{\text{ML}} := \arg\max_{\theta \in \Theta} \ell_Y(\theta). \tag{32}$$

Bemerkungen

- Eine Log-Likelihood-Funktion ist für einen festen Datensatz eine Funktion der Parameter eines Modells.
- Der Funktionswert einer Log-Likelihood-Funktion entspricht der logarithmierten Wahrscheinlichkeitsdichte des Datensatzes unter dem Modell für einen speziellen Parameterwert.
- Ein Maximum-Likelihood-Schätzer maximiert die Log-Likelihood Funktion.
- Eine Einführung zur Generation von Maximum-Likelihood-Schätzern gibt die Einheit (10) Parameterschätzung aus der Vorlesung Wahrscheinlichkeitstheorie und Frequentistische Inferenz des BSc Psychologie.

Theorem (Log-Likelihood-Funktion der konfirmatorischen Faktorenanalyse)

Gegeben sei ein konfimatorisches Faktorenanalysemodell

$$\upsilon = L\xi + \varepsilon \text{ mit } \xi \sim N(0_k, \Phi) \text{ und } \varepsilon \sim N(0_m, \Psi) \tag{33}$$

mit Parametervektor θ und marginalen Kovarianzmatrixparameter Σ_θ. Weiterhin sei $Y := (y_1, \ldots, y_n) \in \mathbb{R}^{m \times n}$ ein zentrierter Datensatz von n unabhängigen Beobachtungen von υ, C seine Stichprobenkovarianzmatrix und

$$S := \frac{n-1}{n} C \tag{34}$$

seine verzerrte Stichprobenkovarianzmatrix. Dann kann die Log-Likelihood-Funktion von Y geschrieben werden als

$$\ell_Y : \Theta \to \mathbb{R}, \theta \mapsto \ell_Y(\theta) := -\frac{n}{2} \ln |\Sigma_\theta| - \frac{n}{2} \text{tr}\left(S\Sigma_\theta^{-1}\right) - \frac{nm}{2} \ln(2\pi) \tag{35}$$

wobei $\text{tr}(\cdot)$ die *Spur*, also die Summe der Diagonalelemente, einer quadratischen Matrix bezeichnet. Eine Maximumstelle von ℓ_Y, also ein Wert $\hat{\theta}_{\text{ML}} \in \Theta$ mit

$$\hat{\theta}_{\text{ML}} = \underset{\theta \in \Theta}{\arg\max}\, \ell_Y(\theta), \tag{36}$$

heißt *Maximum-Likelihood-Schätzer der konfirmatorischen Faktorenanalyse*.

Beweis

Mit der Definition der Log-Likelihood-Funktion als und der marginalen WDF des Faktorenanalysemodells ergibt sich mit den Rechenregeln des Logarithmus sofort

$$\begin{aligned}
\ell_Y(\theta) &= \ln \prod_{i=1}^{n} p_\theta(y_i) \\
&= \prod_{i=1}^{n} \ln N\left(y_i; 0_m, L\Phi L^T + \Psi\right) \\
&= \ln \left(\prod_{i=1}^{n} (2\pi)^{-m/2} |\Sigma_\theta|^{-1/2} \exp\left(-\frac{1}{2} (y_i - 0_m)^T \Sigma_\theta^{-1} (y_i - 0_m)) \right) \right) \\
&= -\frac{mn}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma_\theta| - \frac{1}{2} \sum_{i=1}^{n} y_i^T \Sigma_\theta^{-1} y_i .
\end{aligned} \tag{37}$$

Um die Gleicheit des letzten Terms auf der rechten Seite mit dem letzten Term in der postulierten Funktionsform zu zeigen, halten wir zunächst fest, dass mit elementaren Eigenschaften der Matrixspur gilt, dass

$$\mathsf{tr}\left(\sum_{i=1}^{n} y_i^T \Sigma_\theta^{-1} y_i \right) = \sum_{i=1}^{n} \mathsf{tr}\left(y_i^T \Sigma_\theta^{-1} y_i \right) = \sum_{i=1}^{n} \mathsf{tr}\left(\Sigma_\theta^{-1} y_i y_i^T \right) = \mathsf{tr}\left(\Sigma_\theta^{-1} \sum_{i=1}^{n} y_i y_i^T \right) . \tag{38}$$

