Linearkombinationen

Multiple Regression

Vorhersage einer abhängigen Variablen anhand anderer beobachteter Variablen:

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots$$

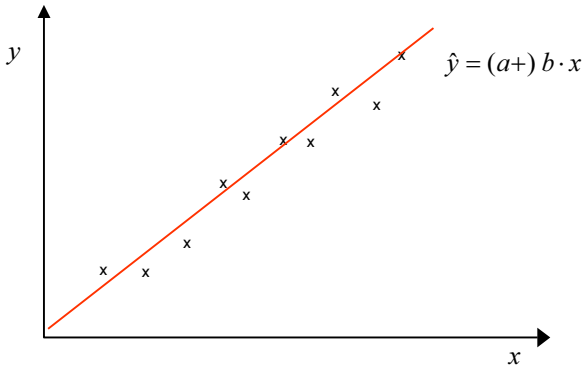
Hauptkomponentenanalyse

Bestimmung von synthetischen Variablen als Linearkombination beobachteter Variablen:

$$HK_1 = c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + \ldots$$
$$HK_2 = d_0 + d_1 \cdot x_1 + d_2 \cdot x_2 + \ldots$$

Dimensionsreduktion

2-d $\rightarrow$ 1-d



Hauptkomponentenanalyse

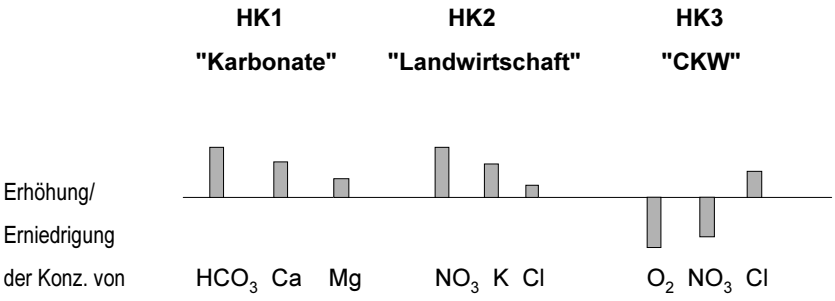
Prinzip:

- fasse verschiedene Variablen zu wenigen synthetischen Linearkombinationen (= ***Hauptkomponenten***) zusammen, die zusammen einen möglichst großen Anteil der Varianz erklären

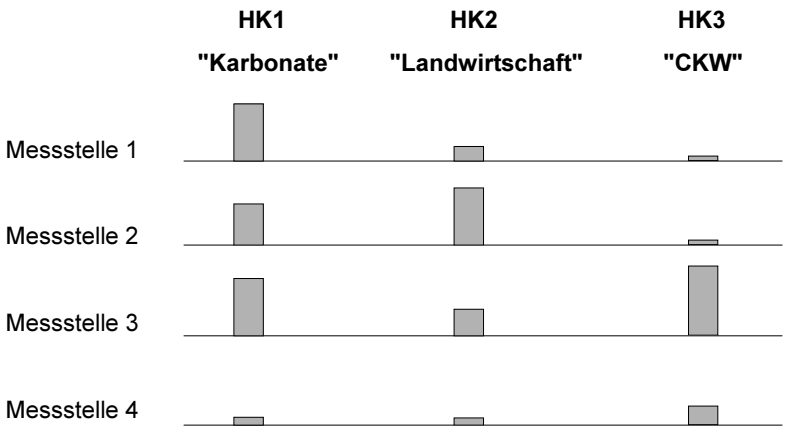
Übliche Anwendung:

- Interpretation der Hauptkomponenten als Einflussfaktoren ("Prozesse")

Beispiel (I)



Beispiel (II)



Orthogonalität: Beispiel

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
1	2	4	4	-1
2	4	2	2	-2
3	6	6	$-\frac{8}{3}$	-3

$$\text{cov}(\mathbf{x}_1;\mathbf{x}_2) = \frac{[(1\quad)\cdot(2\quad)]+[(2\quad)\cdot(4\quad)]+[(3\quad)\cdot(6\quad)]}{\quad}$$

$$\mathbf{x}_1^T \cdot \mathbf{x}_2 = [1\ \ 2\ \ 3]\cdot\begin{bmatrix}2\\4\\6\end{bmatrix} = [(1\cdot 2)+(2\cdot 4)+(3\cdot 6)] = 28$$

$$L(\mathbf{x}_1) = \sqrt{1^2+2^2+3^2} = \sqrt{14}$$
$$L(\mathbf{x}_2) = \sqrt{2^2+4^2+6^2} = \sqrt{56} = 2\cdot\sqrt{14}$$
$$L(\mathbf{x}_1)\cdot L(x_2) = \sqrt{14}\cdot(2\cdot\sqrt{14}) = 28$$

