Multivariate Datenanalyse

MSc Psychologie WiSe 2021/22

Prof. Dr. Dirk Ostwald

(5) Faktorenanalyse

Überblick

Generative Perspektive zu konfirmatorischer und exploratorischer Faktoranalyse

- "Generativ" bedeutet hier probabilistisch und modell-basiert.

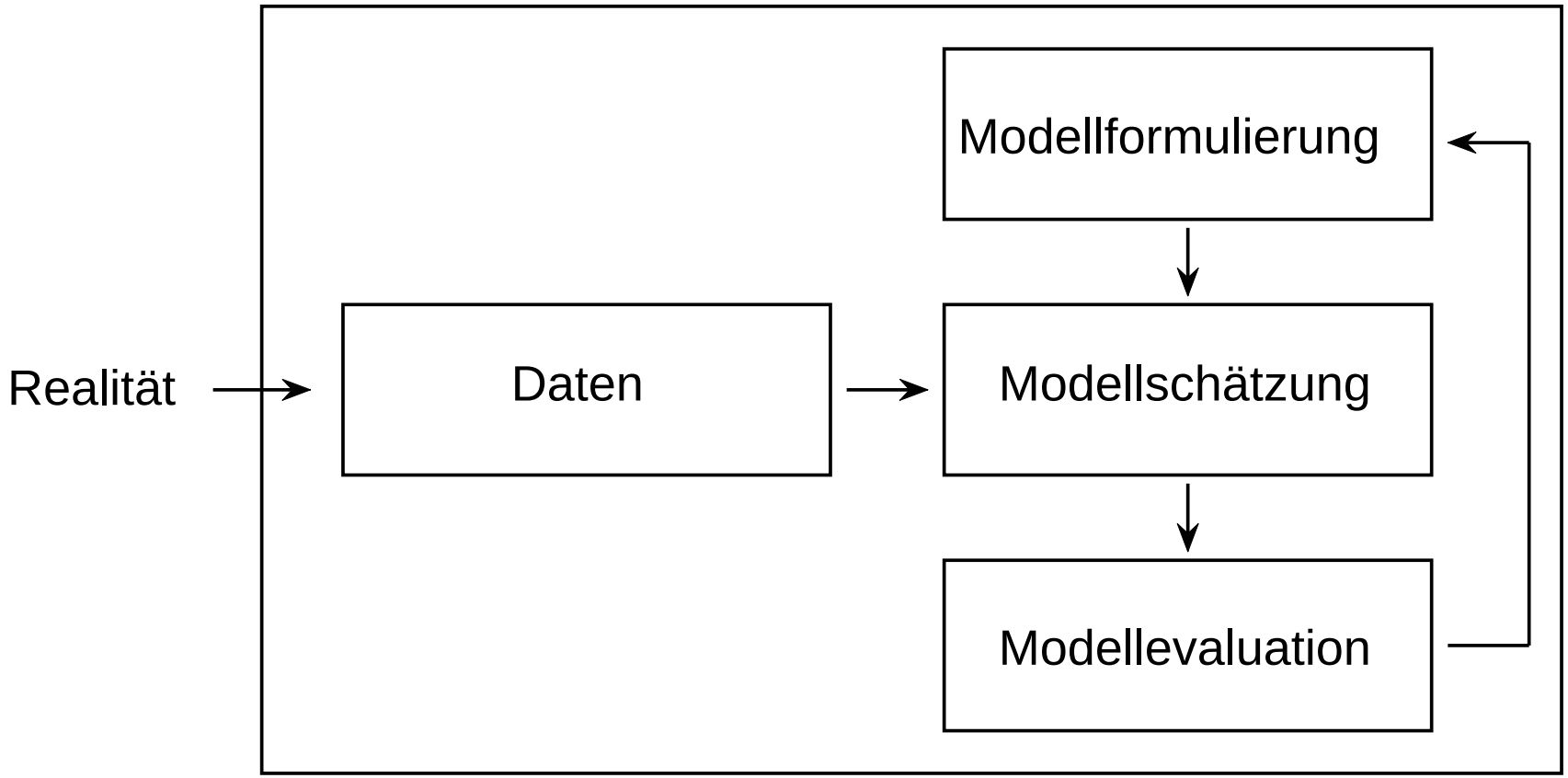
Einführung zum Expectation-Maximization (EM) Algorithmus.

- Generischer Algorithmus zur Schätzung von Parametern in Modellen mit latenten Variablen.
- Moderne Sichtweise von EM als Evidende Lower Bound Maximierung.
- Ein Schritt in Richtung eines Verständnisses von Variational Inference.

$\Rightarrow$ Integrative Perspektive von Inferenz und Lernen in Modellen mit latenten Variablen mit einer natürlichen Generalisierung zu kontemporären Modellen des maschinellen Lernens und der künstlicher Intelligenz (Variational Bayesian Filtering oder "Variational Autoencoders" für die Generation Deep Learning)

$\Rightarrow$ Ein Schritt zu einem Verständnis zeitgenössischer Theorien zur Funktionsweise des Gehirns (Free Energy Principle, Active Inference, Agent-based behavioral models)

Überblick


Realität
Daten
Modellformulierung
Modellschätzung
Modellevaluation

Literaturhinweise

Wir folgen der Darstellung Linearer Normalverteilungsmodelle in Roweis and Ghahramani (1999). Dempster, Laird, and Rubin (1977) bietet eine inhaltsreiche Einführung zum EM Algorithmus. Die Anwendung des EM Algorithmus im Rahmen der Faktorenanalyse geht auf Rubin and Thayer (1982) zurück. Probabilistische Hauptkomponentenanalye wird in Tipping and Bishop (1999) und Roweis (1998) diskutiert und geht auf Arbeiten von Lawley (1953) zurück. Ursprünglich wurde die Hauptkomponentenanalyse von Pearson (1901) vorgeschlagen und insbesondere von Hotelling (1933) verfeinert. Das Anwendungsbeispiel aus dem Gebiet der kognitiven Fähigkeitsforschung geht auf das Beispiel zur konfirmatorischen Faktorenanalyse in Rosseel (2012) basierend auf Joreskog (1969) und Holzinger and Swineford (1939).

Grundlagen

Inferenz und Lernen

Faktorenanalyse

Hauptkomponentenanalyse Revisited

Anwendungsbeispiel

Selbstkontrollfragen

Appendix

Grundlagen

Inferenz und Lernen

Faktorenanalyse

Hauptkomponentenanalyse Revisited

Anwendungsbeispiel

Selbstkontrollfragen

Appendix

Definition (Multivariate Normalverteilung)

X sei ein n-dimensionaler Zufallsvektor mit Ergebnisraum $\mathbb{R}^n$ und WDF

$$p : \mathbb{R}^n \to \mathbb{R}_{>0},\ x \mapsto p(x) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \quad (1)$$

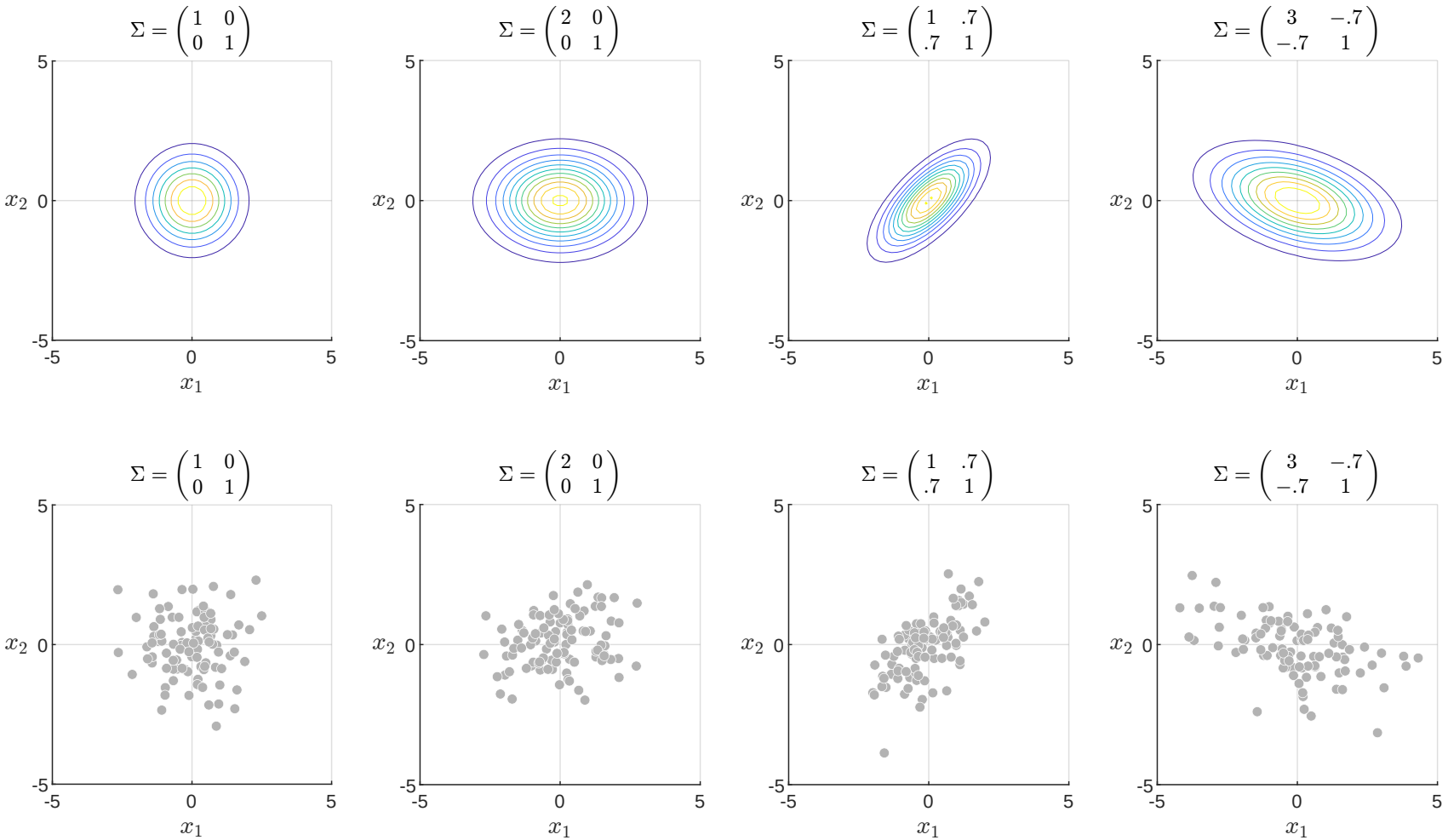
Dann sagen wird, dass X einer *multivariaten (oder n-dimensionalen) Normalverteilung* mit *Erwartungswertparameter* $\mu \in \mathbb{R}^n$ und positive-definitem *Kovarianzmatrixparameter* $\Sigma \in \mathbb{R}^{n \times n}$ unterliegt und nennen X einen *(multivariat) normalverteilten Zufallsvektor*. Wir kürzen dies mit $X \sim N(\mu, \Sigma)$ ab. Die WDF eines multivariat normalverteilten Zufallsvektors bezeichnen wir mit

$$N(x; \mu, \Sigma) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \quad (2)$$

Bemerkungen

- Der Parameter $\mu \in \mathbb{R}^n$ entspricht dem Wert höchster Wahrscheinlichkeitsdichte
- Die Diagonalelemente von Σ spezifizieren die Breite der WDF bezüglich $X_1, \ldots, X_n$.
- Das i, jte Element von Σ spezifiziert die Kovarianz on X_i und X_j.
- Der Term $(2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}}$ ist die Normalsieriungskonstante für den Exponentialfunktionsterm.

Zweidimensionale Normalverteilungen



Definition (Lineares Normalverteilungsmodell)

X sei ein kontinuierlicher nicht-beobachtbarer k-dimensionaler Zufallsvektor und Y sei ein kontinuierlicher beobachbarer m-dimensionaler Zufallsvektor. $B \in \mathbb{R}^{m \times k}$ sei eine Matrix und $R \in \mathbb{R}^{m \times m}$ sei eine positiv-definite Matrix. Dann heißt ein probabilistisches Modell mit WDF

$$p(x, y) = p(y|x)p(x), \tag{3}$$

wobei

$$p(y|x) := N(y; Bx, R) \text{ und } p(x) := N(x; 0_k, I_k) \tag{4}$$

gilt, ein *lineares Normalverteilungsmodell (LNM)*. Die Parametermenge eines LNMs ist $\theta := \{B, R\}$ und wir schreiben die WDFen von LNMen allgemein als

$$p_\theta(x, y) := N(y; Bx, R)N(x; 0_k, I_k). \tag{5}$$

Bemerkungen

- Der Zufallsvektor X heißt auch *Zustandsvektor* oder *latenter Vektor*
- Der Zufallsvektor Y heißt auch *Datenvektor*.
- In hierarchischer Form kann ein LNM geschrieben werden als

$$\begin{aligned} X &= \xi & \xi &\sim N(0_k, I_k) \\ Y &= BX + \eta & \eta &\sim N(0_m, R). \end{aligned} \tag{6}$$

- ξ heißt dabei *Zustandsrauschen*, η heißt dabei *Beobachtungsrauschen/fehler*.
- Samplen eines LNMs resultiert in Realisierungen $(x^{(i)}, y^{(i)})$ mit $i = 1, ..., n$.
- Die $x^{(i)} \in \mathbb{R}^k$ modellieren nicht beobachtbare/latente/virtuelle Daten.
- Die $y^{(i)} \in \mathbb{R}^m$ modellieren beobachtbare Daten.
- LNMe sind spezielle lineare normalverteilte Zustandsraummodelle.

Theorem (LNM Datenverteilung)

Die Datenverteilung des LNMs

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0_k, I_k) \tag{7}$$

ist gegeben durch

$$p_\theta(y) = N(y; 0_m, BB^T + R) \tag{8}$$

und wird auch als *marginale Datenverteilung* oder *marginale Likelihood* bezeichnet.

