

1. Eigenwertprobleme:

Gewöhnliches symmetrisches Eigenwertproblem:

Es sei $\mathbf{B} = (p \times p)$ reell und symmetrisch.

Gesucht: $\mathbf{a} = (p \times 1)$ mit $\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$.

Dieses Problem heißt ein symmetrisches Eigenwertproblem. Der Lösungsvektor \mathbf{a} heißt dann ein Eigenvektor von \mathbf{B} , und die reelle Zahl λ heißt der zugehörige Eigenwert.

Hauptachsentransformation einer symmetrischen Matrix (Spektralzerlegung):

Es sei $\mathbf{B} = (p \times p)$ reell und symmetrisch.

Dann existiert eine orthogonale Matrix $\mathbf{R} = (p \times p)$ und eine Diagonalmatrix Λ (beide reell) mit

$$\mathbf{R}^T \mathbf{B} \mathbf{R} = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p) \quad \text{und} \quad \mathbf{R}^T \mathbf{R} = \mathbf{I}_p \quad (\text{da } \mathbf{R} \text{ orthogonal ist}).$$

Die Matrix \mathbf{R} transformiert die Matrix \mathbf{B} auf Diagonalfom (Hauptachsentransformation, Zusammenhang mit Drehung eines Koordinatensystems, sodaß die Koordinatenachsen mit den Hauptachsen einer allgemeinen Ellipse übereinstimmen).

Folgerungen:

1. Die Matrix \mathbf{B} kann dargestellt werden als $\mathbf{B} = \mathbf{R} \Lambda \mathbf{R}^T$ (Spektraldarstellung von \mathbf{B}).
2. Es gilt: $\mathbf{B}\mathbf{R} = \mathbf{R}\Lambda$, d.h. $\mathbf{B}\mathbf{r}_j = \lambda_j \mathbf{r}_j$, wobei \mathbf{r}_j die Spalte j der Matrix \mathbf{R} bezeichnet. Somit sind die Spaltenvektoren \mathbf{r}_j von \mathbf{R} Eigenvektoren zum obigen Eigenwertproblem jeweils mit dem Eigenwert λ_j .

Verallgemeinertes symmetrisches Eigenwertproblem:

Es sei $\mathbf{B} = (p \times p)$ reell und symmetrisch

$\mathbf{W} = (p \times p)$ reell, symmetrisch und positiv definit (d.h. alle Eigenwerte sind positiv).

Gesucht: $\mathbf{a} = (p \times 1)$ mit $\mathbf{B}\mathbf{a} = \lambda \mathbf{W}\mathbf{a}$.

Dieses Problem heißt ein verallgemeinertes symmetrisches Eigenwertproblem. Der Lösungsvektor \mathbf{a} heißt dann ein Eigenvektor des verallgemeinerten Eigenwertproblems, und die reelle Zahl λ heißt der zugehörige Eigenwert.

Beachte, daß das verallgemeinerte, symmetrische Eigenwertproblem nicht reduziert werden kann auf ein gewöhnliches symmetrisches Eigenwertproblem durch $\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}$, da die Matrix $\mathbf{W}^{-1}\mathbf{B}$ nicht symmetrisch zu sein braucht, auch wenn \mathbf{B} und \mathbf{W} symmetrisch sind!

Simultane Hauptachsentransformation von zwei symmetrischen Matrizen:

Es sei $\mathbf{B} = (p \times p)$ reell und symmetrisch

$\mathbf{W} = (p \times p)$ reell, symmetrisch und positiv definit (d.h. alle Eigenwerte sind positiv).

Dann existiert eine Matrix $\mathbf{A} = (p \times p)$ und eine Diagonalmatrix Λ (beide reell) mit

$$\mathbf{A}^T \mathbf{W} \mathbf{A} = \mathbf{I}_p \quad (\mathbf{I}_p = \text{Einheitsmatrix})$$

$$\mathbf{A}^T \mathbf{B} \mathbf{A} = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p).$$

Die Matrix \mathbf{A} transformiert also beide Matrizen \mathbf{W} und \mathbf{B} gleichzeitig auf Diagonalform (daher simultane Hauptachsentransformation). Die Matrix \mathbf{A} ist aber im allgemeinen nicht orthogonal.

Folgerungen:

Es gilt

$$\mathbf{B} \mathbf{A} = \mathbf{A}^{-T} \Lambda = \mathbf{A}^{-T} \underbrace{\mathbf{A}^T \mathbf{W} \mathbf{A}}_{=\mathbf{I}_p} \Lambda = \mathbf{W} \mathbf{A} \Lambda,$$

d.h. $\mathbf{B} \mathbf{a}_j = \lambda_j \mathbf{W} \mathbf{a}_j$, wobei \mathbf{a}_j die Spalte j der Matrix \mathbf{A} bezeichnet. Somit sind die Spaltenvektoren \mathbf{a}_j von \mathbf{A} Eigenvektoren zum obigen verallgemeinerten Eigenwertproblem jeweils mit dem Eigenwert λ_j . Falls die Matrix \mathbf{B} positiv semidefinit ist, so sind alle Eigenwerte $\lambda_j \geq 0$.