Beweis (fortgeführt)

Weiterhin gilt mit dem Binomischen Lehrsatz

$$\begin{aligned}
\sum_{i=1}^{n} y_i^T \Sigma_\theta^{-1} y_i &= \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y} + \bar{y}) \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^{n} \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma_\theta^{-1} \bar{y} \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^{n} \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 \left(\sum_{i=1}^{n} \left(y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \right)^T \right) \Sigma_\theta^{-1} \bar{y} \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^{n} \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 \left(\sum_{i=1}^{n} y_i^T - \frac{n}{n} \sum_{i=1}^{n} y_i^T \right) \Sigma_\theta^{-1} \bar{y} \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^{n} \bar{y}^T \Sigma_\theta^{-1} \bar{y} + 2 \left(0_m^T \Sigma_\theta^{-1} \bar{y} \right) \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^{n} \bar{y}^T \Sigma_\theta^{-1} \bar{y}
\end{aligned}$$

Beweis (fortgeführt)

Also folgt mit obiger Eigenschaft der Matrixspur, der Zentrierung des Datensatzes $\bar{y} = 0_m$ und der Definition von S

$$\begin{aligned}
\sum_{i=1}^{n} y_i^T \Sigma_\theta^{-1} y_i &= \sum_{i=1}^{n} (y_i - \bar{y})^T \Sigma_\theta^{-1} (y_i - \bar{y}) + \sum_{i=1}^{n} \bar{y}^T \Sigma_\theta^{-1} \bar{y} \\
&= \sum_{i=1}^{n} y_i^T \Sigma_\theta^{-1} y_i \\
&= \mathsf{tr} \left(\sum_{i=1}^{n} y_i^T \Sigma_\theta^{-1} y_i \right) \\
&= \mathsf{tr} \left(\Sigma_\theta^{-1} \sum_{i=1}^{n} y_i y_i^T \right) \\
&= \mathsf{tr} \left(\Sigma_\theta^{-1} n S \right) \\
&= n \, \mathsf{tr} \left(S \Sigma_\theta^{-1} \right)
\end{aligned} \tag{39}$$

Substitution in $\ell_Y(\theta)$ von oben ergibt dann die postulierte funktionale Form der Log-Likelihood Funktion.

Definition (CFA Diskrepanzfunktion und CFA Parameterschätzer)

Gegeben sei ein konfirmatorisches Faktorenanalysemodell

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

mit Parametervektor θ und marginalen Kovarianzmatrixparameter Σ_θ, respektive. Weiterhin sei für einen Datensatz $Y \in \mathbb{R}^{m \times n}$ von n unabhängigen Beobachtungen von υ S die verzerrte Stichprobenkovarianzmatrix von Y. Dann heißt die Funktion

$$F_Y : \Theta \to \mathbb{R}, \theta \mapsto F_Y(\theta) := n \ln |\Sigma_\theta| + n\text{tr}\left(S\Sigma_\theta^{-1}\right) - n \ln |S| - nm \tag{41}$$

die *CFA Diskrepanzfunktion*. Weiterhin heißt ein Wert $\hat{\theta} \in \Theta$ mit

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min}\, F_Y(\theta), \tag{42}$$

also eine Minimumstelle von F_Y, ein *CFA Parameterschätzer*.

Bemerkungen

- Die Log-Likelihood-Funktion und die Diskrepanzfunktion sind nicht identisch.
- Wir zeigen jedoch unten, dass Minimumstellen von F_Y Maximumstellen von ℓ_Y sind.
- Minimierung der Diskrepanzfunktion liefert für die CFA also Maximum Likelihood Schätzer.
- Die Motivation der Diskrepanzfunktion erschließt sich vor dem Hintergrund der Modellevaluation.

Theorem (CFA Maximum-Likelihood-Schätzer)

Eine Minimumstelle $\hat{\theta}$ der CFA Diskrepanzfunktion F_Y maximiert die Log-Likelihood-Funktion ℓ_Y der konfirmatorischen Faktorenanalyse, ein CFA Parameterschätzer ist also ein Maximum-Likelihood-Schätzer $\hat{\theta}_{\mathsf{ML}}$.