Beweis

Wir halten zunächst fest, dass mit dem Theorem zu Gemeinsamen Normalverteilungen aus (3) Wahrscheinlichkeitstheorie direkt folgt, dass die WDF der gemeinsamen Verteilung von X und Y gegeben ist durch

$$p_\theta(x, y) = N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 0_k \\ 0_m \end{pmatrix}, \begin{pmatrix} I_k & B^T \\ B & BB^T + R \end{pmatrix}\right). \tag{9}$$

Mit dem Theorem zu Marginalen Normalverteilungen aus (3) Wahrscheinlichkeitstheorie folgt dann aber soforft, dass die WDF der marginalen Verteilung von Y durch

$$p_\theta(y) = N\left(y; 0_m, BB^T + R\right). \tag{10}$$

gegeben ist. □

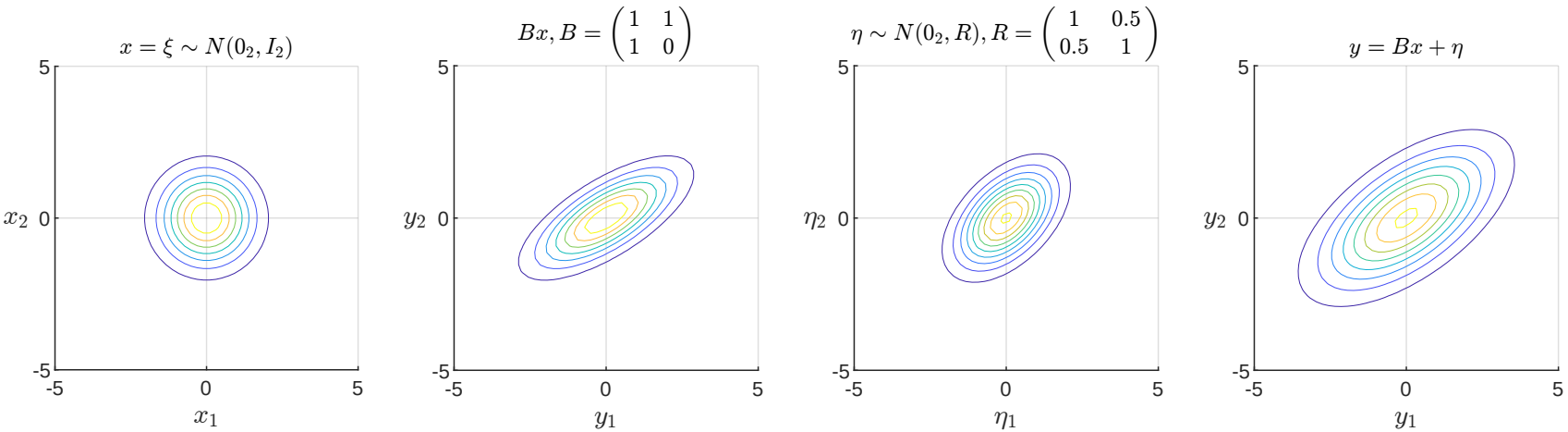
Bemerkungen

- LNMs modellieren zentrierte multivariat normalverteilte Datensätze mit Kovarianzmatrix.

$$\mathbb{C}(Y) = BB^T + R. \tag{11}$$

- "Modellieren" bedeutet hier insbesondere, die Datenkovarianzmatrix $\mathbb{C}(Y)$ zu erklären.
- Die Form von $\mathbb{C}(Y)$ resultiert dabei aus der Transformation einer latenten Normalverteilung.
- Die Matrizen B und R generieren/erklären $\mathbb{C}(Y)$ mechanistisch.
- B und R ermöglichen so oft eine kondensiertere Erklärung als $\mathbb{C}(Y)$ *per se*.
- Verschiedene LNMe haben unterschiedliche Potentiale, Datenkovarianzmatrizen zu erklären.

Grudlagen



- Der latente Zufallsvektor X ist in $\mathbb{R}^k$ (hier $k = 2$) spärisch verteilt.
- Durch B wird diese Sphäre gedehnt, rotiert, und nach $\mathbb{R}^m$ (hier $m = 2$) transformiert.
- Bei $m < p$ sieht die Sphäre von X zum Beispiel wie ein Pfannkuchen aus.
- Dieser Pfannkuchen wird dann noch mit der Kovarianz von des Beobachtungsrauschens η konvolviert.
- Es mag helfen, sich diese Konvolution (Faltung) als “Addition von Rauschen” vorzustellen.

Spezielle lineare Normalverteilungsmodelle

- Das Ziel der Datenmodellierung mit LNMe ist die Erklärung der Datenkovarianmatrixstruktur.
- Die Datenkovarianmatrixstruktur kann durch Wahl von B und R erklärt werden
- Spezielle LNMe entsprechen spezifischen Randbedingungen für R.

  $\Rightarrow$ In der konfirmatorischen Faktorenanalyse wird R als Diagonalmatrix vorausgesetzt.

  $\Rightarrow$ In der probabilistischen PCA wird R als sphärisch vorausgesetzt.

  $\Rightarrow$ In der exploratorischen Faktorenanalyse (PCA) wird $R = 0_{mm}$ vorausgesetzt.

Grundlagen

Inferenz und Lernen

Faktorenanalyse

Hauptkomponentenanalyse Revisited

Anwendungsbeispiel

Selbstkontrollfragen

Appendix

Definition (Verteilungen von LNM Datensätzen)

$$Y := \begin{pmatrix} y^{(1)} & \dots & y^{(n)} \end{pmatrix} \in \mathbb{R}^{m \times n} \text{ und } X := \begin{pmatrix} x^{(1)} & \dots & x^{(n)} \end{pmatrix} \in \mathbb{R}^{k \times n} \tag{12}$$

seien ein beobachteter Datensatz (Realisierungen des beobachtbaren Zufallsvektors) und der assoziierte unbeobachtete Datensatz (Realisierungen des latenten Zufallsvektors). Unter der Annahme unabhängier und identischer Verteilung der gemeinsamen Realisationen $(x^{(i)}, y^{(i)})$ für $i = 1, \dots, n$ ist die WDF der gemeinsamen Verteilung eines LNM Datensatz durch

$$p_\theta(X, Y) = \prod_{i=1}^{n} p_\theta \left(x^{(i)}, y^{(i)} \right) = \prod_{i=1}^{n} N \left(y^{(i)}; Bx^{(i)}, R \right) N \left(x^{(i)}; 0_k, I_k \right) \tag{13}$$

und die marginale WDF des beobachtbaren Datensatzes gegeben durch

$$p_\theta(Y) = \prod_{i=1}^{n} N \left(y^{(i)}; 0_m, BB^T + R \right). \tag{14}$$

Bemerkungen

- X **und** Y **bezeichnen ab jetzt keine Zufallsvektoren mehr.**
- $X \in \mathbb{R}^{k \times n}$ **und** $Y \in \mathbb{R}^{m \times n}$ **bezeichnen ab jetzt Matrizen.**

Inferenz

Was ist die Verteilung der latenten Zufallsvektoren und was sind ihre wahrscheinlichsten Werte für feste Werte der LNM Parameter $\theta := \{B, R\}$?

$\Rightarrow$ Die Antwort gibt das Theorem zu bedingten multivariaten Normalverteilungen.

Lernen

Welche Parameterwerte maximieren die Marginal-Likelihood Funktion

$$L : \Theta \to \mathbb{R}_{\geq 0}, \theta \mapsto L(\theta) := p_\theta(Y) = \int p_\theta(X, Y)\, dX, \tag{15}$$

oder, äquivalent, die Log Marginal-Likelihood Funktion

$$\ell : \Theta \to \mathbb{R}, \theta \mapsto \ell(\theta) := \ln p_\theta(Y) = \ln \int p_\theta(X, Y)\, dX \ ? \tag{16}$$

für einen festen beobachteten Datensatz $Y \in \mathbb{R}^{m \times n}$?

$\Rightarrow$ Die Antwort gibt der Expectation-Maximization Algorithmus.

Theorem (LNM Inferenz)

Es sei

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0, I_k),$$

ein LNM. Dann ist die WDF der bedingten Verteilung des latenten Zufallsvektors gegeben durch

$$p_\theta(x|y) = N\left(x; B^T(BB^T + R)^{-1}y, I_k - B^T(BB^T + R)^{-1}B\right). \tag{17}$$

Bemerkungen

- Die bedingte Verteilung des latenten Zufallsvektors ist eine Normalverteilung.
- Der wahrscheinlichste Wert des latenten ZVs gegeben eine Beobachtung des beobachtbaren ZVs ist also

$$\hat{x} := \mu_{x|y} = B^T(BB^T + R)^{-1}y. \tag{18}$$

- Die mit diesem Wert assoziierte Unsicherheit ist $\Sigma_{x|y} := I_k - B^T(BB^T + R)^{-1}B$.

Beweis

Wir hatten oben bereits gesehen, dass die WDF der gemeinsamen Verteilung von latentem und beobachtbarem Zufallsvektor gegeben ist durch

$$p_\theta(x,y) = N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \begin{pmatrix} 0_k \\ 0_m \end{pmatrix}, \begin{pmatrix} I_k & B^T \\ B & BB^T + R \end{pmatrix}\right) \tag{19}$$

Mit dem Theorem zu Bedingten Normalverteilungen aus Einheit (3) Wahrscheinlichkeitstheorie gilt dann mit Identifikation von

$$\mu_x := 0_k, \mu_y := 0_m, \Sigma_{xx} := I_k, \Sigma_{xy} := B^T, \Sigma_{yx} := B, \text{ und } \Sigma_{yy} := BB^T + R, \tag{20}$$

dass

$$p_\theta(x|y) = N\left(x; \mu_{x|y}, \Sigma_{x|y}\right), \tag{21}$$

wobei

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) = B^T(BB^T + R)^{-1}y, \tag{22}$$

und wobei

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} = I_k - B^T(BB^T + R)^{-1}B. \tag{23}$$

ist. □

Theorem (LNM Datensatzinferenz)

$p_\theta(X, Y)$ sei die gemeinsame Verteilung eines LNM Datenatzes. Dann ist die WDF der bedingten Vertilung von X gegeben Y gegeben durch

$$p_\theta(X|Y) = \prod_{i=1}^{n} N\left(x^{(i)}; B^T(BB^T + R)^{-1}y^{(i)}, I_k - B^T(BB^T + R)^{-1}B\right). \quad (24)$$

Beweis

Mit den Ausdrücken für die WDFen der gemeinsamen und marginalen LNM Datensatzverteilungen gilt

$$p_\theta(X|Y) = \frac{p_\theta(X, Y)}{p_\theta(Y)} = \frac{\prod_{i=1}^{n} p_\theta(x^{(i)}, y^{(i)})}{\prod_{i=1}^{n} p_\theta\left(y^{(i)}\right)} = \prod_{i=1}^{n} \frac{p_\theta(x^{(i)}, y^{(i)})}{p_\theta(y^{(i)})} = \prod_{i=1}^{n} p_\theta\left(x^{(i)}|y^{(i)}\right).$$

Das Theorem folgt dann direkt mit dem Theorem zur LNM Inferenz.

□

Theorem (Evidence Lower Bound)

Für einen Datensatz $Y \in \mathbb{R}^{m \times n}$ sei $\ln p_\theta(Y)$ die WDF der Log Marginal-Likelihood Verteilung eines LNMs. Dann gilt für jede WDF $q(X)$, dass

$$\ln p_\theta(Y) \geq \int q(X) \ln p_\theta(X, Y)\, dX - \int q(X) \ln q(X)\, dX =: \text{ELBO}(q(X), \theta).$$

$\text{ELBO}(q(X), \theta)$ heißt *Evidence Lower Bound*.

Beweis

Mit der Jensenschen Ungleichung (Appendix) gilt

$$\ln p_\theta(Y) := \ln \int p_\theta(X, Y)\, dX = \ln \int q(X) \left(\frac{p_\theta(X, Y)}{q(X)} \right) dX \geq \int q(X) \ln \left(\frac{p_\theta(X, Y)}{q(X)} \right) dX.$$

Damit aber folgt

$$\begin{aligned} \ln p_\theta(Y) \geq \int q(X) \ln \left(\frac{p_\theta(X, Y)}{q(X)} \right) dX &= \int q(X) \left(\ln p_\theta(X, Y) - \ln q(X) \right) dX \\ &= \int q(X) \ln p_\theta(X, Y)\, dX - \int q(X) \ln q(X)\, dX. \end{aligned}$$

Bemerkungen

- Für einen festen Datensatz Y ist ELBO$(q(X), \theta)$ eine Funktion von $q(X)$ und θ.
- Die Bezeichnung Evidence Lower Bound geht auf den Term "Evidence" für $p_\theta(Y)$ zurück.
- In den kognitiven Neurowissenschaften ist die ELBO als "Freie Energie" bekannt.
- Die Signifikanz der ELBO geht weit über LNMe hinaus:
  - Die ELBO ist für die Variational Inference zentral.
  - Variational Inference ist für moderne Theorien zur Funktion des Gehirns zentral.
- Für Einführungen zur Variational Inference siehe zum Beispiel
  - Ostwald et al. (2014), Starke and Ostwald (2017), Blei and Smyth (2017).

Definition (Expectation-Maximization Algorithmus)

Die iterative koordinatenweise Maximierung der ELBO hinsichtlich $q(X)$ und θ heißt *Expectation-Maximization (EM) Algorithmus*. Der Algorithmus hat die allgemeine Form

EM Algorithmus

0. Initialisierung von $q^{(0)}(X)$ und $\theta^{(0)}$

Für $j = 1, 2, \ldots$

1. E Schritt: Setze $q^{(j)}(X) := \arg\max_{q(X)} \text{ELBO}\left(q(X), \theta^{(j-1)}\right)$
2. M Schritt: Setze $\theta^{(j)} := \arg\max_{\theta} \text{ELBO}\left(q^{(j)}(X), \theta\right)$

Nach Konvergenz, nutze $\hat{\theta} := \theta^{(j)}$ als Parameterschätzer.

Bemerkungen

- "Expectation Schritt" ist eine Fehlbezeichnung, es handelt sich auch um einen Maximization Schritt...
- ... allerdings ergibt die Bezeichnung im sogenannten "exakten" EM Algorithmus Sinn.

Theorem (Exakter Expectation-Maximization Algorithmus)

Das Setzen von

$$q^{(j)}(X) := p_{\theta^{(j-1)}}(X|Y) \text{ für alle } j = 1, 2, \ldots \tag{25}$$

im E Schritt des EM Algorithmus maximiert die ELBO hinsichtlich $q(X)$ und heißt *exakter E Schritt*. Der Algorithmus hat dann die folgende Form

Exakter EM Algorithmus

0. Initialisierung von $\theta^{(0)}$

Für $j = 1, 2, \ldots$

1. E Step $q^{(j)}(X) := p_{\theta^{(j-1)}}(X|Y)$
2. M Step $\theta^{(j)} := \arg\max_\theta \int p_{\theta^{(j-1)}}(X|Y) \ln p_\theta(X, Y)\, dX$

Nach Konvergenz, nutze $\hat{\theta} := \theta^{(j)}$ als Parameterschätzer.

Bemerkungen

- Für LNMe kann $p_{\theta^{(j-1)}}(X|Y)$ analytisch evaluiert werden $\Rightarrow$ Inferenz.
- Der M Schritt des exakten EM Algorithmus für LNM Parameterschätzung $\Rightarrow$ Lernen.