Herleitung der simultanen Hauptachsentransformation zu \mathbf{B} und \mathbf{W} :

1. Spektraldarstellung von \mathbf{W} : $\mathbf{W} = \mathbf{R} \mathbf{D} \mathbf{R}^T$ mit $\mathbf{D} = \text{diag}(d_1, \dots, d_p)$, alle $d_j > 0$.
2. Setze $\mathbf{W}^{-1/2} = \mathbf{R} \mathbf{D}^{-1/2} \mathbf{R}^T$ wobei $\mathbf{D}^{-1/2} = \text{diag}(d_1^{-1/2}, \dots, d_p^{-1/2})$. Die Matrix $\mathbf{W}^{-1/2}$ ist reell und symmetrisch.
3. Setze $\mathbf{P} = \mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2}$. Die Matrix \mathbf{P} ist reell und symmetrisch.
4. Spektralzerlegung von \mathbf{P} : $\mathbf{P} = \mathbf{S} \Lambda \mathbf{S}^T$ mit $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$. Falls die Matrix \mathbf{B} positiv semidefinit ist, so ist auch \mathbf{P} positiv semidefinit, und somit sind dann alle Eigenwerte $\lambda_j \geq 0$.
5. Setze $\mathbf{A} = \mathbf{W}^{-1/2} \mathbf{S}$. Dann gilt

$$\mathbf{A}^T \mathbf{W} \mathbf{A} = \mathbf{S}^T \mathbf{W}^{-1/2} \mathbf{W} \mathbf{W}^{-1/2} \mathbf{S} = \mathbf{I}_p.$$

$$\mathbf{A}^T \mathbf{B} \mathbf{A} = \mathbf{S}^T \underbrace{\mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2}}_{=\mathbf{P} = \mathbf{S} \Lambda \mathbf{S}^T} \mathbf{S} = \Lambda.$$

2. Diskriminanzanalyse

Eine Gesamtheit (Population) G mit n Objekten (Untersuchungseinheiten) sei zerlegt in g Teilpopulationen: $G = P_1 \cup \dots \cup P_g$. Die Teilpopulation (Gruppe) P_r besitze n_r Elemente ($r = 1, \dots, g$), wobei $n_1 + \dots + n_g = n$. Bei jedem Objekt der Gesamtheit werden die Variablen (Untersuchungsmerkmale) x_1, \dots, x_p beobachtet. Die Datenmatrix für die Teilpopulation P_r hat die Form

α_i	x_1	\dots	x_p
α_1	x_{11}	\dots	x_{1p}
\vdots	\vdots		\vdots
α_{n_r}	$x_{n_r 1}$	\dots	$x_{n_r p}$

Ziel: Gegeben sei ein neues Objekt γ mit den Merkmalsausprägungen $\mathbf{x}^\gamma = (x_1^\gamma, \dots, x_p^\gamma)^\top$, das zur Gesamtheit G gehört, dessen Populationszugehörigkeit aber unbekannt ist. Gesucht ist eine optimale Zuordnungsvorschrift, die auf dem Datenvektor \mathbf{x}^γ und den Datenmatrizen der Populationen P_1, \dots, P_g beruht, und die angibt, welcher Population das neue Objekt zugeordnet werden soll.

Lineare Diskriminanzfunktion (nach R.A. Fisher):

Es sei $y = \mathbf{a}^\top \mathbf{x} = \sum_{j=1}^p a_j x_j$ eine Linearkombination der Variablen x_1, \dots, x_p . Die Varianz von y kann zerlegt werden wie folgt:

$$(1) \quad \text{Var}(y) = S_w + S_b$$

$$\text{mit} \quad S_w = \frac{1}{n} \sum_{r=1}^g s_r^2 \quad \text{wobei } s_r^2 = \text{Varianz von } y \text{ in } P_r$$

$$S_b = \frac{1}{n} \sum_{r=1}^g n_r (\bar{y}_r - \bar{y})^2 \quad \text{wobei } \bar{y}_r = \text{arithmetisches Mittel von } y \text{ in } P_r.$$

S_w ist die Komponente innerhalb (within) und S_b die Komponente zwischen (between) den Gruppen. Nun soll das folgende Extremalproblem gelöst werden

$$(2) \quad \mathbf{a} = (a_1, \dots, a_p)^\top \text{ so wählen, daß } \frac{S_b}{S_w} = \max.$$

Die Linearkombination $y = \mathbf{a}^\top \mathbf{x}$ heißt dann "lineare Diskriminanzfunktion". Im Falle von zwei Gruppen lautet nun die Zuordnungsvorschrift wie folgt:

Das neue Objekt γ besitze den Datenvektor \mathbf{x}^γ .

Berechne für das neue Objekt den Wert der linearen Diskriminanzfunktion $y^\gamma = \mathbf{a}^\top \mathbf{x}^\gamma$.

Ordne das neue Objekt γ der Population P_1 zu, falls $|y^\gamma - \bar{y}_1| < |y^\gamma - \bar{y}_2|$.