Beweis

Wir halten zunächst fest, dass mit von θ unabhängigen Konstanten $a, b \in \mathbb{R}$ gilt, dass

$$\ell_Y(\theta) = a\left(-F_Y(\theta)\right) + b. \tag{43}$$

Die Log-Likelihood-Funktion ist also eine linear-affine und damit insbesondere monotone Transformation von F_Y. Da monotone Transformation Extremstellen unverändert lassen und ein negatives Vorzeichen eine Minimumstelle in eine Maximumstelle transformiert, ergibt sich das Theorem direkt.

Minimierung der Diskrepanzfunktion

Minima von F_Y werden in populären Analyseprogrammen zur konfirmatorischen Faktorenanalyse mit iterativen Standardverfahren der nichtlinearen Optimierung bestimmt. Allgemein gehen diese Verfahren von einem Startwert $\hat{\theta}^{(0)}$ aus und evaluieren weitere Iteranden rekursiv durch

$$\hat{\theta}^{(i+1)} = f\left(\hat{\theta}^{(i)}, Y\right) \text{ für } i = 0, 1, \ldots \tag{44}$$

mit einer entsprechend gewählten Funktion f des vorherigen Iteranden $\hat{\theta}^{(i)}$ und des Datensatzes Y solange, bis ein entsprechend gewähltes Abbruchkriterium erfüllt ist. Einen Überblick über die zum Beispiel im populären CFA Analyseprogramm lavaan implementierten Optimierungsalgorithmen geben Rosseel (2012) und Rosseel (2021). Exzellente Einführungen in die Theorie der nichtlinearen Optimierung geben zum Beispiel Nocedal and Wright (2006) und Kochenderfer and Wheeler (2019).

Wir wollen die numerische Minimierung der Diskrepanzfunktion hier nicht weiter vertiefen und betrachten stattdessen ein Simulationsbeispiel. In diesem Beispiel geben wir für ein konfirmatorisches Faktorenanalysemodel mit $m = 3$ und $k = 1$ wahre, aber unbekannte, Parameterwerte des Modells vor und generieren durch Samplen der multivariaten Normalverteilung einen Datensatz. Wir betrachten dann den Schätzfehler, also die Abweichung zwischen geschätzten und wahrem Parameterwert basierend auf der lavaan Parameterschätzung für steigenden Stichprobenumfang. Wir bestimmen den Schätzfehler dabei als die Euklidische Distanz zwischen wahren, aber unbekanntem, und geschätztem Parameterwert.