Beweis

Wir zeigen zunächst, dass die ELBO $q^{(j)}(X) := p_{\theta^{(j-1)}}(X|Y)$ den maximalen Wert $\ln p_{\theta^{(j-1)}}(Y)$ annimmt:

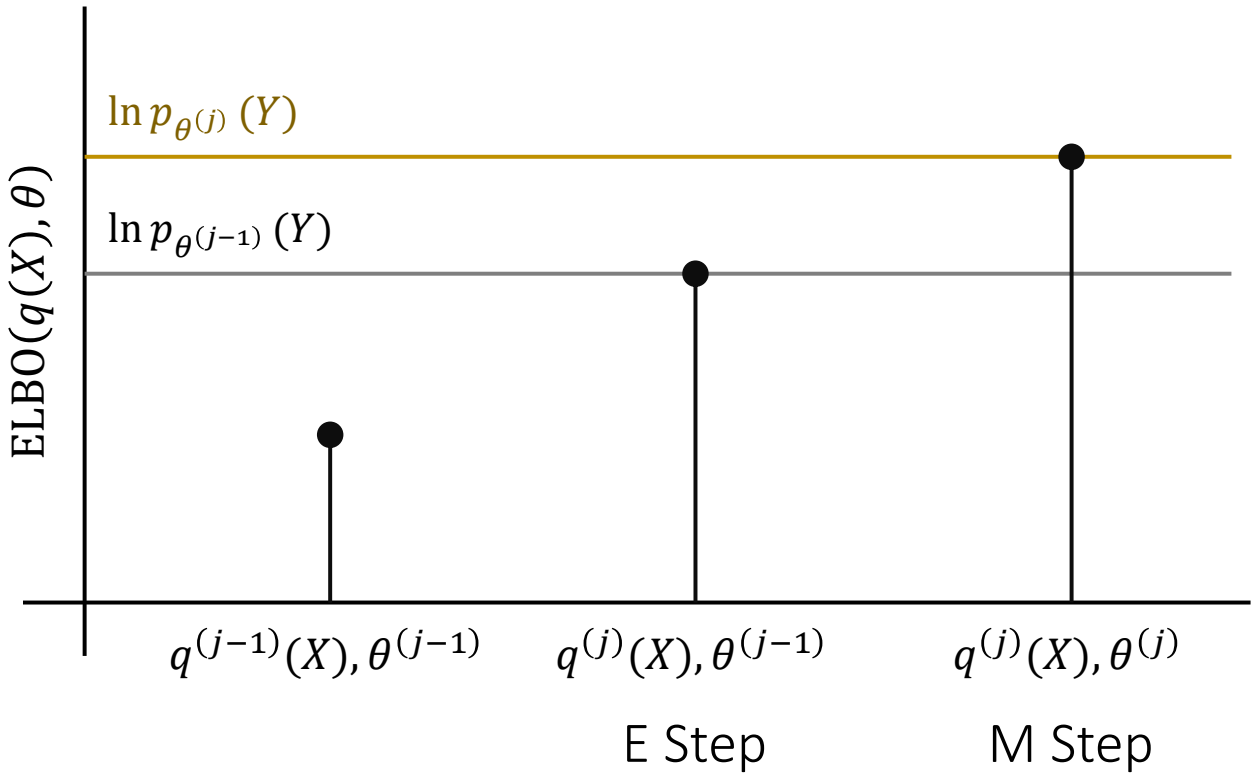
$$\begin{aligned}
\text{ELBO}(p_{\theta^{(j-1)}}(X|Y), \theta) &= \int p_{\theta^{(j-1)}}(X|Y) \ln \left(\frac{p_{\theta^{(j-1)}}(X, Y)}{p_{\theta^{(j-1)}}(X|Y)} \right) dX \\
&= \int p_{\theta^{(j-1)}}(X|Y) \ln \left(\frac{p_{\theta^{(j-1)}}(Y) p_{\theta^{(j-1)}}(X|Y)}{p_{\theta^{(j-1)}}(X|Y)} \right) dX \\
&= \int p_{\theta^{(j-1)}}(X|Y) \ln p_{\theta^{(j-1)}}(Y)\, dX \\
&= \ln p_{\theta^{(j-1)}}(Y) \int p_{\theta^{(j-1)}}(X|Y)\, dX \\
&= \ln p_{\theta^{(j-1)}}(Y).
\end{aligned}$$

Der M Schritt hat dementsprechend die Form

$$\begin{aligned}
\theta^{(j)} &= \arg\max_{\theta} \text{ELBO} \left(p_{\theta^{(j-1)}}(X|Y), \theta \right) \\
&= \arg\max_{\theta} \int p_{\theta^{(j-1)}}(X|Y) \ln p_{\theta}(X, Y)\, dX - \int p_{\theta^{(j-1)}}(X|Y) \ln p_{\theta^{(j-1)}}(X|Y)\, dX.
\end{aligned}$$

Der M Schritt des exakten EM Algorithmus folgt dann damit, dass der zweite Integralterm hier nicht von θ abhängt.

Visuelle Intuition



Weitere Bemerkungen

- Der M Step der $i$ten Iteration des exakten EM Algorithmus entspricht der Maximierung des Erwartungswertes der logarithmierten WDF der gemeinsamen Datenverteilung $p_\theta(X, Y)$ hinsichtlich θ, wobei der Erwartungswert hinsichtlihch der WDF der bedingten Datenverteilung von X gegeben Y basierend auf der Parameterschätzung $\theta^{(j-1)}$, die in der $(j-1)$ten Iteration des exakten EM Algorithmus gewonnen wurde.
- Überraschenderweise garantiert aufgrund der inherenten Logik des EM Algrithmus die Maximierung des Erwartungswertes

$$\mathbb{E}_{p_{\theta^{(j-1)}}(X|Y)}\left(\ln p_\theta(X, Y)\right) = \int p_{\theta^{(j-1)}}(X|Y) \ln p_\theta(X, Y)\, dX \tag{26}$$

  auch die Maximierung der tatsächlichen Funktion von Interesse, $\ln p_\theta(Y)$.
- Für konkrete Algorithmen und für spezifische LNMe muss obiger Erwartungswert analytisch als Funktion von $\theta^{(j-1)}$ ausgewertet werden und dann hinsichtlich θ entweder analytisch oder numerisch maximiert werden um die Parameterschätzung $\theta^{(j)}$. zu erhalten.

Grundlagen

Inferenz und Lernen

Faktorenanalyse

Hauptkomponentenanalyse Revisited

Anwendungsbeispiel

Selbstkontrollfragen

Appendix

Definition (Faktorenanalysemodell)

Es sei

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0, I_k) \tag{27}$$

ein LGM mit diagonaler Beobachtungsrauschen Kovarianzmatrix

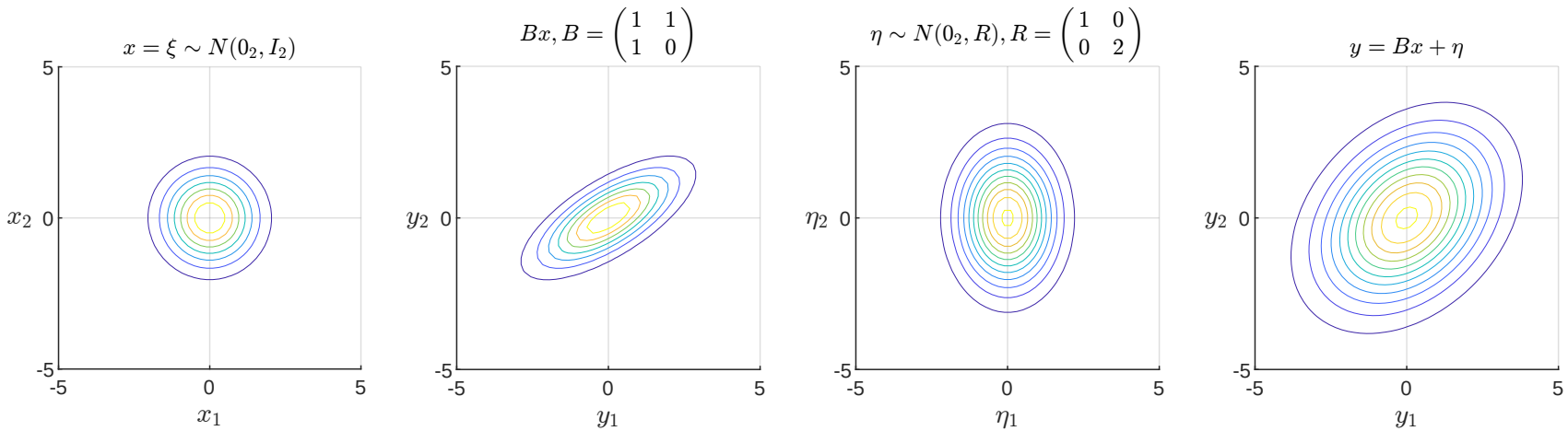
$$R = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2) \in \mathbb{R}^{m \times m}, \sigma_i^2 > 0,\ i = 1, ..., m. \tag{28}$$

Dann heißt $p_\theta(x, y)$ *Faktorenanalysemodell*.

Bemerkungen

- Das Modell heißt auch **Modell der konfirmatorischen Faktorenanalyse**.
- Die Komponenten (Zufallsvariablen) des latenten Zufallsvektors werden **Faktoren** genannt.
- Die Matrix B des Faktorenanalysemodells wird **Faktorenladungsmatrix** genannt.
- Die Diagonalelemente von R werden **Uniquenesses** genannt.

Visuelle Intuition



Bemerkungen

- Faktorenenalysemodelle erklären die Struktur Datenkovarianzmatrizen dadurch, dass
  - alle Korrelationen zwischen Datendimensionen durch B erklärt werden,
  - alle Varianzen dimensionsspezifisch durch R erklärt werden und
  - die Komponenten des beobachtbaren ZVs als bedingt unabhängig angenommen werden.
- Faktorenanalysemodelle behandeln Datenkovarianzen und Varianzen nicht identisch.
- Die Datenkovarianzstruktur wird als bedeutsam angesehen

  $\Leftrightarrow$ Das Beobachtungsrauschen wird als unkorreliert angenommen.
- Exploratorische und konfirmatorische FA entsprechen unterschiedlichen Bedingungen an R.
- Die Parameter des Modells können mit dem exakten EM Algorithmus geschätzt werden.

Theorem (Exakter EM Algorithmus für Faktorenanalysemodelle)

0. Initialisiere $B^{(0)}$ und $R^{(0)}$.

Für $i = 1, 2, \ldots$

1. E Schritt. Setze $\tilde{B} := B^{(j-1)}$ und $\tilde{R} := R^{(j-1)}$ und

$$q^{(j)}(X) := \prod_{j=1}^{n} N\left(x^{(i)}; \hat{x}^{(i)}, \hat{\Sigma}^{(i)}\right), \tag{29}$$

wobei

$$\hat{x}^{(i)} := \tilde{B}^T(\tilde{B}\tilde{B}^T + \tilde{R})^{-1} y^{(i)} \text{ und } \hat{\Sigma}^{(i)} := I_k - \tilde{B}^T(\tilde{B}\tilde{B}^T + \tilde{R})^{-1}\tilde{B}. \tag{30}$$

2. M Schritt. Setze

$$B^{(j)} := \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)^T} \left(\sum_{i=1}^{n} \hat{x}^{(i)} \hat{x}^{(i)^T} + \hat{\Sigma}^{(i)} \right)^{-1} \tag{31}$$

und

$$R^{(j)} := \frac{1}{n} \text{diag} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)^T} B^{(i)^T} \right). \tag{32}$$

Siehe Appendix für einen Beweis.

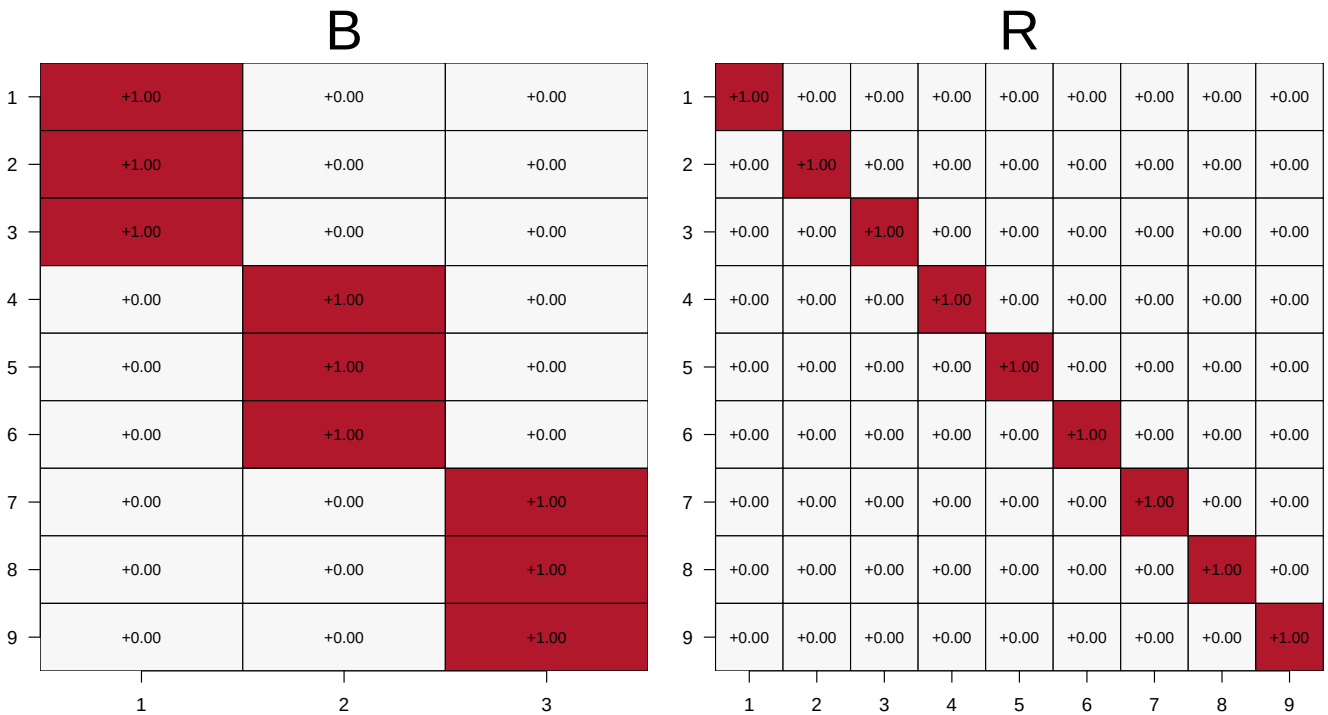
Simulationsbeispiel

Datengeneration

```
# mvrnorm() Paket
library(MASS)

# Paramterspezifikation
k      = 3                                  # Dimension des latenten Zufallsvektors
m      = 9                                  # Dimension des beobachtbaren Zufallsvektors
B      = matrix(c( 1,0,0,                   # Faktorenladungsmatrix
                   1,0,0,
                   1,0,0,
                   0,1,0,
                   0,1,0,
                   0,1,0,
                   0,0,1,
                   0,0,1,
                   0,0,1),
                 nrow  = m,
                 byrow = TRUE)
u      = rep(1,m)                           # Uniquenesses
R      = diag(u)                            # Beobachtungsrauschenkovarianmatrix
mu     = rep(0,k+m)                         # Erwartungswertparameter  p_\theta(x,y)
Sigma  = rbind(cbind(diag(k), t(B)),        # Kovarianzmatrixparameter p_\theta(x,y)
          cbind(B, B %*% t(B) + R))

# Datengeneration
n      = 1e3                                # Beobachtungsanzahl
XY     = t(mvrnorm(n,mu,Sigma))             # p_theta(x,y) sampling und z-score normalization
X      = XY[1:k,]                           # virtuelle Daten    X \in R^{k x n}
Y      = XY[(k+1):nrow(XY),]                # Beobachtete Daten  Y \in R^{m x n}
```