Der Idealfall $S_w = 0$ und $S_b > 0$ bedeutet, daß die lineare Diskriminanzfunktion für alle Objekte der Population P_1 und P_2 jeweils denselben Wert annimmt, und daß die beiden Werte verschieden sind.

Lösung des Extremalproblems:

Es sei \mathbf{C} die Kovarianzmatrix der Gesamtheit G und \mathbf{C}_r die Kovarianzmatrix der Population P_r :

$$(3) \quad \mathbf{C} = E(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^\top \quad \text{mit} \quad E(\cdots) = \frac{1}{n} \sum_{i=1}^n \cdots$$

$$(4) \quad \mathbf{C}_r = E_r(\mathbf{x} - \bar{\mathbf{x}}_r)(\mathbf{x} - \bar{\mathbf{x}}_r)^\top \quad \text{mit} \quad E_r(\cdots) = \frac{1}{n_r} \sum_{i \in P_r} \cdots$$

Dann gilt die Zerlegung (analog zur Varianzzerlegung):

$$(5) \quad \mathbf{C} = \mathbf{C}_w + \mathbf{C}_b$$

$$\text{mit} \quad \mathbf{C}_w = \frac{1}{n} \sum_{r=1}^g n_r \mathbf{C}_r$$

$$\mathbf{C}_b = \frac{1}{n} \sum_{r=1}^g n_r (\bar{x}_r - \bar{x})(\bar{x}_r - \bar{x})^\top.$$

Wenn $y = \mathbf{a}^\top \mathbf{x}$ ist, dann gilt

$$(6) \quad \text{Var}(y) = \mathbf{a}^\top \mathbf{C} \mathbf{a} = \underbrace{\mathbf{a}^\top \mathbf{C}_w \mathbf{a}}_{=S_w} + \underbrace{\mathbf{a}^\top \mathbf{C}_b \mathbf{a}}_{=S_b}$$

und damit lautet unser Extremalproblem:

$$(7) \quad \mathbf{a} = (a_1, \dots, a_p)^\top \quad \text{so wählen, daß} \quad \frac{\mathbf{a}^\top \mathbf{C}_b \mathbf{a}}{\mathbf{a}^\top \mathbf{C}_w \mathbf{a}} = \max.$$

Als notwendige Bedingung für eine Extremalstelle finden wir

$$(8) \quad \mathbf{C}_b \mathbf{a} = \lambda \mathbf{C}_w \mathbf{a} \quad \text{mit} \quad \lambda = \frac{\mathbf{a}^\top \mathbf{C}_b \mathbf{a}}{\mathbf{a}^\top \mathbf{C}_w \mathbf{a}}.$$

Der gesuchte Vektor \mathbf{a} ist also der Eigenvektor zum größten Eigenwert dieses verallgemeinerten symmetrischen Eigenwertproblems.

Vollständiger Satz von kanonischen Variablen:

Es sei s die Anzahl der positiven Eigenwerte von (8):

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s > 0, \lambda_{s+1} = \dots = \lambda_p = 0$$

und $\mathbf{a}_1, \dots, \mathbf{a}_s$ seien die zugehörigen Eigenvektoren. Es ist $s = rk(\mathbf{C}_b) \leq g - 1$. Dann definieren wir die s kanonischen Variablen wie folgt: $y_j = \mathbf{a}_j^\top \mathbf{x}$, $j = 1, \dots, s$. Die kanonischen Variablen sind alle unkorreliert und haben alle die gleiche Varianzkomponente innerhalb der Gruppen, da $\mathbf{A}^\top \mathbf{C}_w \mathbf{A} = \mathbf{I}_p$, wobei $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_s)$. Im allgemeinen Fall von g Gruppen lautet nun die Zuordnungsvorschrift wie folgt:

Das neue Objekt γ besitze den Datenvektor \mathbf{x}^γ .

Berechne für das neue Objekt die Werte der s kanonischen Variablen $y_j^\gamma = \mathbf{a}_j^\top \mathbf{x}^\gamma$.

Ordne das neue Objekt γ jener Population P_r zu, für welche $\sum_{j=1}^s (y_j^\gamma - \bar{y}_{jr})^2 = \min$.

Klassische Diskriminanzanalyse bei Vorliegen einer Normalverteilung:

Wir nehmen jetzt an, daß die Variablen x_1, \dots, x_p eine p -dimensionale Normalverteilung besitzen und zwar

$$\text{in } P_r : (x_1, \dots, x_p) \sim N(\mu_r, \Gamma),$$

d.h. der Vektor $\mathbf{x} = (x_1, \dots, x_p)^T$ besitzt in jeder Population P_r eine p -dimensionale Normalverteilung mit der gleichen Kovarianzmatrix Γ und jeweils verschiedenen Erwartungswertvektoren μ_r . Nun seien n unabhängige Realisationen von $\mathbf{x} = (x_1, \dots, x_p)^T$ gegeben und zwar n_r Realisationen aus P_r , $n_1 + \dots + n_g = n$. Die unbekannt Parameter μ_r und Γ werden wie folgt geschätzt:

$$\hat{\mu}_r = \bar{\mathbf{x}}_r \quad \text{arithmetisches Mittel in Population } P_r$$

$$\hat{\Gamma} = \frac{1}{n} \sum_{r=1}^g n_r \mathbf{C}_r = \mathbf{C}_w \quad \text{Komponente innerhalb der Gruppen (gepoolte Kovarianzmatrix).}$$

Nun lautet die Zuordnungsvorschrift der Maximum-Likelihood-Methode wie folgt:

Das neue Objekt γ besitze den Datenvektor \mathbf{x}^γ .