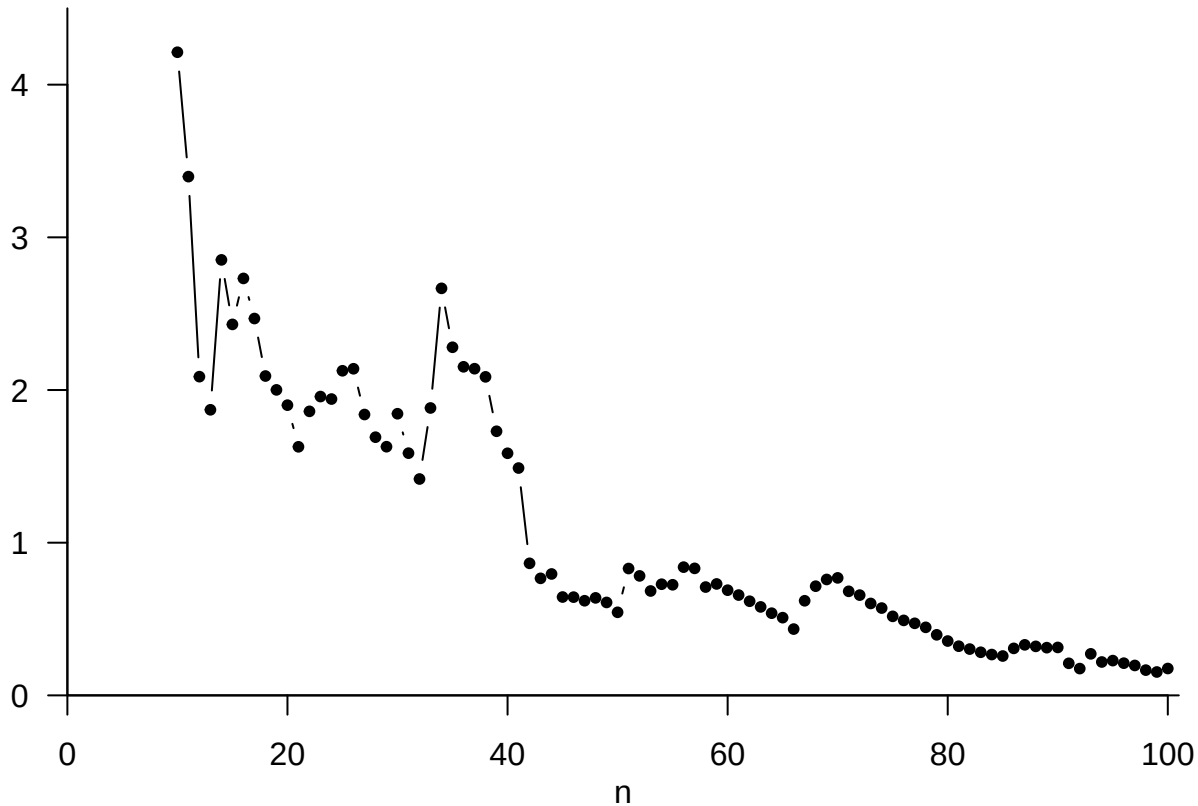
Simulationsbeispiel

```r
# Modellformulierung
set.seed(2)
k      = 1                                          # Dimension des latenten Faktorenzufallsvektors
m      = 3                                          # Dimension des beobachtbaren Zufallsvektors
L      = matrix(c(1,2,3), nrow = m)                 # Faktorladungsmatrix
Phi    = 2                                          # Faktorkovarianzmatrix
Psi    = diag(c(1,2,3))                             # Beobachtungsrauschenkovarianmatrix
theta  = c(as.vector(L),diag(Psi),as.vector(Phi))   # wahrer, aber unbekannter, Parametervektor \theta
p      = length(theta)                              # Anzahl Parameter

# Modellrealisierungen
library(MASS)                                       # Normalverteilungspaket
n      = 100                                        # Beobachtungsanzahl
Y      = matrix(rep(NaN,m*n), nrow = m)             # Simulierte beobachtete Datenmatrix
for(i in 1:n){
  x      = mvrnorm(1,rep(0,k),Phi)                  # 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
}

# Datenformatierung für lavaan
Y            = as.data.frame(t(Y))
colnames(Y)  = c(paste("y", 1:m, sep = ""))         # Indikatorvariablennamen
rownames(Y)  = 1:n                                  # Beobachtungslabels

# Modellschätzung
library(lavaan)                                     # lavaan Paket
na           = seq(1e1,n,1e0)                       # analysierte Stichprobenumfänge
ns           = length(na)                           # Stichprobensampleanzahl
cfa_mod      = 'x1 =~ y1 + y2 + y3'                 # CFA Model Spezifikation
cfa_est      = matrix(rep(NaN,ns*p), nrow = p)      # CFA Parameterschätzer
cfa_err      = rep(NaN,ns)                          # CFA Schätzfehler
for(i in 1:length(na)){
  cfa_fit     = cfa(cfa_mod, data = Y[1:na[i],])    # CFA Modellschätzung
  cfa_sta     = parameterestimates(cfa_fit)         # CFA Parameterschätzer
  cfa_est[,i] = cfa_sta$est                         # CFA Parameterschätzer
  cfa_err[i]  = norm(cfa_est[,i]-theta, type = "2") # CFA Schätzfehler
}
```