Simulationsbeispiel

Wahre, aber unbekannte, Parameter



Simulationsbeispiel

Marginale Kovarianzmatrix und marginale Korrelationsmatrix

Σ_y

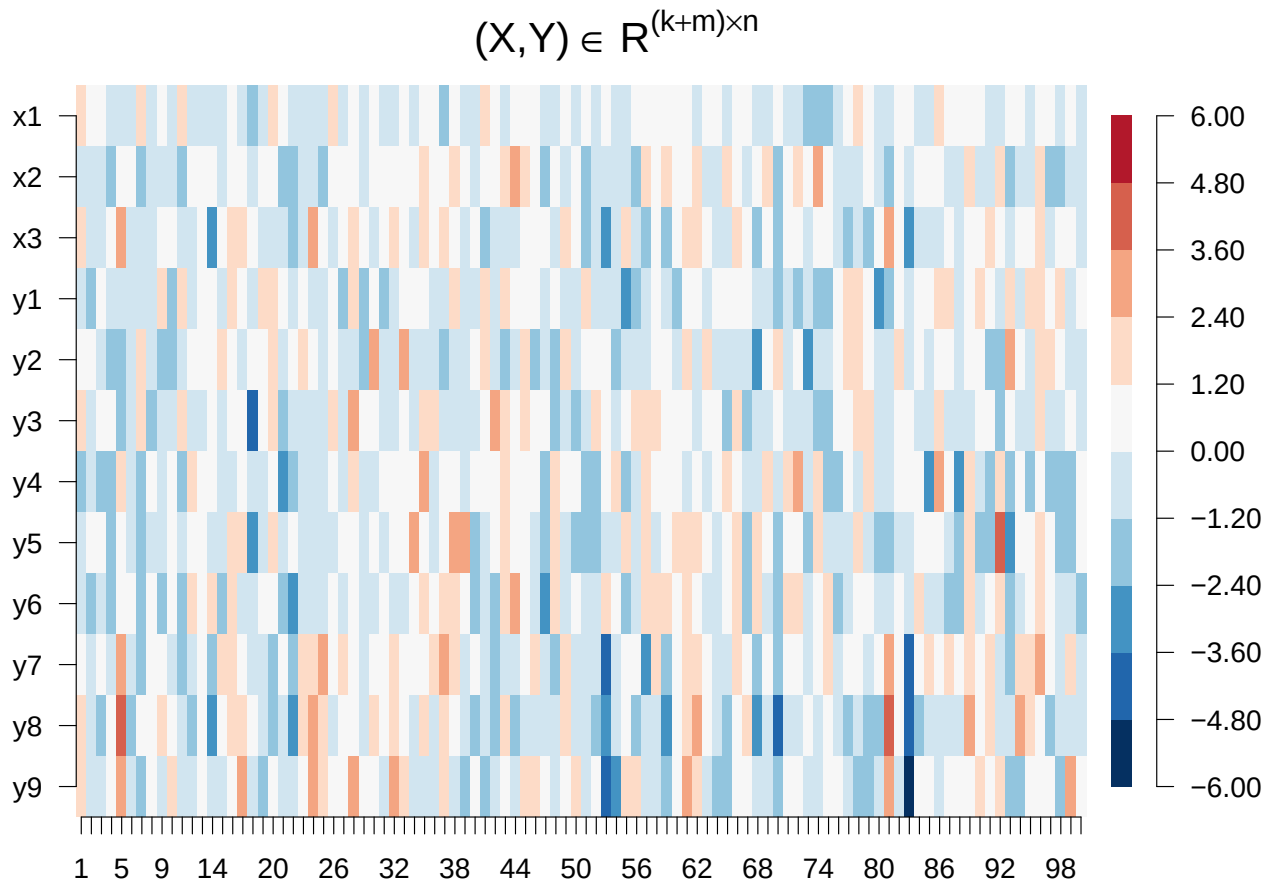
	1	2	3	4	5	6	7	8	9
1	+2.0	+1.0	+1.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0
2	+1.0	+2.0	+1.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0
3	+1.0	+1.0	+2.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0
4	+0.0	+0.0	+0.0	+2.0	+1.0	+1.0	+0.0	+0.0	+0.0
5	+0.0	+0.0	+0.0	+1.0	+2.0	+1.0	+0.0	+0.0	+0.0
6	+0.0	+0.0	+0.0	+1.0	+1.0	+2.0	+0.0	+0.0	+0.0
7	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+2.0	+1.0	+1.0
8	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+1.0	+2.0	+1.0
9	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+1.0	+1.0	+2.0

ρ_y

	1	2	3	4	5	6	7	8	9
1	+1.0	+0.5	+0.5	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0
2	+0.5	+1.0	+0.5	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0
3	+0.5	+0.5	+1.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0
4	+0.0	+0.0	+0.0	+1.0	+0.5	+0.5	+0.0	+0.0	+0.0
5	+0.0	+0.0	+0.0	+0.5	+1.0	+0.5	+0.0	+0.0	+0.0
6	+0.0	+0.0	+0.0	+0.5	+0.5	+1.0	+0.0	+0.0	+0.0
7	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+1.0	+0.5	+0.5
8	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.5	+1.0	+0.5
9	+0.0	+0.0	+0.0	+0.0	+0.0	+0.0	+0.5	+0.5	+1.0

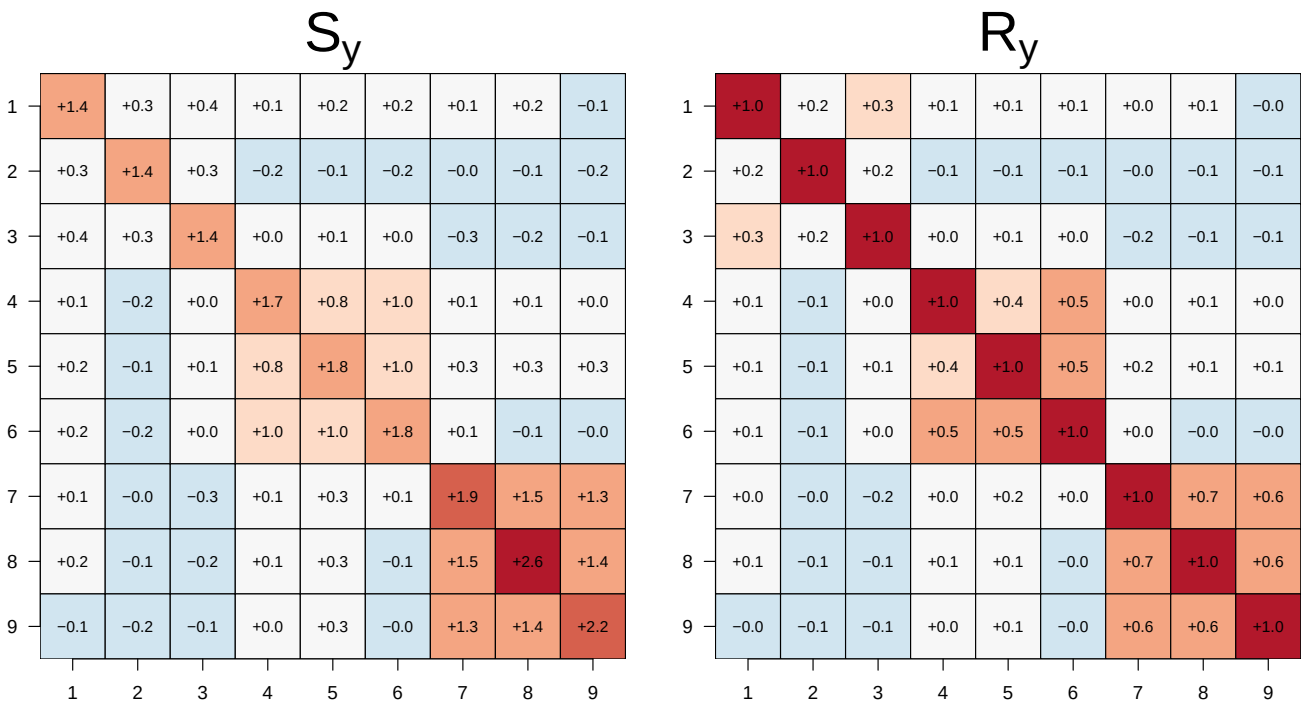
Simulationsbeispiel

Vollständiger Datensatz (n = 100)



Simulationsbeispiel

Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix



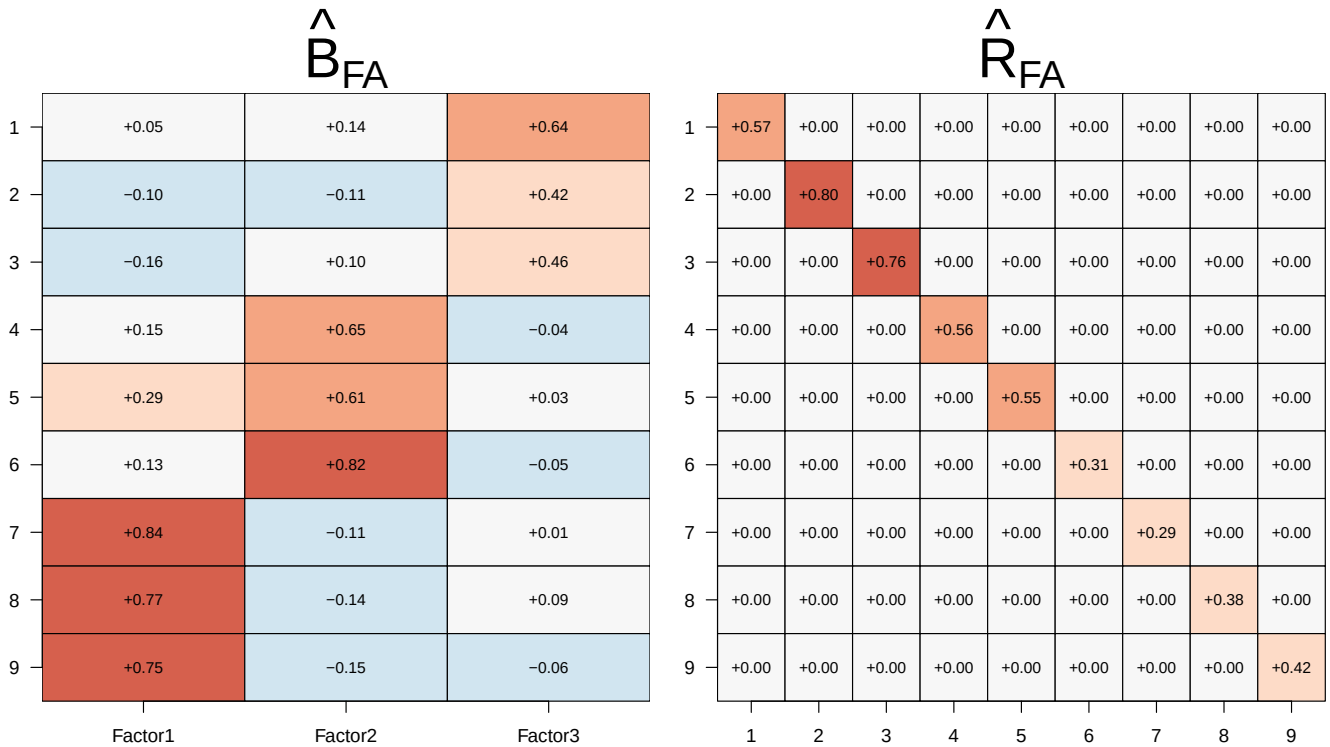
Simulationsbeispiel

Parameterschätzung mit R's `factanal()` Funktion

```
# Lawley & Maxwell (1971) Schätzverfahren für Faktorenanalyse
fa              = factanal(t(Y), factors = 3, rotation = "none")  # Parameterschätzung
B_hat_fa        = fa$loadings[1:9,1:3]                            # \hat{B}
R_hat_fa        = diag(fa$uniquenesses)                           # \hat{R}

# Parametschätzer-basierte Kovarianz und Korrelationsmatrix
Sigma_hat_y_fa  = B_hat_fa %*% t(B_hat_fa) + R_hat_fa
D_hat_y_fa      = diag(1/sqrt(diag(Sigma_hat_y_fa)))
rho_hat_y_fa    = D_hat_y_fa %*% Sigma_hat_y_fa %*% D_hat_y_fa
```

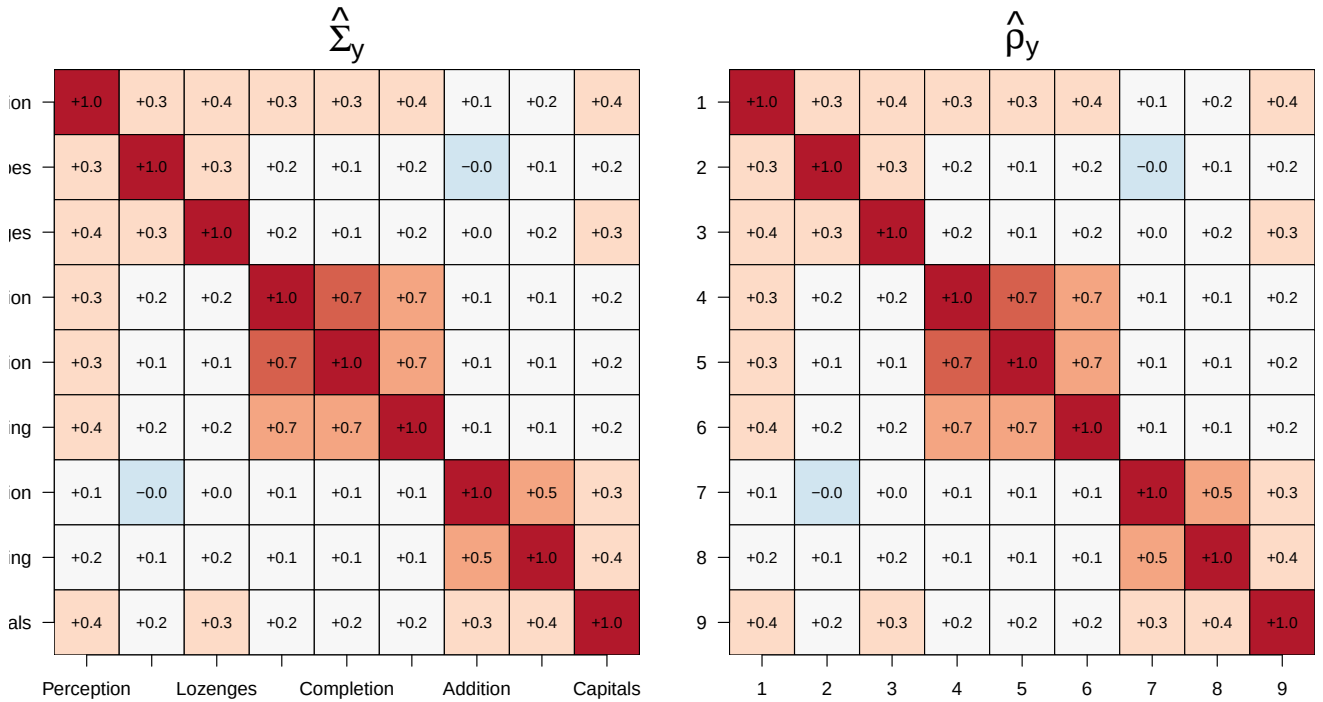
Simulationsbeispiel

`factanal()` basierte Parameterschätzer



Simulationsbeispiel

`factanal()` basierte marginale Kovarianmatrix und Korrelationsmatrix



Simulationsbeispiel

EM Algorithmus

```
library(matlib)                                    # Matrizenrechnungspaket
max_j = 2^2                                        # Maximale Anzahl Iterationen

# EM Initialisierung
B_j = matrix(runif(m*k), nrow = m)
R_j = diag(runif(0,m))

# Iterations
for(j in 1:max_j){

  # E Schritt
  X_hat_j      = t(B_j) %*% inv(B_j %*% t(B_j) + R) %*% Y
  Sigma_hat_j  = diag(k) - (t(B_j) %*% inv(B_j %*% t(B_j) + R) %*% B_j)

  # M Schritt
  yxT = Y %*% t(X_hat_j)
  xxT = X_hat_j %*% t(X_hat_j)
  yyT = Y %*% t(Y)
  B_j = yxT %*% inv(xxT + Sigma_hat_j)
  R_j = (1/n) * diag(diag((yyT - (yxT %*% t(B_j)))))
}
print(yyT)
print(Y %*% t(Y))
# Parameterschätzer
B_hat_em = B_j
R_hat_em = R_j

# Parametschätzer-basierte Kovarianz und Korrelationsmatrix
Sigma_hat_y_em  = B_hat_em %*% t(B_hat_em) + R_hat_em
D_hat_y_em      = diag(1/sqrt(diag(Sigma_hat_y_em)))
rho_hat_y_em    = D_hat_y_em %*% Sigma_hat_y_em %*% D_hat_y_em
```