Berechne für das neue Objekt die Werte der g Dichtefunktionen $f_r(\mathbf{x}^\gamma | \hat{\mu}_r, \hat{\Gamma})$.

Ordne das neue Objekt γ jener Population P_r zu, welche die größte Dichte besitzt.

Diese Zuordnungsregel ist äquivalent zu:

Das neue Objekt γ besitze den Datenvektor \mathbf{x}^γ .

Berechne für das neue Objekt die g Mahalanobis-Abstände

$$d(\mathbf{x}^\gamma, \bar{\mathbf{x}}_r) = (\mathbf{x}^\gamma - \bar{\mathbf{x}}_r)^T \mathbf{C}_w^{-1} (\mathbf{x}^\gamma - \bar{\mathbf{x}}_r).$$

Ordne das neue Objekt γ der Population P_r mit dem kleinsten Mahalanobis-Abstand zu.

Die Zuordnungsvorschrift der Maximum-Likelihood-Methode ist äquivalent zur obigen Zuordnungsvorschrift mit dem vollständigen Satz von kanonischen Variablen.

Bemerkungen zur Diskriminanzanalyse:

1. Klare Zielsetzung und eindeutige mathematische Lösung.
2. Lösung beruht auf dem Mahalanobis-Abstand (multivariate Standardisierung), und sie ist daher unabhängig von der Skalierung der Variablen.
3. Normalverteilungsannahme ist unwesentlich (vgl. kanonische Variablen), wohl aber sollte die Kovarianzmatrix in den einzelnen Populationen annähernd gleich sein.

3. Kanonische Korrelation

Gegeben sind zwei Gruppen von Variablen:

$$x_1, \dots, x_p$$

$$y_1, \dots, y_q$$

mit der Datenmatrix

α_i	x_1	\dots	x_p	y_1	\dots	y_q
α_1	x_{11}	\dots	x_{1p}	y_{11}	\dots	y_{1q}
\vdots	\vdots		\vdots	\vdots		\vdots
α_n	x_{n1}	\dots	x_{np}	y_{n1}	\dots	y_{nq}

1. Schritt: Suche zwei neue Variablen

$$\begin{aligned} \xi &= \mathbf{a}^\top \mathbf{x} \\ \eta &= \mathbf{b}^\top \mathbf{y} \end{aligned} \quad \text{mit} \quad \underbrace{\text{Corr}(\xi, \eta)}_{=r} = \max.$$

Die Kovarianzmatrix des $(p+q)$ -Vektors $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ sei partitioniert in folgender Form

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{\mathbf{xx}} & \mathbf{C}_{\mathbf{xy}} \\ \mathbf{C}_{\mathbf{yx}} & \mathbf{C}_{\mathbf{yy}} \end{pmatrix} = \begin{pmatrix} (p \times p) & (p \times q) \\ (q \times p) & (q \times q) \end{pmatrix}.$$

Dann gilt

$$r^2 = \frac{(\text{Cov}(\xi, \eta))^2}{(\text{Var} \xi)(\text{Var} \eta)} = \frac{(\mathbf{a}^\top \mathbf{C}_{\mathbf{xy}} \mathbf{b})^2}{(\mathbf{a}^\top \mathbf{C}_{\mathbf{xx}} \mathbf{a})(\mathbf{b}^\top \mathbf{C}_{\mathbf{yy}} \mathbf{b})},$$

und als notwendige Bedingungen für eine Extremalstelle von r^2 haben wir

$$(1) \quad \mathbf{C}_{\mathbf{xy}} \mathbf{C}_{\mathbf{yy}}^{-1} \mathbf{C}_{\mathbf{yx}} \mathbf{a} = r^2 \mathbf{C}_{\mathbf{xx}} \mathbf{a}$$

$$(2) \quad \mathbf{C}_{\mathbf{yx}} \mathbf{C}_{\mathbf{xx}}^{-1} \mathbf{C}_{\mathbf{xy}} \mathbf{b} = r^2 \mathbf{C}_{\mathbf{yy}} \mathbf{b}$$

sowie

$$(3) \quad \mathbf{b} = \frac{1}{r} \mathbf{C}_{\mathbf{yy}}^{-1} \mathbf{C}_{\mathbf{yx}} \mathbf{a}.$$

Es sei \mathbf{a} der Eigenvektor zu (1) mit dem größten Eigenwert r^2 , und \mathbf{b} sei definiert gemäß (3).

Dann bilden $(\mathbf{a}, \mathbf{b}, r^2)$ eine Lösung des obigen Extremalproblems.