Matrix-Multiplikation

$$\mathbf{A}_{(m;n)} \cdot \mathbf{B}_{(n;q)} = \mathbf{C}_{(m;q)} = \sum_n a_{ij} b_{jk} \quad i = 1, \ldots, m; \quad k = 1, \ldots, q$$

$$\mathbf{A}_{(3;2)} = \begin{bmatrix} 1 & 3 \\ 5 & 4 \\ 2 & -2 \end{bmatrix} \qquad \mathbf{B}_{(2;4)} = \begin{bmatrix} 4 & 6 & 2 & 3 \\ 3 & -1 & 7 & 1 \end{bmatrix}$$

$$\mathbf{A}_{(3;2)} \cdot \mathbf{B}_{(2;4)} = \mathbf{C}_{(3;4)} = \begin{bmatrix} (1\cdot 4)+(3\cdot 3) & (1\cdot 6)+(3\cdot -1) & (1\cdot 2)+(3\cdot 7) & (1\cdot 3)+(3\cdot 1) \\ (5\cdot 4)+(4\cdot 3) & (5\cdot 6)+(4\cdot -1) & (5\cdot 2)+(4\cdot 7) & (5\cdot 3)+(4\cdot 1) \\ (2\cdot 4)+(-2\cdot 3) & (2\cdot 6)+(-2\cdot -1) & (2\cdot 2)+(-2\cdot 7) & (2\cdot 3)+(2\cdot 1) \end{bmatrix}$$

Orthogonalität: Beispiel

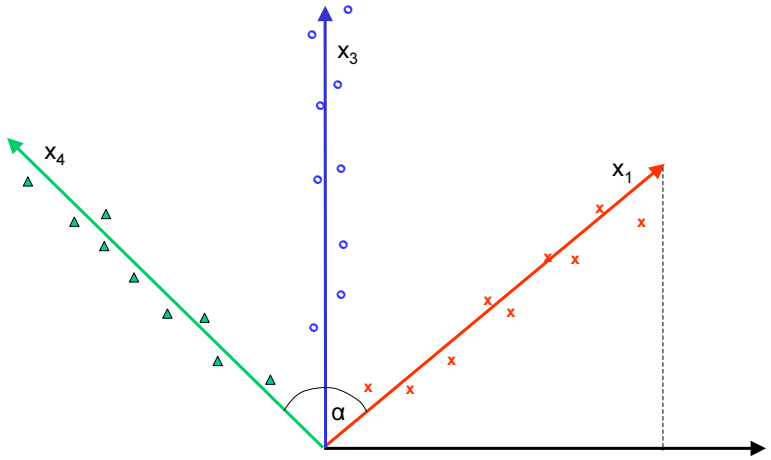
$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
1	2	4	4	−1
2	4	2	2	−2
3	6	6	$-8/3$	−3

$$\mathbf{x}_1^T \cdot \mathbf{x}_2 = [1 \;\; 2 \;\; 3] \cdot \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} = [(1\cdot 2)+(2\cdot 4)+(3\cdot 6)] = 28$$

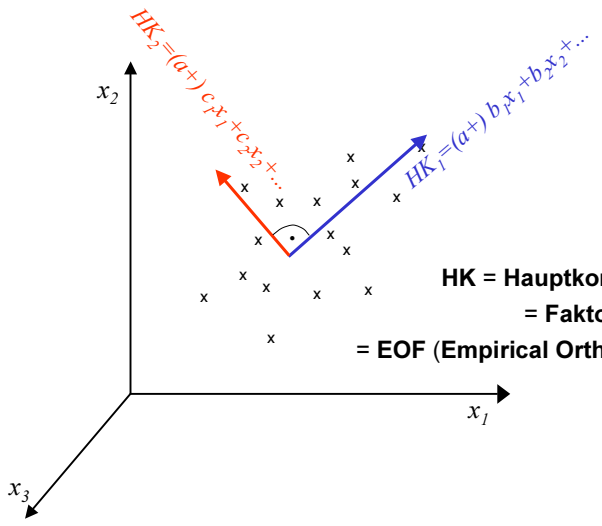
$$\mathbf{x}_1^T \cdot \mathbf{x}_3 = [1 \;\; 2 \;\; 3] \cdot \begin{bmatrix} 4 \\ 2 \\ 6 \end{bmatrix} = [(1\cdot 4)+(2\cdot 2)+(3\cdot 6)] = 24$$

$$\mathbf{x}_1^T \cdot \mathbf{x}_4 = [1 \;\; 2 \;\; 3] \cdot \begin{bmatrix} 4 \\ 2 \\ -8/3 \end{bmatrix} = [(1\cdot 4)+(2\cdot 2)+(3\cdot -8/3)] = 0$$