Simulationsbeispiel

EM Algorithmus-basierte Parameterschätzer

$\hat{B}_{EM}$

−0.47	+0.99	+1.71	1
−0.55	+1.16	+1.77	2
−0.41	+0.95	+1.94	3
+1.14	+2.02	−0.30	4
+1.22	+1.89	−0.50	5
+1.36	+1.84	−0.47	6
+1.76	−1.21	+0.95	7
+1.62	−1.30	+0.98	8
+1.70	−1.44	+0.99	9
1	2	3	

$\hat{R}_{EM}$

+0.69	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
+0.00	+0.62	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
+0.00	+0.00	+0.63	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
+0.00	+0.00	+0.00	+0.61	+0.00	+0.00	+0.00	+0.00	+0.00
+0.00	+0.00	+0.00	+0.00	+0.64	+0.00	+0.00	+0.00	+0.00
+0.00	+0.00	+0.00	+0.00	+0.00	+0.61	+0.00	+0.00	+0.00
+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.67	+0.00	+0.00
+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.65	+0.00
+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.68
1	2	3	4	5	6	7	8	9

Simulationsbeispiel

EM Algorithmus basierte marginale Kovarianmatrix und Korrelationsmatrix

$\hat{\Sigma}_y^{EM}$

	1	2	3	4	5	6	7	8	9
1	+4.8	+4.4	+4.4	+0.9	+0.4	+0.4	−0.4	−0.4	−0.5
2	+4.4	+5.4	+4.8	+1.2	+0.6	+0.5	−0.7	−0.7	−0.9
3	+4.4	+4.8	+5.4	+0.9	+0.3	+0.3	−0.0	+0.0	−0.2
4	+0.9	+1.2	+0.9	+6.1	+5.4	+5.4	−0.7	−1.1	−1.3
5	+0.4	+0.6	+0.3	+5.4	+6.0	+5.4	−0.6	−1.0	−1.2
6	+0.4	+0.5	+0.3	+5.4	+5.4	+6.1	−0.3	−0.7	−0.8
7	−0.4	−0.7	−0.0	−0.7	−0.6	−0.3	+6.1	+5.4	+5.7
8	−0.4	−0.7	+0.0	−1.1	−1.0	−0.7	+5.4	+5.9	+5.6
9	−0.5	−0.9	−0.2	−1.3	−1.2	−0.8	+5.7	+5.6	+6.6

$\hat{\rho}_y^{EM}$

	1	2	3	4	5	6	7	8	9
1	+1.0	+0.9	+0.9	+0.2	+0.1	+0.1	−0.1	−0.1	−0.1
2	+0.9	+1.0	+0.9	+0.2	+0.1	+0.1	−0.1	−0.1	−0.1
3	+0.9	+0.9	+1.0	+0.2	+0.1	+0.0	−0.0	+0.0	−0.0
4	+0.2	+0.2	+0.2	+1.0	+0.9	+0.9	−0.1	−0.2	−0.2
5	+0.1	+0.1	+0.1	+0.9	+1.0	+0.9	−0.1	−0.2	−0.2
6	+0.1	+0.1	+0.0	+0.9	+0.9	+1.0	−0.0	−0.1	−0.1
7	−0.1	−0.1	−0.0	−0.1	−0.1	−0.0	+1.0	+0.9	+0.9
8	−0.1	−0.1	+0.0	−0.2	−0.2	−0.1	+0.9	+1.0	+0.9
9	−0.1	−0.1	−0.0	−0.2	−0.2	−0.1	+0.9	+0.9	+1.0

Grundlagen

Inferenz und Lernen

Faktorenanalyse

Hauptkomponentenanalyse Revisited

Selbstkontrollfragen

Appendix

Definition (Modell der probabilistischen Hauptkomponentenanalyse)

Es sei

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0_k, I_k) \tag{33}$$

ein LNM mit der sphärischen Beobachtungsrauschen Kovarianzmatrix

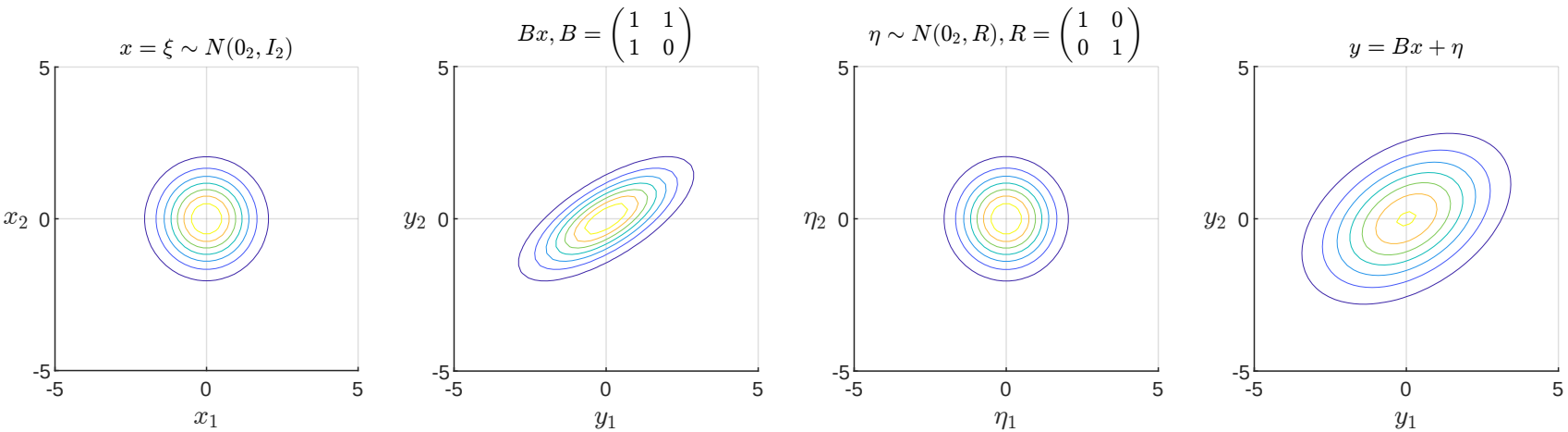
$$R := \sigma^2 I_m . \tag{34}$$

Dann heißt $p_\theta(x, y)$ *probabilistisches Hauptkomponentanalysemodell*.

Bemerkungen

- Die Matrix B bedingt die Beziehung zur klassischen Hauptkomponentenanalyse.
- σ^2 heißt *globales Rauschlevel*.
- B and σ^2 können durch den EM Algorithmus geschätzt werden.
- B and σ^2 können auch durch direkte Maximierung der marginalen Likelihood Funktion geschätzt werden.

Visuelle Intuition zur probabilistischen Hauptkomponentenanalyse



Theorem (Parameter der probabilistischen Hauptkomponentnanalyse)

Es sei

$$p_\theta(x, y) = N(y; Bx, \sigma^2 I_m) N(x; 0, I_k) \tag{35}$$

ein probabilitisches Hauptkomponentenanalysemodell und es sei

$$\mathbb{C}(Y) = Q \Lambda Q^T \tag{36}$$

die Orthogonalzerlegung der Kovarianzmatrix des beobachtbaren Zufallsvektors. Dann kann der Parameter B des probabilistischen Hauptkomponentnanalysemodells geschrieben werden als

$$B = Q \left(\Lambda - \sigma^2 I_m \right)^{1/2}. \tag{37}$$

Bemerkungen

- Wir haben $\mathbb{C}(Y) = Q \Lambda Q^T$ Hauptkomponentenanalyse von $\mathbb{C}(Y)$ genannt.
- Q besteht aus den Eigenvektoren von $\mathbb{C}(y)$, Λ aus den assoziierten Eigenwerten.
- Wir haben die Eigenvektoren von $\mathbb{C}(Y)$ die Hauptkomponenten von $\mathbb{C}(Y)$ genannt.
- Die Spalten von B sind die Hauptkomponenten gewichtet mit $(\Lambda - \sigma^2 I_m)^{1/2}$.

Beweis

Wir halten zunächst fest, dass die marginale Datenverteilung eines probabilistischen Hauptkomponentenanalysemodells gegeben ist durch

$$p_\theta(y) = N(y; 0_m, BB^T + \sigma^2 I_m) \text{ and thus } \mathbb{C}(Y) = BB^T + \sigma^2 I_m. \tag{38}$$

Einsetzen von $B = Q(\Lambda - \sigma^2 I_m)^{1/2}$ ergibt dann

$$\begin{aligned}
\mathbb{C}(Y) &= BB^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m)^{1/2}(Q(\Lambda - \sigma^2 I_m)^{1/2})^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m)^{1/2}((\Lambda - \sigma^2 I_m)^{1/2})^T Q^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m)^{1/2}(\Lambda - \sigma^2 I_m)^{1/2} Q^T + \sigma^2 I_m \\
&= Q(\Lambda - \sigma^2 I_m) Q^T + \sigma^2 I_m \\
&= (Q\Lambda - \sigma^2 Q I_m) Q^T + \sigma^2 I_m \\
&= Q\Lambda Q^T - \sigma^2 Q Q^T + \sigma^2 I_m \\
&= Q\Lambda Q^T - \sigma^2 I_m + \sigma^2 I_m \\
&= Q\Lambda Q^T.
\end{aligned} \tag{39}$$

Also ergibt sich die Äquivalenz

$$\mathbb{C}(y) = Q\Lambda Q^T \Leftrightarrow B = Q(\Lambda - \sigma^2 I_m)^{\frac{1}{2}}. \tag{40}$$

Theorem (Direkte Marginal-Maximum Likelihood Schätzung)

Es sei

$$p_\theta(x, y) = N(y; Bx, \sigma^2 I_m) N(x; 0, I_k) \tag{41}$$

ein probabilistisches Hauptkomponentenanalyse model und $Y \in \mathbb{R}^{m \times n}$ sei ein von diesem Modell generierter Datensatz. Weiterhin sei

$$C = \frac{1}{n} Y Y^T \text{ und } C = Q \Lambda Q^T \tag{42}$$

die unkorrigierte Stichprobenkovarianzmatrix und ihre Orhtogonazerlegung, respektive. Schließlich seien $Q_q \in \mathbb{R}^{m \times q}$ und $\Lambda_q \in \mathbb{R}^{q \times q}$ die Matrizen die durch die spaltenweise Konkatenation der höchsten Eigenwerte und ihrer assoziierten Eigenvektoren entstehen. Dann sind Maximum Marginal Likelihood Schätzer von B und σ^2 durch

$$\hat{B} = Q_q (\Lambda_q - \sigma^2 I_m)^{1/2} \text{ and } \hat{\sigma}^2 = \frac{1}{m - l} \sum_{j=l+1}^{m} \lambda_j, \tag{43}$$

gegeben.

Siehe Appendix für einen Beweis.

Definition (Modell der Hauptkomponentenanalyse)

Es sei

$$p_\theta(x, y) = N(y; Bx, R)N(x; 0, I_k) \tag{44}$$

ein LNM mit der Beobachtungsrauschenkovarianzmatrix

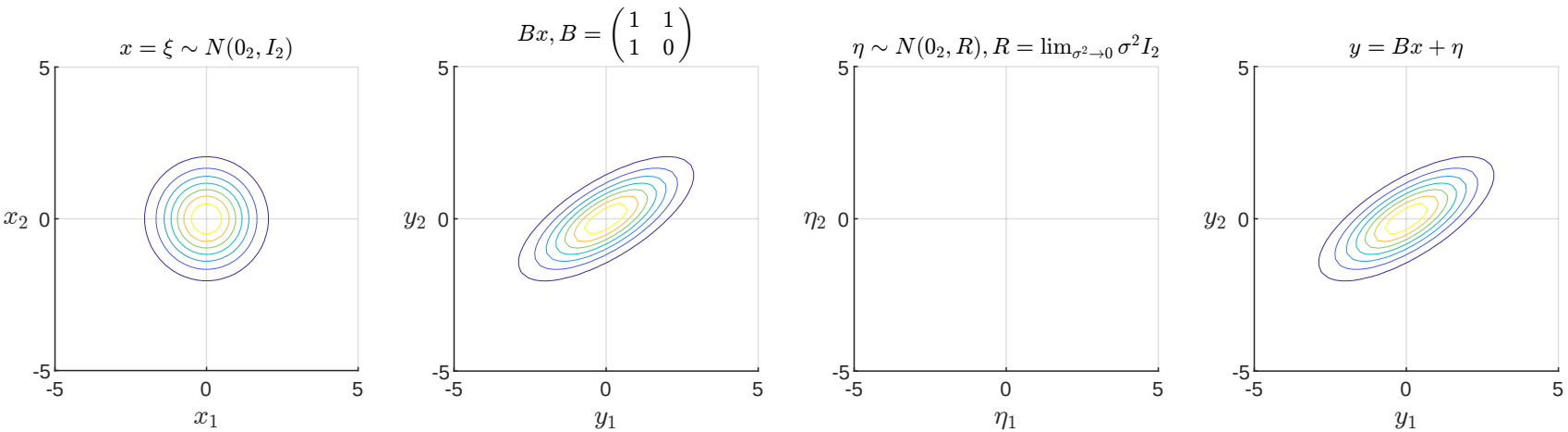
$$R := \lim_{\sigma^2 \to 0} \sigma^2 I_m \in \mathbb{R}^{m \times m}. \tag{45}$$

Dann heißt $p_\theta(x, y)$ *Modell der Hauptkomponentenanalyse*.

Bemerkungen

- B enkodiert die Beziehung zur klassischen Hauptkomponentenanalyse.
- Das Beobachtungsrauschen wird als nicht-existent angenommen.