Simulationsbeispiel

Abnahme des Schätzfehlers als Funktion des Stichprobenumfangs

$$||\theta - \hat{\theta}||$$



Anwendungsbeispiel

Schätzung des restringierten Modells nach Rosseel (2012) für den Holzinger and Swineford (1939) Datensatz

```
library(lavaan)                                  # Lavaan SEM Paket
data(HolzingerSwineford1939)                     # Datensatz
YT           = HolzingerSwineford1939[,7:15]     # transponierte Datenmatrix
colnames(YT) = paste("y", 1:9, sep = "")         # vorlesungskonsistente Variablennamen
rownames(YT) = paste("i =", 1:nrow(YT))          # vorlesungskonsistente Variablennamen
cfa_mod      = 'x_1 =~  y1 + y2 + y3             #  Faktor x_1 mit Nicht-null Ladungen für y_1,y_2,y_3
                x_2 =~  y4 + y5 + y6             # Faktor x_2 mit Nicht-null Ladungen für y_4,y_5,y_6
                x_3 =~  y7 + y8 + y9'            # Faktor x_3 mit Nicht-null Ladungen für y_7,y_8,y_9
cfa_mod      = cfa(cfa_mod, data = YT)           # CFA Modellformulierung und -schätzung
theta_hat    = lavInspect(cfa_mod, what = "est") # Inspektion der geschätzten Parameter
L_hat        = theta_hat$lambda                  # Geschätzte Faktorladungsmatrix
Phi_hat      = theta_hat$psi                     # Geschätzte Beobachtungsrauschenmatrix
Psi_hat      = theta_hat$theta                   # Geschätzte Faktorrauschenmatrix
```

Anwendungsbeispiel

Geschätzte Parameterwerte

```
> L_hat =

>      x_1   x_2  x_3
> y1 1.000 0.000 0.00
> y2 0.554 0.000 0.00
> y3 0.729 0.000 0.00
> y4 0.000 1.000 0.00
> y5 0.000 1.113 0.00
> y6 0.000 0.926 0.00
> y7 0.000 0.000 1.00
> y8 0.000 0.000 1.18
> y9 0.000 0.000 1.08

> Psi_hat =

>    y1    y2    y3    y4    y5    y6    y7    y8    y9
> y1 0.549
> y2 0.000 1.134
> y3 0.000 0.000 0.844
> y4 0.000 0.000 0.000 0.371
> y5 0.000 0.000 0.000 0.000 0.446
> y6 0.000 0.000 0.000 0.000 0.000 0.356
> y7 0.000 0.000 0.000 0.000 0.000 0.000 0.799
> y8 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.488
> y9 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.566

> Phi_hat =

>     x_1   x_2   x_3
> x_1 0.809
> x_2 0.408 0.979
> x_3 0.262 0.173 0.384
```

Anwendungsbeispiel

Das geschätzte restringierte Modell für den Holzinger and Swineford (1939) Datensatz hat also die Form

$$\upsilon = \hat{L}\xi + \varepsilon \Leftrightarrow \begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \upsilon_3 \\ \upsilon_4 \\ \upsilon_5 \\ \upsilon_6 \\ \upsilon_7 \\ \upsilon_8 \\ \upsilon_9 \end{pmatrix} = \begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.55 & 0.00 & 0.00 \\ 0.73 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 1.11 & 0.00 \\ 0.00 & 0.92 & 0.00 \\ 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 1.18 \\ 0.00 & 0.00 & 1.08 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix} \tag{45}$$

mit

$$\xi \sim N(0_3, \hat{\Phi}) \text{ und } \varepsilon \sim N(0_9, \hat{\Psi}) \tag{46}$$

und

$$\hat{\Phi} := \begin{pmatrix} 0.81 & 0.41 & 0.26 \\ 0.41 & 0.97 & 0.17 \\ 0.26 & 0.17 & 0.38 \end{pmatrix} \text{ und } \hat{\Psi} := \text{diag}\left(0.55, 1.13, 0.84, 0.37, 0.45, 0.36, 0.80, 0.49, 0.57\right). \tag{47}$$

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Überblick

$Y := (y_1, ..., y_n) \in \mathbb{R}^{m \times n}$ sei ein Datensatz von n unabhängigen Beobachtungen von υ mit $\bar{y} = 0_m$ und uvec$(A, B, ...)$ sei die konkatenisierte Vektorisierung der unikalen Werte der Matrizen $A, B, ...$ sowie uvec$^{-1}(A, B, ...)$ ihre Umkehrung.