4. Hauptkomponenten (Principal Components)

Gegeben sind die Variablen x_1, \dots, x_p mit der Datenmatrix

α_j	x_1	\dots	x_p
α_1	x_{11}	\dots	x_{1p}
\vdots	\vdots		\vdots
α_n	x_{n1}	\dots	x_{np}

Erste Hauptkomponente:

- (9) Suche eine neue Variable $\mathbf{y} = \mathbf{a}^\top \mathbf{x}$ mit der größtmöglichen Varianz unter der Nebenbedingung $\mathbf{a}^\top \mathbf{a} = 1$.

Wenn \mathbf{C} die Kovarianzmatrix der Variablen x_1, \dots, x_p bezeichnet, dann ist das obige Extremalproblem äquivalent zu:

- (10) Wähle den Vektor $\mathbf{a} = (a_1, \dots, a_p)$ so, daß $\frac{\mathbf{a}^\top \mathbf{C} \mathbf{a}}{\mathbf{a}^\top \mathbf{a}} = \max$.

Dies führt zum Eigenwertproblem

- (11) $\mathbf{C} \mathbf{a} = \lambda \mathbf{a}$ wobei $\lambda = \frac{\mathbf{a}^\top \mathbf{C} \mathbf{a}}{\mathbf{a}^\top \mathbf{a}}$.

Der gesuchte Vektor \mathbf{a} ist also der normierte ($\mathbf{a}^\top \mathbf{a} = 1$) Eigenvektor zum größten Eigenwert λ .

Die zugehörige Variable $\mathbf{y} = \mathbf{a}^\top \mathbf{x}$ heißt erste Hauptkomponente und sie hat die Varianz λ .

Vollständiger Satz aller Hauptkomponenten:

- (12) Suche neue Variablen $\mathbf{y} = \mathbf{R}^\top \mathbf{x}$, sodaß y_1, \dots, y_p unkorreliert sind unter der Nebenbedingung $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_p$.

Die Lösung ist gegeben durch die Hauptachsentransformation \mathbf{R} :

- (13) $\mathbf{R}^\top \mathbf{C} \mathbf{R} = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ mit $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_p$.

Die zugehörigen Variablen y_1, \dots, y_p mit $\mathbf{y} = \mathbf{R}^\top \mathbf{x}$ sind unkorreliert mit den Varianzen $\lambda_1, \dots, \lambda_p$, und sie bilden den vollständigen Satz aller Hauptkomponenten.

Bemerkungen:

- Es gilt $\sum_{j=1}^p \text{Var}(x_j) = \sum_{j=1}^p \text{Var}(y_j) = \sum_{j=1}^p \lambda_j$, da die Spur bei einer orthogonalen Transformation invariant bleibt.
- Die Hauptkomponenten und ihre Varianzen $\lambda_1, \dots, \lambda_p$ sind *nicht* skaleninvariant.

5. Faktorenanalyse

Gegeben sind die Variablen x_1, \dots, x_p mit der Datenmatrix

α_i	x_1	\dots	x_p
α_1	x_{11}	\dots	x_{1p}
\vdots	\vdots		\vdots
α_n	x_{n1}	\dots	x_{np}

Frage: Kann die Abhängigkeit zwischen den Variablen x_1, \dots, x_p erklärt werden durch wenige latente (verborgene) gemeinsame Faktoren?

Faktorenmodell:

$$(14) \quad \mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{e}$$

wobei $\mathbf{x} = (p \times 1)$, $\mathbf{A} = (p \times m)$, $\mathbf{f} = (m \times 1)$ und $\mathbf{e} = (p \times 1)$:

x_1, \dots, x_p beobachtbare Variablen

f_1, \dots, f_m gemeinsame Faktoren (latent d.h. nicht direkt beobachtbar)

e_1, \dots, e_p spezifische Faktoren (latent)

$\mathbf{A} = (a_{jk}) = (p \times m)$ Ladungsmatrix (Koeffizienten der gemeinsamen Faktoren)

Modellannahmen:

Die wesentliche Modellannahme des Faktorenmodells lautet:

Alle $m + p$ Faktoren sind unkorreliert (bzw. normalverteilt und unabhängig).

Da wir uns bei der Faktorenanalyse für die Abhängigkeitsstruktur (Kovarianzen) der Variablen interessieren und nicht für die Mittelwerte, und da das Faktorenmodell im wesentlichen invariant bleibt bei einer Neuskalierung der Variablen, treffen wir die folgenden Annahmen:

1. alle Variablen x_1, \dots, x_p sind standardisiert (arithmetisches Mittel = 0, Varianz = 1).
2. alle $m + p$ Faktoren sind unkorreliert (bzw. normalverteilt und unabhängig)
3. die gemeinsamen Faktoren f_1, \dots, f_m sind standardisiert
4. die spezifischen Faktoren e_1, \dots, e_p haben die Varianzen u_1^2, \dots, u_p^2

Folgerungen aus dem Faktorenmodell:

Wenn \mathbf{C} die Kovarianzmatrix der Variablen x_1, \dots, x_p bezeichnet, dann folgt aus den Annahmen zum Faktorenmodell: $\mathbf{C} = \mathbf{A}\mathbf{A}^T + \mathbf{U}^2$, wobei $\mathbf{U}^2 = \text{diag}(u_1^2, \dots, u_p^2)$.