Visuelle Intuition zur Hauptkomponentenanalyse



Theorem (Parameter der Hauptkomponentenanalyse)

Es sei

$$p_\theta(x, y) = N\left(y; Bx, \lim_{\sigma^2 \to 0} \sigma^2 I_m\right) N(x; 0_k, I_k) \tag{46}$$

ein Hauptkomponentenanalysemodell und es sei

$$\mathbb{C}(Y) = Q^T \Lambda Q \tag{47}$$

die Hauptkomponentenanalyse der zugehörigen marginalen Datenverteilung. Dann gilt

$$B = Q\Lambda^{\frac{1}{2}}. \tag{48}$$

Beweis

Wir halten zunächst fest, dass die marginale Datenverteilung des Hauptkomponentenanalysemodells im Limit $\sigma^2 \to 0$ gegeben ist durch

$$p_\theta(y) = N(y; 0_m, BB^T) \text{ and thus } \mathbb{C}(y) = BB^T. \tag{49}$$

Es ergibt sich also

$$\mathbb{C}(y) = BB^T \Leftrightarrow Q\Lambda Q^T = BB^T \Leftrightarrow B = Q\Lambda^{\frac{1}{2}}. \tag{50}$$

□

Grundlagen

Inferenz und Lernen

Faktorenanalyse

Hauptkomponentenanalyse Revisited

Anwendungsbeispiel

Selbstkontrollfragen

Appendix

Nine mental ability tests | Intelligenzforschungsdatensatz

Holzinger and Swineford (1939)

Visualization

1. Visual Perception
2. Cubes
3. Lozenges

Verbal intelligence

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

Speed

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

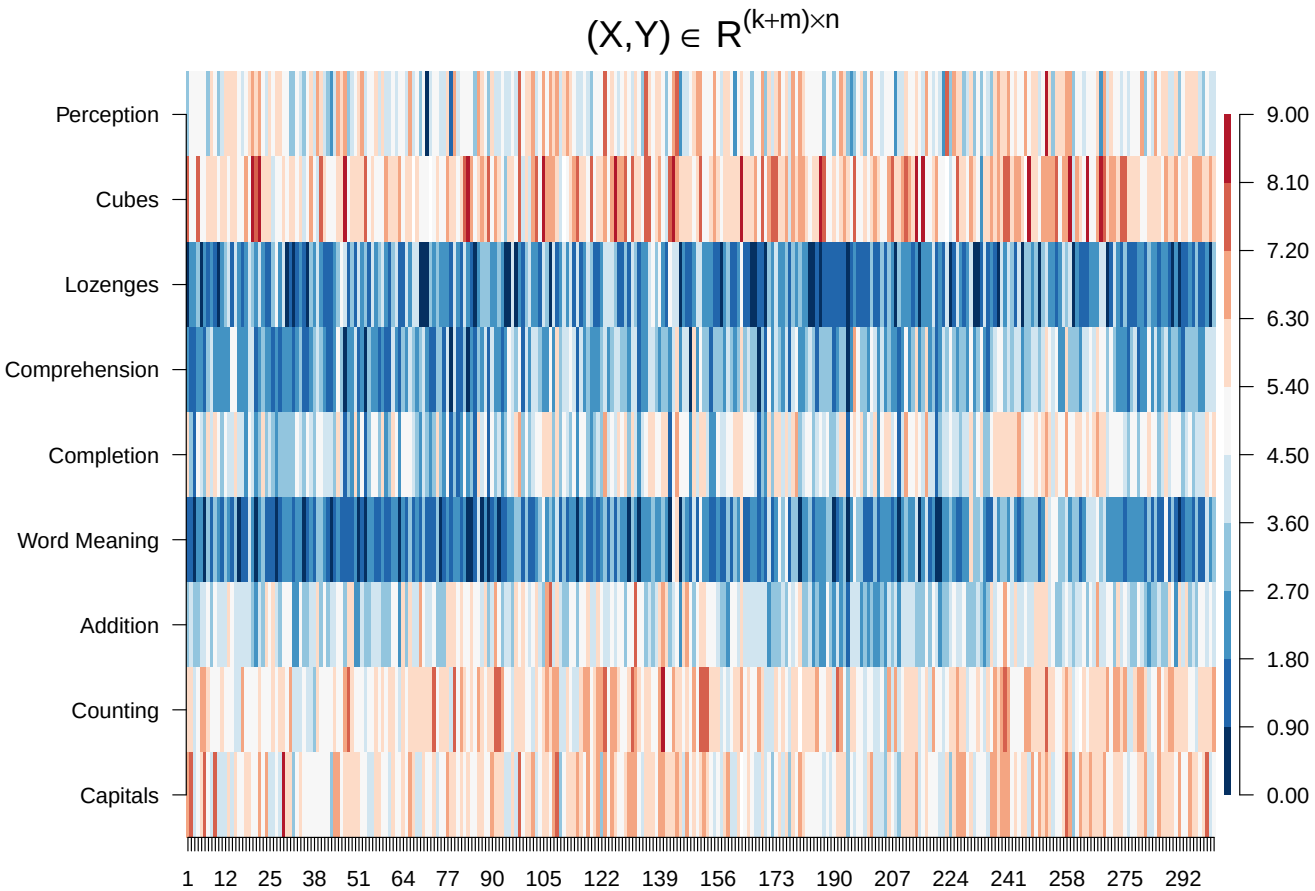
Beobachteter Datensatz (n = 301)

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

Beobachteter Datensatz (n = 301)



Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix

S_y

	Perception	Cubes	Lozenges	Comprehension	Completion	Word Meaning	Addition	Counting	Capitals
Perception	+1.4	+0.4	+0.6	+0.5	+0.4	+0.5	+0.1	+0.3	+0.5
Cubes	+0.4	+1.4	+0.5	+0.2	+0.2	+0.2	−0.1	+0.1	+0.2
Lozenges	+0.6	+0.5	+1.3	+0.2	+0.1	+0.2	+0.1	+0.2	+0.4
Comprehension	+0.5	+0.2	+0.2	+1.4	+1.1	+0.9	+0.2	+0.1	+0.2
Completion	+0.4	+0.2	+0.1	+1.1	+1.7	+1.0	+0.1	+0.2	+0.3
Word Meaning	+0.5	+0.2	+0.2	+0.9	+1.0	+1.2	+0.1	+0.2	+0.2
Addition	+0.1	−0.1	+0.1	+0.2	+0.1	+0.1	+1.2	+0.5	+0.4
Counting	+0.3	+0.1	+0.2	+0.1	+0.2	+0.2	+0.5	+1.0	+0.5
Capitals	+0.5	+0.2	+0.4	+0.2	+0.3	+0.2	+0.4	+0.5	+1.0

R_y

	Perception	Cubes	Lozenges	Comprehension	Completion	Word Meaning	Addition	Counting	Capitals
Perception	+1.0	+0.3	+0.4	+0.4	+0.3	+0.4	+0.1	+0.2	+0.4
Cubes	+0.3	+1.0	+0.3	+0.2	+0.1	+0.2	−0.1	+0.1	+0.2
Lozenges	+0.4	+0.3	+1.0	+0.2	+0.1	+0.2	+0.1	+0.2	+0.3
Comprehension	+0.4	+0.2	+0.2	+1.0	+0.7	+0.7	+0.2	+0.1	+0.2
Completion	+0.3	+0.1	+0.1	+0.7	+1.0	+0.7	+0.1	+0.1	+0.2
Word Meaning	+0.4	+0.2	+0.2	+0.7	+0.7	+1.0	+0.1	+0.1	+0.2
Addition	+0.1	−0.1	+0.1	+0.2	+0.1	+0.1	+1.0	+0.5	+0.3
Counting	+0.2	+0.1	+0.2	+0.1	+0.1	+0.1	+0.5	+1.0	+0.4
Capitals	+0.4	+0.2	+0.3	+0.2	+0.2	+0.2	+0.3	+0.4	+1.0

Modellschätzung

Die Art der Tests motiviert ein 3-Faktor-Modell mit Faktoren

- Visualisierungsvermögen
- Verbale Intelligenz
- Schnelligkeit

```
# Lawley & Maxwell (1971) Schätzverfahren für Faktorenanalyse
fa              = factanal(t(Y), factors = 3, rotation = "none")  # Parameterschätzung
B_hat_fa        = fa$loadings[1:9,1:3]                            # \hat{B}
R_hat_fa        = diag(fa$uniquenesses)                           # \hat{R}

# Parametschätzer-basierte Kovarianz und Korrelationsmatrix
Sigma_hat_y_fa  = B_hat_fa %*% t(B_hat_fa) + R_hat_fa
D_hat_y_fa      = diag(1/sqrt(diag(Sigma_hat_y_fa)))
rho_hat_y_fa    = D_hat_y_fa %*% Sigma_hat_y_fa %*% D_hat_y_fa
```

Modellschätzung

$\hat{B}_{FA}$

	Factor1	Factor2	Factor3
1	+0.15	+0.54	−0.39
2	+0.05	+0.70	−0.33
3	+0.07	+0.76	−0.21
4	−0.40	+0.31	+0.52
5	−0.53	+0.29	+0.40
6	−0.57	+0.34	+0.44
7	+0.59	−0.01	+0.23
8	+0.74	+0.15	+0.43
9	+0.63	+0.26	+0.20

$\hat{R}_{FA}$

	1	2	3	4	5	6	7	8	9
1	+0.54	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
2	+0.00	+0.40	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
3	+0.00	+0.00	+0.38	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00
4	+0.00	+0.00	+0.00	+0.47	+0.00	+0.00	+0.00	+0.00	+0.00
5	+0.00	+0.00	+0.00	+0.00	+0.47	+0.00	+0.00	+0.00	+0.00
6	+0.00	+0.00	+0.00	+0.00	+0.00	+0.36	+0.00	+0.00	+0.00
7	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.59	+0.00	+0.00
8	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.24	+0.00
9	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.00	+0.49

Marginale Kovarianmatrix und Korrelationsmatrix

$\hat{\Sigma}_y$

	Perception	Cubes	Lozenges	Comprehension	Completion	Word Meaning	Addition	Counting	Capitals
Perception	+1.0	+0.3	+0.4	+0.3	+0.3	+0.4	+0.1	+0.2	+0.4
Cubes	+0.3	+1.0	+0.3	+0.2	+0.1	+0.2	−0.0	+0.1	+0.2
Lozenges	+0.4	+0.3	+1.0	+0.2	+0.1	+0.2	+0.0	+0.2	+0.3
Comprehension	+0.3	+0.2	+0.2	+1.0	+0.7	+0.7	+0.1	+0.1	+0.2
Completion	+0.3	+0.1	+0.1	+0.7	+1.0	+0.7	+0.1	+0.1	+0.2
Word Meaning	+0.4	+0.2	+0.2	+0.7	+0.7	+1.0	+0.1	+0.1	+0.2
Addition	+0.1	−0.0	+0.0	+0.1	+0.1	+0.1	+1.0	+0.5	+0.3
Counting	+0.2	+0.1	+0.2	+0.1	+0.1	+0.1	+0.5	+1.0	+0.4
Capitals	+0.4	+0.2	+0.3	+0.2	+0.2	+0.2	+0.3	+0.4	+1.0

$\hat{\rho}_y$

	Perception	Cubes	Lozenges	Comprehension	Completion	Word Meaning	Addition	Counting	Capitals
Perception	+1.0	+0.3	+0.4	+0.3	+0.3	+0.4	+0.1	+0.2	+0.4
Cubes	+0.3	+1.0	+0.3	+0.2	+0.1	+0.2	−0.0	+0.1	+0.2
Lozenges	+0.4	+0.3	+1.0	+0.2	+0.1	+0.2	+0.0	+0.2	+0.3
Comprehension	+0.3	+0.2	+0.2	+1.0	+0.7	+0.7	+0.1	+0.1	+0.2
Completion	+0.3	+0.1	+0.1	+0.7	+1.0	+0.7	+0.1	+0.1	+0.2
Word Meaning	+0.4	+0.2	+0.2	+0.7	+0.7	+1.0	+0.1	+0.1	+0.2
Addition	+0.1	−0.0	+0.0	+0.1	+0.1	+0.1	+1.0	+0.5	+0.3
Counting	+0.2	+0.1	+0.2	+0.1	+0.1	+0.1	+0.5	+1.0	+0.4
Capitals	+0.4	+0.2	+0.3	+0.2	+0.2	+0.2	+0.3	+0.4	+1.0

$\Rightarrow$ Modellierung des Datensatzes mit einem 3-Faktor Modell ist möglich.

Modellvergleich | Welche Anzahl von Faktoren zwischen 1 und 5 ist am besten?

Bayesian Information Criterion (cf. Horvath et al. (2021), Schwarz (1978))

$$\text{BIC} := \ln p_{\hat{\theta}}(Y) - \frac{l}{2} \ln n \tag{51}$$

$\ln p_{\hat{\theta}}(Y)$

- Logarithmierte marginale Datenwahrscheinlichkeit(sdichte) bei optimierten Parametern
- Maß für die Passungsgüte des Modells

$\frac{l}{2} \ln n$

- Stichprobengröße gewichtete Anzahl an Parametern l
- Maß für die Komplexität des Modells

$$\text{BIC} = \text{Passungsgüte} - \text{Komplexität} \tag{52}$$

Modellvergleich | Welche Anzahl von Faktoren zwischen 1 und 5 ist am besten?

Theorem (BIC für Faktorenanalysemodelle)

$p_\theta(X, Y)$ sei die WDF der gemeinsamen Verteilung eines Faktorenanalysemodelldatensatzes mit $X \in \mathbb{R}^{k \times n}$ und $Y \in \mathbb{R}^{m \times n}$. Weiterhin seien $\hat{\theta} := \{\hat{B}, \hat{R}\}$ (marginale) Maximum Likelihood Schätzer von $\theta := \{B, R\}$ und es sei

$$\hat{\Sigma} := \hat{B}\hat{B}^T + \hat{R}. \tag{53}$$

Dann ergibt sich das Bayesian Information Criterion zu

$$\text{BIC} = -\frac{nm}{2} \ln 2\pi - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{1}{2} \text{tr}\left(\hat{\Sigma}^{-1} Y Y^T\right) - \frac{mk + m}{2} \ln n \tag{54}$$

Modellvergleich | Welche Anzahl von Faktoren zwischen 1 und 5 ist am besten?

Beweis

Nach Definition der Verteilung von LNM Datensätzen gilt

$$p_{\hat{\theta}}(Y) = \prod_{i=1}^{n} N\left(y^{(i)}; 0_m, \hat{B}\hat{B}^T + \hat{R}\right). \tag{55}$$

Die Anzahl der Parameter eines Faktorenanalysemodells ergibt sich als $l = mk + m$, wobei mk die Anzahl der Einträge in B und m die Anzahl der Einträge in R sind. Mit der Eigenschaft

$$x^T Ax = \mathsf{tr}(Axx^T) \tag{56}$$

der Matrix Spur

$$\mathsf{tr}(A) := \sum_{i=1}^{n} a_i i \text{ für } A := (a_{ij})_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n} \tag{57}$$

und der Definition von

$$\hat{\Sigma} := \hat{B}\hat{B}^T + \hat{R} \tag{58}$$

ergibt sich dann

$$\mathsf{BIC} = \ln p_{\hat{\theta}}(Y) - \frac{l}{2} \ln n = \ln \left(\prod_{i=1}^{n} N\left(y^{(i)}; 0_m, \hat{\Sigma}\right) \right) - \frac{mk + m}{2} \ln n \tag{59}$$

Modellvergleich | Welche Anzahl von Faktoren zwischen 1 und 5 ist am besten?

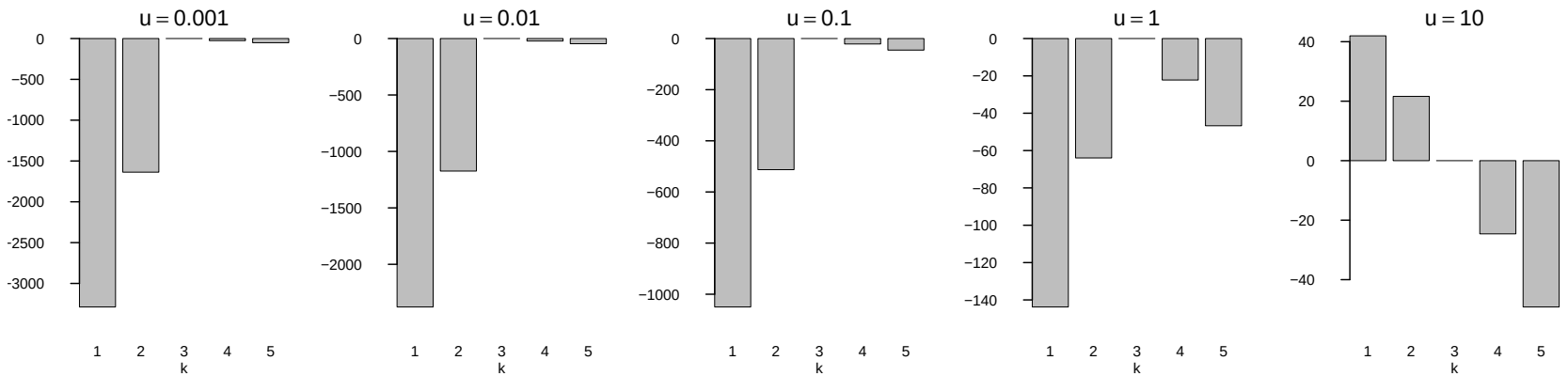
Beweis (fortgeführt)

Es ergibt sich also

$$
\begin{aligned}
\text{BIC} &= \sum_{i=1}^{n} \ln N\left(y^{(i)}; 0_m, \hat{\Sigma}\right) - \frac{mk+m}{2} \ln n \\
&= \sum_{i=1}^{n} \ln\left((2\pi)^{-\frac{m}{2}} |\hat{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} {y^{(i)}}^T \Sigma^{-1} y^{(i)}\right)\right) - \frac{mk+m}{2} \ln n \\
&= \sum_{i=1}^{n} \left(-\frac{m}{2} \ln 2\pi - \frac{1}{2} \ln |\hat{\Sigma}| - \frac{1}{2} {y^{(i)}}^T \Sigma^{-1} y^{(i)}\right) - \frac{mk+m}{2} \ln n \\
&= -\frac{nm}{2} \ln 2\pi - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} {y^{(i)}}^T \Sigma^{-1} y^{(i)} - \frac{mk+m}{2} \ln n \\
&= -\frac{nm}{2} \ln 2\pi - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{1}{2} \text{tr}\left(\Sigma^{-1} \sum_{i=1}^{n} y^{(i)} {y^{(i)}}^T\right) - \frac{mk+m}{2} \ln n \\
&= -\frac{nm}{2} \ln 2\pi - \frac{n}{2} \ln |\hat{\Sigma}| - \frac{1}{2} \text{tr}\left(\hat{\Sigma}^{-1} Y Y^T\right) - \frac{mk+m}{2} \ln n
\end{aligned}
\tag{60}
$$

Modellvergleich | Welche Anzahl von Faktoren zwischen 1 und 5 ist am besten?