Häufig möchte man bei der konfirmatorischen Faktorenanalyse basierend auf Y zwei Modelle vergleichen.

(M1) Ein konfirmatorisches Faktorenanalysemodell, das die Ordnungsrelation erfüllt und identifizierbar ist,

$$\upsilon \sim N(0, \Sigma_\theta) \text{ mit } \Sigma_\theta = L\Phi L^T + \Psi, \theta = \text{uvec}(L, \Phi, \Psi) \in \Theta, \Theta \subset \mathbb{R}^p \text{ und } p \leq m(m+1)/2. \tag{48}$$

(M2) Ein multivariates Normalverteilungsmodell mit beliebigem Kovarianzmatrixparameter,

$$\upsilon \sim N(0, \Sigma_\gamma) \text{ mit } \Sigma_\gamma = \text{uvec}^{-1}(\gamma), \gamma \in \Gamma \text{ und } \Gamma \subset \mathbb{R}^{m(m+1)/2}. \tag{49}$$

Ein häufig genutztes Kriterium für diesen Modellvergleich ist das Log-Likelihood-Ratio-Kriterium

$$\Lambda_Y := \ln\left(\frac{\max_{\theta\in\Theta}\prod_{i=1}^n N(y_i; 0_m, \Sigma_\theta)}{\max_{\gamma\in\Gamma}\prod_{i=1}^n N(y_i; 0_m, \Sigma_\gamma)}\right) \tag{50}$$

Das $\Lambda_Y \in \mathbb{R}$ setzt die maximierten Wahrscheinlichkeitsdichten von Y unter (M1) und (M2) ins Verhältnis. Große Werte von Λ_Y bedeuten, dass Y unter (M1) eine größere Wahrscheinlichkeit(sdichte) besitzt als unter (M2). Dies wird allgemein als Evidenz dafür verstanden, dass Y eher von (M1) als von (M2) generiert wurden. Dabei ist Λ_Y letztlich die zentrale Motivation für die funktionale Form der Diskrepanzfunktion (cf. Lawley (1940)).

Theorem (Diskrepanzfunktion)

Für einen Datensatz $Y := (y_1, ..., y_n) \in \mathbb{R}^{m \times n}$ mit $\bar{y} = 0_m$ von n unabhängigen Beobachtungen eines Zufallsvektors υ sei das *Log-Likelihood-Ratio-Kriterium der konfirmatorischen Faktorenanalyse* gegeben durch

$$\Lambda_Y := \ln \left(\frac{\max_{\theta \in \Theta} \prod_{i=1}^{n} N(y_i; 0_m, \Sigma_\theta)}{\max_{\gamma \in \Gamma} \prod_{i=1}^{n} N(y_i; 0_m, \Sigma_\gamma)} \right), \tag{51}$$

wobei

$$\Sigma_\theta = L \Phi L^T + \Psi, \theta = \text{uvec}(L, \Phi, \Psi) \in \Theta, \Theta \subset \mathbb{R}^p \text{ und } p \leq m(m+1)/2 \tag{52}$$

und

$$\Sigma_\gamma = \text{uvec}^{-1}(\gamma), \gamma \in \Gamma \text{ und } \Gamma \subset \mathbb{R}^{m(m+1)/2} \tag{53}$$

seien. Weiterhin sei für die verzerrte Stichprobenkovarianzmatrix S von Y

$$F_Y(\theta) := n \ln |\Sigma_\theta| + n \text{tr} \left(S \Sigma_\theta^{-1} \right) - n \ln |S| - nm \tag{54}$$

die Diskrepanzfunktion der CFA und $\hat{\theta}$ eine Minimumstelle von F_Y. Dann gilt

$$-2\Lambda_Y = F_Y(\hat{\theta}) \tag{55}$$

Beweis

Wir halten zunächst fest, dass

$$\max_{\theta \in \Theta} \prod_{i=1}^{n} N(y_i; 0_m, \Sigma_\theta) = \prod_{i=1}^{n} N\left(y_i; 0_m, \Sigma_{\hat{\theta}}\right) \tag{56}$$

weil eine Minimumstelle $\hat{\theta}$ von F_Y wie oben gesehen die Log-Likelihood-Funktion und damit auch die Likelihood-Funktion der exploratorischen Faktorenanalyse maximiert. Weiterhin halten wir fest, dass