Rotationsproblem:

Es sei $\mathbf{R} = (m \times m)$ eine beliebige orthogonale Matrix und $\tilde{\mathbf{f}} = \mathbf{R}^T \mathbf{f}$. Dann sind die Komponenten $\tilde{f}_1, \dots, \tilde{f}_m$ wieder unkorreliert (bzw. normalverteilt und unabhängig) und es gilt:

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{e} = \underbrace{\mathbf{A}\mathbf{R}}_{\tilde{\mathbf{A}}} \underbrace{\mathbf{R}^T \mathbf{f}}_{\tilde{\mathbf{f}}} + \mathbf{e} = \tilde{\mathbf{A}}\tilde{\mathbf{f}} + \mathbf{e}, \text{ d.h. auch in den Variablen } \tilde{f}_1, \dots, \tilde{f}_m \text{ haben wir wieder}$$

ein Faktorenmodell mit der Ladungsmatrix $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R}$, für das alle vier obigen Annahmen erfüllt sind. Die Faktoren f_1, \dots, f_m und die Ladungsmatrix \mathbf{A} sind durch das Faktorenmodell (14) mit den vier Annahmen also nicht eindeutig festgelegt. Dieser Spielraum (Modellparameter nicht eindeutig identifizierbar) wird ausgenutzt zur Erleichterung der Interpretation (Rotationsmatrix so wählen, daß ein gewünschtes Faktormuster herauskommt!?)

Maximum-Likelihood-Methode:

Annahme: $(x_1, \dots, x_p) \sim N_p(0, \Gamma)$.

$$(15) \quad \hat{\mathbf{A}}, \hat{\mathbf{U}}^2 \text{ so: } L(\hat{\Gamma} | \mathbf{X}) = \max, \text{ wobei } \hat{\Gamma} = \hat{\mathbf{A}}\hat{\mathbf{A}}^T + \hat{\mathbf{U}}^2 \text{ und } L = \text{Likelihood-Funktion.}$$

Die Bedingung (15) ist äquivalent zu

$$(16) \quad \hat{\mathbf{A}}, \hat{\mathbf{U}}^2 \text{ so: } d(\hat{\Gamma}, \mathbf{C}) = \min,$$

wobei

$$(17) \quad d(\hat{\Gamma}, \mathbf{C}) = \text{tr}(\hat{\Gamma}^{-1}\mathbf{C}) - \ln \det(\hat{\Gamma}^{-1}\mathbf{C}) - p.$$

Es gilt $d(\hat{\Gamma}, \mathbf{C}) \geq 0$, und die "Abstandsfunktion" $d(\hat{\Gamma}, \mathbf{C})$ ist symmetrisch in den beiden Argumenten. Eine rein algebraische Lösung zur Maximum-Likelihood-Methode ist nicht bekannt, man benötigt numerische Iterationsverfahren. Falls die Lösung am Rande des zulässigen Bereichs liegt (eine spezifische Varianz wird Null), so spricht man vom Heywood-Fall.

Varianten zur Maximum-Likelihood-Methode:

Anstelle der "Abstandsfunktion" (17) werden u.a. die folgenden Varianten vorgeschlagen:

- ungewogene kleinste Quadrate:

$$d(\hat{\Gamma}, \mathbf{C}) = \text{tr}(\hat{\Gamma} - \mathbf{C})^2 = \sum_{j,k=1}^p (\hat{\gamma}_{jk} - c_{jk})^2$$

- verallgemeinerte kleinste Quadrate:

$$d(\hat{\Gamma}, \mathbf{C}) = \text{tr}(\hat{\Gamma}^{-1}\mathbf{C} - \mathbf{I})^2.$$

Hauptkomponentenmethode:

Es sei

\mathbf{C}	die empirische Kovarianzmatrix zu x_1, \dots, x_p
$\mathbf{C} = \mathbf{R} \mathbf{\Lambda} \mathbf{R}^T$	die Spektraldarstellung von \mathbf{C}
$\mathbf{y} = \mathbf{R}^T \mathbf{x}$	der Vektor der Hauptkomponenten
$\mathbf{x} = \mathbf{R} \mathbf{y}$	die Darstellung der x -Variablen als Linearkombination der Hauptkomponenten
$\mathbf{x} = \underbrace{\mathbf{R} \mathbf{\Lambda}^{1/2}}_{\mathbf{A}} \underbrace{\mathbf{\Lambda}^{-1/2} \mathbf{y}}_{\mathbf{f}}$	die Darstellung der x -Variablen als Linearkombination aller standardisierten Hauptkomponenten