Simulationsbasierte Validierung des BIC bei $k = 3, n = 301$ und $R := uI_m$ mit $u > 0$.



$\Rightarrow$ BIC erlaubt Model Recovery in Szenarien mit niedrigem bis mittlerem Beobachtungsrauschen.

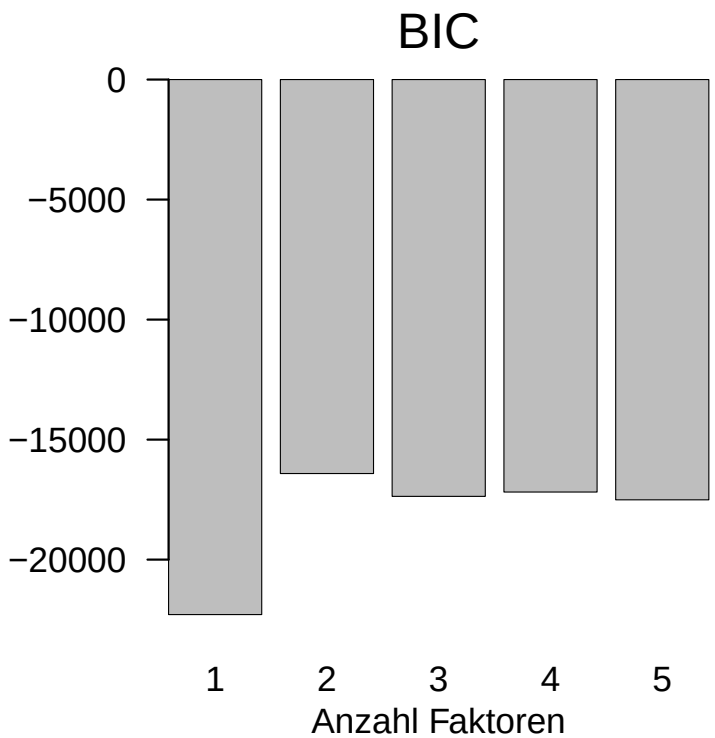
Modellvergleich | Welche Anzahl von Faktoren zwischen 1 und 5 ist am besten?

BIC Modellevaluation

```
# Datensatz
library(lavaan)                                          # Lavaan SEM Paket
data(HolzingerSwineford1939)                             # Datensatz
Y              = t(HolzingerSwineford1939[,7:15])        # Datenmatrix
m              = nrow(Y)                                 # Anzahl Tests/Variablen
n              = ncol(Y)                                 # Anzahl Proband:innen
max_k          = 5                                       # Maximale Faktorenanzahl
BIC            = rep(NaN,max_k)                          # BIC

# Modell Iterationen
for(k in 1:max_k){

  # Modellformulierung und Modellschätzung
  fa         = factanal(t(Y), factors = k, rotation = "none")
  B_hat      = fa$loadings[,1:k]
  R_hat      = diag(fa$uniquenesses)

  # Modellschätzer
  Sigma_hat = B_hat %*% t(B_hat) + R_hat

  # Modellevaluation
  BIC[k]      = (-((n*m)/2)*log(2*pi)
                 -(n/2*log(det(Sigma_hat)))
                 -(1/2)*sum(diag(inv(Sigma_hat)%*%Y%*%t(Y)))
                 -(((m*k)+m)/2)*log(n))
}
```

Modellvergleich | Welche Anzahl von Faktoren zwischen 1 und 5 ist am besten?

BIC Modellevaluation



$\Rightarrow$ Das BIC Modellvergleichskriterium legt ein Modell mit 2 Faktoren nahe.

Selbstkontrollfragen

1. Definieren Sie den Begriff des Linearen Normalverteilungsmodells (LNMs).
2. Erläutern Sie die hierarchische Form eines LNMs.
3. Geben Sie das Theorem zur LNM Datenverteilung wieder.
4. Was modellieren/erklären LNMs?
5. Definieren Sie die WDFen der gemeinsamen und marginalen Datenverteilungen eines LNM Datensatzes.
6. Erläutern Sie die Fragestellungen der Inferenz und des Lernens bei der Schätzung von LNMen.
7. Geben Sie das Theorem zum Exakten Expectation-Maximization Algorithmus wieder.
8. Erläutern Sie das Theorem zum Exakten Expectation-Maximization Algorithmus.
9. Geben Sie die Definition eines Faktorenanalysemodells wieder.
10. Erläutern Sie das Verhältnis von Datenkomponentenkovarianzen und -varianzen im Faktorenanalysemodell.
11. Geben Sie die Definition eines Hauptkomponentenanalysemodells wieder.
12. Geben Sie das Theorem zum Parameter des Hauptkomponentenanalysemodells wieder.
13. Erläutern Sie den Zusammenhang von modell-freier PCA (Einheit (4)) und modell-basierter PCA (Einheit (5)).
14. Definieren Sie das Bayesian Information Criterion (BIC).
15. Erläutern Sie die Intuition zum BIC im Kontext von Modellvergleichen.

Grundlagen

Inferenz und Lernen

Faktorenanalyse

Hauptkomponentenanalyse Revisited

Selbstkontrollfragen

Appendix

Theorem (Jensen's inequality)

Let x be a random variable and g be a convex function, i.e.

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda)g(x_2). \tag{61}$$

Then

$$\mathbb{E}(g(x)) \geq g(\mathbb{E}(x)). \tag{62}$$

Conversely, let g be a concave function, i.e.

$$g(\lambda x_1 + (1-\lambda)x_2) \geq \lambda g(x_1) + (1-\lambda)g(x_2). \tag{63}$$

Then

$$\mathbb{E}(g(x)) \leq g(\mathbb{E}(x)). \tag{64}$$

Proof

By adapting the proof of Theorem 4.7.8 in Casella and Berger (2012) , we show the inequality for the concave case. Let f be a tangent line at the point $g(\mathbb{E}(x))$, i.e. is a linear-affine function of the form $f(x) := ax + b$ for some $a, b \in \mathbb{R}$ with $f(\mathbb{E}(x)) = g(\mathbb{E}(x))$. Because g is concave, we have $g(x) \leq ax + b$ for all $x \in \mathbb{R}$ and thus als $g(x) \leq ax + b$. Hence,

$$\mathbb{E}(g(x)) \leq \mathbb{E}(ax + b) = a\mathbb{E}(x) + b = f(\mathbb{E}(x)) = g(\mathbb{E}(x)). \tag{65}$$

Remarks

- For convex g the function's graph lies below the straight line $g(x_1)$ to $g(x_2)$.
- For concave g the function's graph lies above the straight line $g(x_1)$ to $g(x_2)$.
- The logarithm is a concave function, hence $\mathbb{E}(\ln x) \leq \ln \mathbb{E}(x)$.

Beweis des Theorems zum exakten EM Algorithmus eines Faktorenanalysemodells

The E Step of the algorithm follows directly with the LGM data set inference theorem. We thus focus on the derivation of the M Step. To this end, recall that the M Step of the exact EM algorithm takes the form

$$\theta^{(j)} := \arg\max_{\theta} \int p_{\theta^{(j-1)}}(X|Y) \ln p_{\theta}(X, Y)\, dX \tag{66}$$

For LGMs, the maximization of the expected joint likelihood with respect to θ can be carried out analytically and in the sense of the necessary condition for a maximum.

To ease notation, we will write $\tilde{\theta} := \theta^{(j-1)}$ in the following and denote the expectation of a function f of the unobserved data X under the conditional distribution $p_{\tilde{\theta}}(X|Y)$ as a conditional expectation:

$$\mathbb{E}_{\tilde{\theta}}(f(X)|Y) := \mathbb{E}_{p_{\tilde{\theta}}(X|Y)}(f(X)) = \int f(X)\, p_{\tilde{\theta}}(X|Y)\, dX. \tag{67}$$

With these simplifications, we thus aim to evaluate

$$\frac{\partial}{\partial B}\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln p_{\theta}(X, Y)\right) \text{ and } \frac{\partial}{\partial R}\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln p_{\theta}(X, Y)\right), \tag{68}$$

set the results to zero, and solve for update equations for $B^{(j)}$ and $R^{(j)}$.

We proceed in four steps. (1) We first use the IID data assumption and the linearity of conditional expectations and derivatives to simplify the problem of evaluating the expected data set joint likelihood and its partial derivatives for a single data point $(x^{(i)}, y^{(i)})$. We then (2) evaluate the conditional expectation and (3) evaluate the respective partial derivatives. By capitalizing on the results form the first step, we then (4) evaluate and simplify the ensuing parameter update equations.

Appendix

(1) Expected joint likelihood partial derivatives under IID data assumptions

$$
\begin{aligned}
\frac{\partial}{\partial B} \mathbb{E}_{p_{\tilde{\theta}}(X|Y)} \left(\ln p_\theta(X, Y) \right) &= \frac{\partial}{\partial B} \left(\mathbb{E}_{p_{\tilde{\theta}}(X|Y)} \left(\ln \prod_{i=1}^{n} p_\theta \left(x^{(i)}, y^{(i)} \right) \right) \right) \\
&= \frac{\partial}{\partial B} \left(\mathbb{E}_{p_{\tilde{\theta}}(X|Y)} \left(\sum_{i=1}^{n} \ln p_\theta \left(x^{(i)}, y^{(i)} \right) \right) \right) \\
&= \frac{\partial}{\partial B} \sum_{i=1}^{n} \mathbb{E}_{\prod_{i=1}^{n} p_{\tilde{\theta}} \left(x^{(i)}, y^{(i)} \right)} \left(\ln p_\theta \left(x^{(i)}, y^{(i)} \right) \right) \qquad (69) \\
&= \frac{\partial}{\partial B} \sum_{i=1}^{n} \mathbb{E}_{p_{\tilde{\theta}} \left(x^{(i)}, y^{(i)} \right)} \left(\ln p_\theta \left(x^{(i)}, y^{(i)} \right) \right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial B} \mathbb{E}_{p_{\tilde{\theta}} \left(x^{(i)}, y^{(i)} \right)} \left(\ln p_\theta \left(x^{(i)}, y^{(i)} \right) \right),
\end{aligned}
$$

$$
\frac{\partial}{\partial R} \mathbb{E}_{p_{\tilde{\theta}}(X|Y)} \left(\ln p_\theta(X, Y) \right) = \sum_{i=1}^{n} \frac{\partial}{\partial R} \mathbb{E}_{p_{\tilde{\theta}} \left(x^{(i)}, y^{(i)} \right)} \left(\ln p_\theta \left(x^{(i)}, y^{(i)} \right) \right). \qquad (70)
$$

(2) Expected joint likelihood for a single data point

For ease of notation, we omit the (i) superscript indexing the data realizations in this step.

$$
\begin{aligned}
&\mathbb{E}_{p_{\tilde{\theta}}(x|y)}\left(\ln p_\theta(x,y)\right) \\
&= \mathbb{E}_{\tilde{\theta}}\left(\ln p_\theta(x,y)|y\right) \\
&= \mathbb{E}_{\tilde{\theta}}\left(\ln\left(N(y;Bx,R)N(x;0,I_k)\right)|y\right) \\
&= \mathbb{E}_{\tilde{\theta}}\left(\ln N(y;Bx,R) + \ln N(x;0,I_k)|y\right) \\
&= \mathbb{E}\left(\ln\left((2\pi)^{-\frac{m}{2}}|R|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y-Bx)^T R^{-1}(y-Bx)\right)\right) + \ln\left((2\pi)^{-\frac{k}{2}}|I_k|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}x^T x\right)\right)|y\right) \\
&= \mathbb{E}_{\tilde{\theta}}\left(-\frac{m+k}{2}\ln 2\pi - \frac{1}{2}\ln|R| - \frac{1}{2}\left(y^T R^{-1}y - 2y^T R^{-1}Bx + x^T B^T R^{-1}Bx\right) - \frac{1}{2}\ln 1 - \frac{1}{2}x^T x|y\right) \\
&= \mathbb{E}_{\tilde{\theta}}\left(-\frac{m+k}{2}\ln 2\pi - \frac{1}{2}\ln|R| - \frac{1}{2}y^T R^{-1}y + y^T R^{-1}Bx - \frac{1}{2}x^T B^T R^{-1}Bx - \frac{1}{2}x^T x|y\right) \\
&= \mathbb{E}_{\tilde{\theta}}\left(-\frac{m+k}{2}\ln 2\pi - \frac{1}{2}\ln|R| - \frac{1}{2}y^T R^{-1}y + y^T R^{-1}Bx - \frac{1}{2}\mathrm{tr}\left(B^T R^{-1}Bxx^T\right) - \frac{1}{2}x^T x|y\right) \\
&= -\frac{m+k}{2}\ln 2\pi - \frac{1}{2}\ln|R| - \frac{1}{2}y^T R^{-1}y + y^T R^{-1}B\mathbb{E}_{\tilde{\theta}}(x|y) - \frac{1}{2}\mathrm{tr}\left(B^T R^{-1}B\mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right)\right) - \frac{1}{2}\mathbb{E}_{\tilde{\theta}}\left(x^T x|y\right)
\end{aligned}
$$

where in the 7th equality, we made use of the fact that $x^T Ax = \mathrm{tr}(Axx^T)$ for $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$.