$$\max_{\gamma \in \Gamma} \prod_{i=1}^{n} N(y_i; 0_m, \Sigma_\gamma) = \prod_{i=1}^{n} N(y_i; 0_m, S) \tag{57}$$

weil die verzerrte Stichprobenkovarianz der Maximum-Likelihood-Schätzer des Kovarianzmatrixparameters einer multivariaten Normalverteilung ist (cf. (6) Multivariate Normalverteilungen). Die Logarithmuseigenschaften ergeben dann

$$\Lambda_Y = \ln\left(\frac{\prod_{i=1}^{n} N\left(y_i; 0_m, \Sigma_{\hat{\theta}}\right)}{\prod_{i=1}^{n} N(y_i; 0_m, S)}\right) = \sum_{i=1}^{n} \ln N\left(y_i; 0_m, \Sigma_{\hat{\theta}}\right) - \sum_{i=1}^{n} \ln N(y_i; 0_m, S) \tag{58}$$

Beweis (fortgeführt)

Substition der funktionalen Form der Log-Likelihood-Funktion der konfirmatorischen Faktorenanalyse und der funktionalen Form der WDF der multivariaten Normalverteilung ergibt

$$\begin{aligned}
\Lambda_Y &= \sum_{i=1}^{n} \ln N\left(y_i; 0_m, \Sigma_{\hat{\theta}}\right) - \sum_{i=1}^{n} \ln N(y_i; 0_m, S) \\
&= -\frac{mn}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma_\theta| - \frac{n}{2}\mathsf{tr}\left(S\Sigma_\theta^{-1}\right) - \frac{n}{2}\bar{y}^T\Sigma_\theta^{-1}\bar{y} \\
&\quad + \frac{mn}{2}\ln(2\pi) + \frac{n}{2}\ln|S| + \frac{n}{2}\mathsf{tr}\left(SS^{-1}\right) + \frac{n}{2}\bar{y}^T S^{-1}\bar{y} \qquad (59) \\
&= -\frac{n}{2}\ln|\Sigma_\theta| - \frac{n}{2}\mathsf{tr}\left(S\Sigma_\theta^{-1}\right) - \frac{n}{2}0_m^T\Sigma_\theta^{-1}0_m \\
&\quad + \frac{n}{2}\ln|S| + \frac{n}{2}\mathsf{tr}\left(SS^{-1}\right) + \frac{n}{2}0_m^T S^{-1}0_m \\
&= -\frac{n}{2}\ln|\Sigma_\theta| - \frac{n}{2}\mathsf{tr}\left(S\Sigma_\theta^{-1}\right) + \frac{n}{2}\ln|S| + \frac{mn}{2}
\end{aligned}$$

Multiplikation mit -2 ergibt schließlich

$$-2\Lambda_Y = n\ln|\Sigma_\theta| + n\,\mathsf{tr}\left(S\Sigma_\theta^{-1}\right) - n\ln|S| - mn = F_Y(\theta). \qquad (60)$$

Anwendungsbeispiel

Evaluation des retringierten Modells nach Rosseel (2012) für den Holzinger and Swineford (1939) Datensatz

```
library(lavaan)                                  # Lavaan SEM Paket
data(HolzingerSwineford1939)                     # Datensatz
YT           = HolzingerSwineford1939[,7:15]     # transponierte Datenmatrix
m            = ncol(YT)                          # Datendimension
n            = nrow(YT)                          # Stichprobenumfang
colnames(YT) = paste("y", 1:9, sep = "")         # vorlesungskonsistente Variablennamen
rownames(YT) = paste("i =", 1:nrow(YT))          # vorlesungskonsistente Variablennamen
cfa_mod      = 'x_1 =~  y1 + y2 + y3             # Faktor x_1 mit Nicht-null Ladungen für y_1,y_2,y_3
                x_2 =~  y4 + y5 + y6             # Faktor x_2 mit Nicht-null Ladungen für y_4,y_5,y_6
                x_3 =~  y7 + y8 + y9'            # Faktor x_3 mit Nicht-null Ladungen für y_7,y_8,y_9
cfa_mod      = cfa(cfa_mod, data = YT)           # CFA Modellformulierung und -schätzung
theta_hat    = lavInspect(cfa_mod, what = "est") # Inspektion der geschätzten Parameter
L_hat        = theta_hat$lambda                  # Geschätzte Faktorladungsmatrix
Phi_hat      = theta_hat$psi                     # Geschätzte Beobachtungsrauschenmatrix
Psi_hat      = theta_hat$theta                   # Geschätzte Faktorrauschenmatrix
Sigma_hat    = L_hat%*%Phi_hat%*%t(L_hat)+Psi_hat # Geschätzte marginale Datenkovarianzmatrix
S            = (n-1)/n*cov(YT)                   # Verzerrte Stichprobenkovarianzmatrix
F_Y          = n*(  log(det(Sigma_hat))          # Diskrepanzfunktionsevaluation
                  + sum(diag(S%*%solve(Sigma_hat)))
                  - log(det(S))
                  - m)
```