Nun sollen nur die wichtigsten Hauptkomponenten zur "Erklärung" der x -Variablen herangezogen werden, d.h. in der letzten Gleichung sollen nur die ersten m Faktoren auf der rechten Seite berücksichtigt werden. Dann ergibt sich das Faktorenmodell der Hauptkomponentenmethode:

$$\mathbf{x} = (\mathbf{A}_1 | \mathbf{A}_2) \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix} = \mathbf{A}_1 \mathbf{f}_1 + \mathbf{e}$$

wobei

$$\mathbf{A}_1 = \text{erste } m \text{ Spalten von } \mathbf{A} = \mathbf{R} \mathbf{\Lambda}^{1/2}$$

$$\mathbf{f}_1 = \text{erste } m \text{ Komponenten des Vektors } \mathbf{f} = \mathbf{\Lambda}^{-1/2} \mathbf{y}$$

$$\mathbf{e} = \mathbf{A}_2 \mathbf{f}_2.$$

Hier sind zwar alle gemeinsamen Faktoren f_1, \dots, f_m (= Komponenten von \mathbf{f}_1) unkorreliert und standardisiert, und weiter besteht keine Korrelation zwischen den gemeinsamen und den spezifischen Faktoren (alle Komponenten von \mathbf{f}_1 und \mathbf{f}_2 sind unkorreliert), aber die spezifischen Faktoren e_1, \dots, e_p sind i.a. nicht unkorreliert.

Schlußbemerkungen:

- Das Faktorenmodell $\mathbf{x} = \mathbf{A} \mathbf{f} + \mathbf{e}$ ist ein fragwürdiges Modell, da die Faktoren \mathbf{f} und die Ladungsmatrix \mathbf{A} bei diesem Modell nicht eindeutig identifizierbar sind (Rotationsproblem, fragwürdige Freiheit).
- Die Maximum-Likelihood-Methode beruht auf der Normalverteilungsannahme; es gibt verschiedene weitere Schätzmethode, die auf Abstandsmaßen zwischen Matrizen beruhen.
- Die Hauptkomponenten-Methode führt zu keinem korrekten Faktorenmodell (spezifische Faktoren nicht unkorreliert); weiter ist diese Methode nicht skaleninvariant, während das Faktorenmodell skaleninvariant ist.
- Eine einfache Clusteranalyse (Variablen als Objekte, Betrag des Korrelationskoeffizienten zwischen den Variablen als Ähnlichkeitsmaß) führt oft zu ähnlichen Ergebnissen ohne fragwürdige Annahmen.

6. Clusteranalyse

Gegeben sind die Variablen x_1, \dots, x_p mit der Datenmatrix

α_i	x_1	\dots	x_p
α_1	x_{11}	\dots	x_{1p}
\vdots	\vdots		\vdots
α_n	x_{n1}	\dots	x_{np}

Ziele:

1. Clusterbildung mit Objekten: Die Gesamtheit $G = \{\alpha_1, \dots, \alpha_n\}$ der n Objekte (Untersuchungseinheiten) soll in Teilgesamtheiten (Cluster, Gruppen) mit möglichst ähnlichen Objekten zerlegt werden. Die Anzahl der Teilgesamtheiten ist i.a. nicht vorgegeben.
2. Clusterbildung mit Variablen: Hier soll die Gesamtheit $G = \{x_1, \dots, x_p\}$ der p Variablen (Untersuchungsmerkmale) in möglichst homogene Cluster zerlegt werden.

Wir beschreiben im folgenden vor allem die Clusterbildung mit Objekten.

Hierarchische Clusterverfahren:

1. Divisive Verfahren: Die Zerlegung wird schrittweise verfeinert, d.h. die Anzahl der Cluster wird schrittweise erhöht: $1 \rightarrow 2 \rightarrow \dots \rightarrow n-1 \rightarrow n$. (Kaum verfügbar in Programmpaketen.)
2. Agglomerative Verfahren: Die Zerlegung wird schrittweise vergrößert, d.h. die Anzahl der Cluster wird schrittweise reduziert: $n \rightarrow n-1 \rightarrow \dots \rightarrow 2 \rightarrow 1$. (Übliche Verfahren in Programmpaketen.)

Abstandsmaße zwischen Objekten:

Objekt α mit Merkmalsausprägungen x_1, \dots, x_p (Zeile der Datenmatrix)

Objekt β mit Merkmalsausprägungen y_1, \dots, y_p (Zeile der Datenmatrix)

1. Euklidischer Abstand:
$$d(\alpha, \beta) = \sqrt{\sum_{j=1}^p (x_j - y_j)^2}$$
2. Quadrierter euklidischer Abstand:
$$d(\alpha, \beta) = \sum_{j=1}^p (x_j - y_j)^2$$
3. Tschebyscheff-Abstand:
$$d(\alpha, \beta) = \max_j |x_j - y_j|$$
4. City-Block-Abstand:
$$d(\alpha, \beta) = \sum_{j=1}^p |x_j - y_j|$$
5. Minkowsky-Abstand:
$$d(\alpha, \beta) = \sqrt[r]{\sum_{j=1}^p |x_j - y_j|^r} \quad (r \text{ wählbar})$$
6. Benutzerdefiniert (customized):
$$d(\alpha, \beta) = \sqrt[r]{\sum_{j=1}^p |x_j - y_j|^s} \quad (r, s \text{ wählbar})$$

Abstandsmaße zwischen Variablen:

Variable x mit Merkmalsausprägungen x_1, \dots, x_n (Spalte der Datenmatrix)

Variable y mit Merkmalsausprägungen y_1, \dots, y_n (Spalte der Datenmatrix)

1. Korrelationskoeffizient von Pearson:

$$d(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Es gilt: $-1 \leq d(x, y) \leq 1$.