(3) Partial derivatives

To evaluate the partial derivatives of the conditional expected joint likelihood with respect to the matrices B and R, we require the following identities from matrix calculus (z.B. Petersen and Pedersen (2012)):

$$\frac{\partial}{\partial X} A^T X B = A B^T, \frac{\partial}{\partial X} \mathsf{tr}(XA) = A^T, \frac{\partial}{\partial X} \mathsf{tr}(X^T A X B) = A X B + A^T X B^T, \frac{\partial}{\partial X} \ln |X| = \left(X^{-1}\right)^T. \tag{71}$$

We then have

$$\begin{aligned}
\frac{\partial}{\partial B} \mathbb{E}_{p_{\tilde{\theta}}(x|y)} \left(\ln p_\theta(x, y)\right) &= \frac{\partial}{\partial B} \left(y^T R^{-1} B \mathbb{E}_{\tilde{\theta}}(x|y) - \frac{1}{2} \mathsf{tr}\left(B^T R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right)\right)\right) \\
&= \frac{\partial}{\partial B} \left(y^T R^{-1} B \mathbb{E}_{\tilde{\theta}}(x|y)\right) - \frac{1}{2} \frac{\partial}{\partial B} \mathsf{tr}\left(B^T R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right)\right) \\
&= \left(y^T R^{-1}\right)^T \mathbb{E}_{\tilde{\theta}}(x|y)^T - \frac{1}{2} R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right) - \frac{1}{2} \left(R^{-1}\right)^T B \mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right)^T \\
&= R^{-1} y \mathbb{E}_{\tilde{\theta}}(x|y)^T - \frac{1}{2} R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right) - \frac{1}{2} R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right) \\
&= R^{-1} y \mathbb{E}_{\tilde{\theta}}(x|y)^T - R^{-1} B \mathbb{E}_{\tilde{\theta}}\left(xx^T|y\right).
\end{aligned} \tag{72}$$

Similarly, we have

$$
\begin{aligned}
&\frac{\partial}{\partial R^{-1}} \mathbb{E}_{p_{\tilde{\theta}}(x|y)} \left(\ln p_\theta(x, y) \right) \\
&= \frac{\partial}{\partial R^{-1}} \left(-\frac{1}{2} \ln |R| - \frac{1}{2} y^T R^{-1} y + y^T R^{-1} B \mathbb{E}_{\tilde{\theta}}(x|y) - \frac{1}{2} \mathsf{tr} \left(B^T R^{-1} B \mathbb{E}_{\tilde{\theta}} \left(x x^T | y \right) \right) \right) \\
&= -\frac{1}{2} \frac{\partial}{\partial R^{-1}} \ln |R| - \frac{1}{2} \frac{\partial}{\partial R^{-1}} y^T R^{-1} y + \frac{\partial}{\partial R^{-1}} y^T R^{-1} C \mathbb{E}_{\tilde{\theta}}(x|y) - \frac{1}{2} \frac{\partial}{\partial R^{-1}} \mathsf{tr} \left(R^{-1} C \mathbb{E}_{\tilde{\theta}} \left(x x^T | y \right) B^T \right) \\
&= \frac{1}{2} R - \frac{1}{2} y y^T + y \left(B \mathbb{E}_{\tilde{\theta}}(x|y) \right)^T - \frac{1}{2} \left(B \mathbb{E}_{\tilde{\theta}} \left(x x^T | y \right) B^T \right)^T \\
&= \frac{1}{2} R - \frac{1}{2} y y^T + y \mathbb{E}_{\tilde{\theta}}(x|y)^T B^T - \frac{1}{2} B \mathbb{E}_{\tilde{\theta}} \left(x x^T | y \right) B^T .
\end{aligned}
$$

(4) Parameter update equations

Re-substitution then yields for the partial derivative with respect to B

$$\begin{aligned}
\frac{\partial}{\partial B}\mathbb{E}_{p_{\tilde{\theta}}(X|Y)}\left(\ln p_\theta(X,Y)\right) &= \sum_{i=1}^{n}\frac{\partial}{\partial B}\mathbb{E}_{p_{\tilde{\theta}}\left(x^{(i)},y^{(i)}\right)}\left(\ln p_\theta\left(x^{(i)},y^{(i)}\right)\right) \\
&= \sum_{i=1}^{n} R^{-1}{y^{(i)}}\mathbb{E}_{\tilde{\theta}}\left(x^{(i)}|y^{(i)}\right)^T - R^{-1}B\mathbb{E}_{\tilde{\theta}}\left(x^{(i)}{x^{(i)}}^T|y^{(i)}\right) \\
&= R^{-1}\sum_{i=1}^{n} y^{(i)}\mathbb{E}_{\tilde{\theta}}\left(x^{(i)}|y^{(i)}\right)^T - R^{-1}B\sum_{i=1}^{n}\mathbb{E}_{\tilde{\theta}}\left(x^{(i)}{x^{(i)}}^T|y^{(i)}\right).
\end{aligned} \tag{73}$$

Setting to zero and solving for $B^{(j)}$ then yields

$$R^{-1} B^{(j)} \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) = R^{-1} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T .$$

$$\Leftrightarrow B^{(j)} = \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T \left(\sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) \right)^{-1} . \tag{74}$$

Similarly, re-substitution yields for the partial derivative with respect to R

$$
\begin{aligned}
&\frac{\partial}{\partial R^{-1}} \mathbb{E}_{p_{\tilde{\theta}}(X|Y)} \left(\ln p_\theta(X, Y) \right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial R^{-1}} \mathbb{E}_{p_{\tilde{\theta}}\left(x^{(i)}, y^{(i)}\right)} \left(\ln p_\theta \left(x^{(i)}, y^{(i)} \right) \right) \\
&= \sum_{i=1}^{n} \frac{1}{2} R - \frac{1}{2} y^{(i)} y^{(i)T} + y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T - \frac{1}{2} B \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)T} | y^{(i)} \right) B^T \\
&= \frac{n}{2} R - \frac{1}{2} \sum_{i=1}^{n} y^{(i)} y^{(i)T} + \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T - \frac{1}{2} B \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)T} | y^{(i)} \right) B^T .
\end{aligned}
\tag{75}
$$

Beweis (fortgeführt)

Setting to zero and solving for $R^{(j)}$ then yields

$$\frac{n}{2} R^{(j)} = \frac{1}{2} \sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T + \frac{1}{2} B \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) B^T$$

$$R^{(j)} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \frac{2}{n} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^T + \frac{1}{n} B \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) B^T$$

Substitution of the update equation for B further yields

$$
\begin{aligned}
R^{(j)} =& \frac{1}{n} \sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \frac{2}{n} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(j)^T} \\
&+ \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T \left(\sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) \right)^{-1} \sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) B^{(j)^T} \\
=& \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)^T} - 2 \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(j)^T} + \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(j)^T} \right) \\
=& \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(j)^T} \right)
\end{aligned}
$$

We thus obtained the parameter update equations

$$B^{(j)} = \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T \left(\sum_{i=1}^{n} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) \right)^{-1}$$

$$R^{(j)} = \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)^T} - \sum_{i=1}^{n} y^{(i)} \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right)^T B^{(j)^T} \right)$$

The computational forms of these update equations can be further simplified by noting that with

$$\mathbb{E} \left(x x^T | y \right) = \mu_{x|y} \mu_{x|y}^T + \Sigma_{x|y} \tag{76}$$

and the LGM inference theorem it holds that

$$\mathbb{E}_{\tilde{\theta}} \left(x^{(i)} | y^{(i)} \right) = \hat{x}^{(i)} \text{ and } \mathbb{E}_{\tilde{\theta}} \left(x^{(i)} x^{(i)^T} | y^{(i)} \right) = \hat{x}^{(i)} \hat{x}^{(i)^T} + \hat{\Sigma}^{(i)}. \tag{77}$$

We thus obtain

$$B^{(j)} = \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)T} \left(\sum_{i=1}^{n} \hat{x}^{(i)} \hat{x}^{(i)T} + \hat{\Sigma}^{(i)} \right)^{-1}$$

$$R^{(j)} = \frac{1}{n} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)T} - \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)T} B^{(j)T} \right).$$

Finally, enforcing the diagonality constraint on R can be achieved by setting

$$R^{(j)} := -\frac{1}{n} \mathsf{diag} \left(\sum_{i=1}^{n} y^{(i)} y^{(i)T} + \sum_{i=1}^{n} y^{(i)} \hat{x}^{(i)T} B^{(j)T} \right) \tag{78}$$

Beweis des Theorems zur direkten Maximum Marginal Likelihood Schätzung des probabilistischen PCA Modells

We only show that $\hat{B}$ as defined in the theorem corresponds to maximum of the log marginal likelihood function. To this end, we closely follow the respective proof in Tipping and Bishop (1999), and proceed in three steps: (1) We first rewrite the marginal data set log likelihood function of the PPCA model in a suitable manner. (2) We then evaluate its gradient with respect to B and (3) finally evaluate the resulting maximum marginal likelihood estimator.

(1) Log likelihood function

We first rewrite the marginal data set log likelihood function of the PPCA model

$$\ell : \Theta \to \mathbb{R}, \theta \mapsto \ell(\theta) := \ln p_\theta(Y) = \sum_{i=1}^{n} \ln N\left(y^{(i)}; 0_m, BB^T + \sigma^2 I_m\right) \tag{79}$$

with $\theta := \{B, \sigma^2\}$ in a way more amenable to direct maximization.

With the definitions of

$$\Sigma := BB^T + \sigma^2 I_m \text{ and } C := \frac{1}{n} \sum_{i=1}^{n} y^{(i)} y^{(i)T} \tag{80}$$

and the trace operator properties

$$x^T A x = \mathrm{tr}(A x x^T) \text{ and } \mathrm{tr}(A) + \mathrm{tr}(B) = \mathrm{tr}(A + B), \tag{81}$$

we have

$$
\begin{aligned}
\ln p_\theta(Y) &= \sum_{i=1}^{n} \ln N\left(y^{(i)}; 0_m, \Sigma\right) \\
&= \sum_{i=1}^{n} \ln\left((2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} {y^{(i)}}^T \Sigma^{-1} y^{(i)}\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \sum_{i=1}^{n} \frac{1}{n} {y^{(i)}}^T \Sigma^{-1} y^{(i)}\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \sum_{i=1}^{n} \mathrm{tr}\left(\Sigma^{-1} \frac{1}{n} y^{(i)} {y^{(i)}}^T\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \mathrm{tr}\left(\sum_{i=1}^{n} \Sigma^{-1} \frac{1}{n} y^{(i)} {y^{(i)}}^T\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \mathrm{tr}\left(\Sigma^{-1} \frac{1}{n} \sum_{i=1}^{n} y^{(i)} {y^{(i)}}^T\right)\right) \\
&= -\frac{n}{2}\left(m \ln 2\pi + \ln|\Sigma| + \mathrm{tr}\left(\Sigma^{-1} C\right)\right).
\end{aligned}
\tag{82}
$$

(2) Gradient of the log marginal likelihood function

With

$$\frac{\partial}{\partial X} \ln |X| = \left(X^{-1}\right)^T, \frac{\partial}{\partial X} X X^T = 2X, \text{ and } \frac{\partial}{\partial X} \text{tr}(XA) = A^T, \tag{83}$$

the gradient of the log marginal likelihood function with respect to B evaluates to

$$\begin{aligned}
\frac{\partial}{\partial B} \ell(\theta) &= -\frac{n}{2} \frac{\partial}{\partial B} \left(m \ln 2\pi + \ln |BB^T + \sigma^2 I_m| + \text{tr}\left(\left(BB^T + \sigma^2 I_m\right)^{-1} C \right) \right) \\
&= -\frac{n}{2} \frac{\partial}{\partial B} \ln |BB^T + \sigma^2 I_m| - \frac{n}{2} \frac{\partial}{\partial B} \text{tr}\left(\left(BB^T + \sigma^2 I_m\right)^{-1} C \right) \\
&= -\frac{n}{2} 2 \left(\left(BB^T + \sigma^2 I_m\right)^{-1} B \right)^T + \frac{n}{2} 2 \left(BB^T + \sigma^2 I_m\right)^{-1} C \left(BB^T + \sigma^2 I_m\right)^{-1} B \\
&= n \left(-\Sigma^{-1} B + \Sigma^{-1} C \Sigma^{-1} B \right) \\
&= n \left(\Sigma^{-1} C \Sigma^{-1} B - \Sigma^{-1} B \right).
\end{aligned} \tag{84}$$

(3) Maximum marginal likelihood estimator evaluation

Setting the gradient of ℓ with respect to B to zero then yields

$$\Sigma^{-1} C \Sigma^{-1} \hat{B} - \Sigma^{-1} \hat{B} = 0 \Leftrightarrow \Sigma^{-1} \hat{B} = \Sigma^{-1} C \Sigma^{-1} \hat{B} \Leftrightarrow \hat{B} = C \Sigma^{-1} \hat{B}. \tag{85}$$

We consider solutions of this necessary condition for a stationary point of the log marginal likelihood function with $B \neq 0$ and $\Sigma \neq C$. To find these, we first express $\hat{B}$ in terms of its singular value decomposition

$$\hat{B} = U L V^T \tag{86}$$

where $U = (u_1, u_2, ..., u_l)$ is an $m \times q$ matrix of orthonormal column vectors, $L = \text{diag}(l_1, l_2, ..., l_q)$ is a $q \times q$ diagonal matrix of singular values, and V is a $q \times q$ orthogonal matrix. Substitution in the necessary condition for a stationary point then yields

$$CUL = U(\sigma^2 I_m + L^2)L. \tag{87}$$

For $l_j \neq 0$, eq. (87) implies that

$$Cu_j = (\sigma^2 + l_j^2)u_j \tag{88}$$

Hence, each column of U must be an eigenvector of C with corresponding eigenvalue $\lambda_j = \sigma^2 + l_j$, and thus

$$\lambda_j = \sigma^2 + l_j^2 \Leftrightarrow l_j^2 = \lambda_j - \sigma^2 \Leftrightarrow l_j = (\lambda_j - \sigma^2)^{\frac{1}{2}}. \tag{89}$$

For $l_j = 0$, u_j is arbitrary. Under the assumption that $l_j \neq 0$ for $j = 1, ..., m$, all potential solutions for $\hat{B}$ can thus be written in the form

$$\hat{B} = U_q \left(\Lambda_q - \sigma^2 I_m \right)^{\frac{1}{2}} R, \tag{90}$$

where U_q is a $m \times q$ matrix whose q columns are the eigenvectors of C, Λ_q is the diagonal matrix of the corresponding eigenvalues, and R is an arbitrary $q \times q$ orthogonal matrix, for example, $R = I_q$.

□

References I

Blei, David M., and Padhraic Smyth. 2017. "Science and Data Science." *Proceedings of the National Academy of Sciences* 114 (33): 8689–92. https://doi.org/10.1073/pnas.1702076114.

Casella, G, and R Berger. 2012. *Statistical Inference*. Duxbury.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Holzinger, K. J., and F. Swineford. 1939. *A Study in Factor Analysis: The Stability of a Bifactor Solution*. Vol. 48. Supplementary Educational Monographs. University of Chicago.

Horvath, Lilla, Stanley Colcombe, Michael Milham, Shruti Ray, Philipp Schwartenbeck, and Dirk Ostwald. 2021. "Human Belief State-Based Exploration and Exploitation in an Information-Selective Symmetric Reversal Bandit Task." *Computational Brain & Behavior*, August. https://doi.org/10.1007/s42113-021-00112-3.

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

Joreskog, K. G. 1969. "A General Approach to Confirmatory Maximum Likelihood Factor Analysis." *Psychometrika* 34: 183–202.

Lawley, N. D. 1953. "A Modified Method of Estimation in Factor Analysis and Some Large Sample Results." *Nord. Psyko. Monogr. Ser* 3: 35–42.

Ostwald, Dirk, Evgeniya Kirilina, Ludger Starke, and Felix Blankenburg. 2014. "A Tutorial on Variational Bayes for Latent Linear Stochastic Time-Series Models." *Journal of Mathematical Psychology* 60 (June): 1–19. https://doi.org/10.1016/j.jmp.2014.04.003.

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.

Petersen, Kaare Brandt, and Michael Syskind Pedersen. 2012. "The Matrixcookbook," 72.

References II

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.

Roweis, Sam. 1998. "EM Algorithms for PCA and SPCA," 7.

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.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64.

Starke, Ludger, and Dirk Ostwald. 2017. "Variational Bayesian Parameter Estimation Techniques for the General Linear Model." *Frontiers in Neuroscience* 11 (September). https://doi.org/10.3389/fnins.2017.00504.

Tipping, Michael E, and Christopher M Bishop. 1999. "Probabilistic Principal Component Analysis," 12.