Anwendungsbeispiel

Evaluation des retringierten Modells nach Rosseel (2012) für den Holzinger and Swineford (1939) Datensatz

```
> n             : 301
> F_Y_theta_hat : 85.3
```

Lavaan Output

```
show(cfa_mod)
```

```
> lavaan 0.6-12 ended normally after 35 iterations
>
>   Estimator                                         ML
>   Optimization method                           NLMINB
>   Number of model parameters                        21
>
>   Number of observations                           301
>
> Model Test User Model:
>
>   Test statistic                                85.306
>   Degrees of freedom                                24
>   P-value (Chi-square)                           0.000
```

Modellformulierung

Modellschätzung

Modellevaluation

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition des Modells der konfirmatorischen Faktorenanalyse (CFA) wieder.
2. Erläutern Sie das Modell der CFA.
3. Geben Sie das Theorem zur WDF der gemeinsamen Verteilung von Faktoren und Daten wieder.
4. Geben Sie das Theorem zur WDF und Eigenschaften der marginalen Datenverteilung wieder.
5. Erläutern Sie den Begriff der Modellidentifizierbarkeit.
6. Geben Sie die Definition eines Identifizierbaren Faktorenanalysemodells wieder.
7. Geben Sie die Definition der Ordnungsbedingung wieder.
8. Geben Sie die Definition der Log-Likelihood-Funktion wieder.
9. Geben Sie die Definition eines Maximum-Likelihood-Schätzers wieder.
10. Erläutern Sie den Zusammenhang zwischen der Log-Likelihood-Funktion und der Diskrepanzfunktion der CFA.
11. Geben Sie das Theorem zum CFA Maximum-Likelihood-Schätzer wieder.
12. Erläutern die Diskrepanzfunktion der CFA vor dem Hintegrund der CFA Modellevaluation.

References

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.

Holzinger, K. J., and F. Swineford. 1939. "A Study in Factor Analysis: The Stability of a Bi-Factor Solution." *Supplementary Educational Monographs* 48 (xi + 91).

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

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

Jöreskog, Karl Gustuv. 1967. "Some Contributions to Maximum Likelihood Factor Analysis."

———. 1969. "A General Approach to Confirmatory Maximum Likelihood Factor Analysis."

Kochenderfer, Mykel J., and Tim A. Wheeler. 2019. *Algorithms for Optimization*. Cambridge, Massachusetts: The MIT Press.

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*.

Nocedal, Jorge, and Stephen J. Wright. 2006. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research. New York: Springer.

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.

———. 2021. "Evaluating the Observed Log-Likelihood Function in Two-Level Structural Equation Modeling with Missing Data: From Formulas to R Code." *Psych* 3 (2): 197–232. https://doi.org/10.3390/psych3020017.

Roweis, Sam, and Zoubin Ghahramani. 1999. "A Unifying Review of Linear Gaussian Models." *Neural Computation* 11 (2): 305–45. https://doi.org/10.1162/089976699300016674.

Rubin, Donald B., and Dorothy T. Thayer. 1982. "EM Algorithms for ML Factor Analysis." *Psychometrika* 47 (1): 69–76. https://doi.org/10.1007/BF02293851.

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.