2. Betrag des Korrelationskoeffizienten von Pearson:

$$d(x, y) = \frac{|\sigma_{xy}|}{\sigma_x \sigma_y} = \frac{\left| \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right|}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Es gilt: $0 \leq d(x, y) \leq 1$.

3. Kosinus-Abstand (bei SPSS):

$$d(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} \quad (= \text{Korrelationskoeffizient von Pearson ohne Zentrierung!})$$

Es gilt: $-1 \leq d(x, y) \leq 1$.

4. Betrag des Kosinus-Abstandes (bei SPSS):

$$d(x, y) = \frac{\left| \sum_{i=1}^n x_i y_i \right|}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$$

Es gilt: $0 \leq d(x, y) \leq 1$.

Grundsätzliches Vorgehen bei agglomerativen Cluster-Verfahren (mit Objekten):

1. Jedes Objekt bildet einen Cluster. Die Abstände zwischen den n Clustern entsprechen den Abständen zwischen den n Objekten.

2. Die nächsten beiden Cluster C_k und C_l werden zu einem neuen Cluster C_m verschmolzen:

$$d(C_k, C_l) = \min_{r,s} d(C_r, C_s) \quad \Rightarrow \quad C_m = C_k \cup C_l.$$

3. Die Abstandstabelle muß aktualisiert werden: Die beiden Cluster C_k und C_l werden aus der Abstandstabelle gestrichen und der neue Cluster C_m wird eingefügt. Die Abstände zwischen dem neuen Cluster C_m und den übrigen Clustern C_j muß berechnet werden nach einer der unten beschriebenen Berechnungsmethoden (linkage method).

4. Zurück zu Punkt 2, bis alle Cluster verschmolzen sind zur Gesamtheit $G = \{\alpha_1, \dots, \alpha_n\}$.

Berechnung des Abstands zwischen zwei Clustern (linkage methods):

Zur Notation: Mit $d(\cdot, \cdot)$ bezeichnen wir

- a) die Abstandsfunktion zwischen Objekten $d(\alpha_j, \alpha_k)$;
- b) die Abstandsfunktion zwischen Datenzeilen $d(\bar{x}_1, \bar{x}_2)$, definiert wie bei Objekten;
- c) die Abstandsfunktion zwischen Clustern $d(C_1, C_2)$, definiert mittels $d(\alpha_j, \alpha_k)$.

1. Methode des nächsten Nachbarn (single linkage):

$$d(C_1, C_2) = \min_{\substack{\alpha_j \in C_1 \\ \alpha_k \in C_2}} d(\alpha_j, \alpha_k).$$

2. Methode des entferntesten Nachbarn (complete linkage):

$$d(C_1, C_2) = \max_{\substack{\alpha_j \in C_1 \\ \alpha_k \in C_2}} d(\alpha_j, \alpha_k).$$

3. Durchschnittlicher paarweiser Abstand zwischen den Clustern (between groups linkage):

$$d(C_1, C_2) = \text{durchschnittlicher paarweiser Abstand zwischen } C_1 \text{ und } C_2.$$

4. Durchschnittlicher paarweiser Abstand innerhalb des neuen Clusters (within groups linkage):

$$d(C_1, C_2) = \text{durchschnittlicher paarweiser Abstand innerhalb } C_1 \cup C_2.$$

5. Zentroid-Methode:

$$d(C_1, C_2) = d(\bar{x}_1, \bar{x}_2),$$

wobei \bar{x}_i = Vektor der arithmetischen Mittel in Cluster i . Hier kann das "Inversionsproblem" auftreten: Nach dem Verschmelzen von zwei Clustern kann der minimale Abstand zwischen dem neuen und den übrigen Clustern kleiner werden als vor dem Verschmelzen!

6. Median-Methode

$$d(C_1, C_2) = d(\tilde{x}_1, \tilde{x}_2),$$

wobei \tilde{x}_i = Vektor der Mediane (Zentralwerte) in Cluster i . Auch hier kann das "Inversionsproblem" auftreten.

7. Methode von Ward:

Es sei

\bar{x}_j = Vektor der arithmetischen Mittel in C_j ,

$$D_j = \sum_{\alpha_i \in C_j} d(\alpha_i, \bar{x}_j) = \text{"Streuung" innerhalb } C_j,$$

wobei $d(\alpha_i, \bar{x}_j)$ den Abstand zwischen der Datenzeile zu α_i und \bar{x}_j bezeichnet,

$$D = \sum_{\text{alle Cluster}} D_j = \text{Gesamtstreuung.}$$

Nun werden jene beiden Cluster verschmolzen, für welche der Zuwachs von D nach der Verschmelzung minimal ist.