Korrelation: Vektordarstellung



Bestimmung orthogonaler Funktionen



HK = Hauptkomponenten
= Faktoren
= EOF (Empirical Orthogonal Functions)

Eigenwertzerlegung einer Matrix

Finde einen Vektor ***x***, so dass gilt: $\mathbf{Ax} = \lambda \cdot \mathbf{x}$

d.h.: Das Vektorprodukt der Datenmatrix ***A*** mit dem Vektor ***x*** führt lediglich zu einer Stauchung / Dehnung des resultierenden Vektors, nicht aber zu einer Richtungsänderung

Gegenbeispiel: Rotationsmatrix

$$\mathbf{A} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

Eigenwertzerlegung einer Matrix

Beispiel: $\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 2 & 4 \\ 3 & 0 & 6 \end{bmatrix}$ gesucht: $\mathbf{x}$, für das gilt $\mathbf{Ax} = \lambda \cdot \mathbf{x}$

$$\mathbf{A} \cdot \mathbf{x} = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 2 & 4 \\ 3 & 0 & 6 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} = 2 \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{A} \cdot \mathbf{x} = \begin{bmatrix} 1 & 0 & 2 \\ 2 & 2 & 4 \\ 3 & 0 & 6 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2.8 \\ 3 \end{bmatrix} = \begin{bmatrix} 1+0+6 \\ 2+5.6+12 \\ 3+0+18 \end{bmatrix} = \begin{bmatrix} 7 \\ 19.6 \\ 21 \end{bmatrix} = 7 \cdot \begin{bmatrix} 1 \\ 2.8 \\ 3 \end{bmatrix} = 29.57 \cdot \begin{bmatrix} 0.237 \\ 0.663 \\ 0.710 \end{bmatrix}$$

Kovarianz-Matrix

$$\mathbf{A} = \begin{bmatrix} \mathrm{cov}(x_1, x_1) & \mathrm{cov}(x_1, x_2) & \ldots & \mathrm{cov}(x_1, x_n) \\ \mathrm{cov}(x_2, x_1) & \mathrm{cov}(x_2, x_2) & \ldots & \mathrm{cov}(x_2, x_n) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(x_n, x_1) & \mathrm{cov}(x_n, x_2) & \ldots & \mathrm{cov}(x_n, x_n) \end{bmatrix}$$

Eigenwert-Zerlegung

$$\mathbf{A}\ \mathbf{v} = \lambda\ \mathbf{v}; \quad \mathbf{v} \neq 0$$

$$\begin{bmatrix} \mathrm{cov}(x_1, x_1) & \mathrm{cov}(x_1, x_2) & \ldots & \mathrm{cov}(x_1, x_n) \\ \mathrm{cov}(x_2, x_1) & \mathrm{cov}(x_2, x_2) & \ldots & \mathrm{cov}(x_2, x_n) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(x_n, x_1) & \mathrm{cov}(x_n, x_2) & \ldots & \mathrm{cov}(x_n, x_n) \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \ldots \\ v_n \end{bmatrix} = \lambda \cdot \begin{bmatrix} v_1 \\ v_2 \\ \ldots \\ v_n \end{bmatrix}$$

$$\begin{bmatrix} \mathrm{cov}(x_1, x_1) \cdot v_1 + \mathrm{cov}(x_1, x_2) \cdot v_2 + \ldots + \mathrm{cov}(x_1, x_n) \cdot v_n \\ \mathrm{cov}(x_2, x_1) \cdot v_1 + \mathrm{cov}(x_2, x_2) \cdot v_2 + \ldots + \mathrm{cov}(x_2, x_n) \cdot v_n \\ \ldots \\ \mathrm{cov}(x_n, x_1) \cdot v_1 + \mathrm{cov}(x_n, x_2) \cdot v_2 + \ldots + \mathrm{cov}(x_n, x_n) \cdot v_n \end{bmatrix} = \begin{bmatrix} \lambda \cdot v_1 \\ \lambda \cdot v_2 \\ \ldots \\ \lambda \cdot v_n \end{bmatrix}$$

Eigenwert vs. Kommunalität (I)

Eigenwert (einer Hauptkomponente)**:**
= Beitrag einer **Hauptkomponente** zur Erklärung der Varianz aller Variablen
= Summe der quadrierten **Ladungen** ***aller*** Variablen hinsichtlich ***einer*** Hauptkomponente F_j: $\sum_n a_{nj}^2$

Kommunalität (einer Variablen)**:**
= Beitrag aller **Hauptkomponenten** zur Erklärung der Varianz einer Variablen
= Summe der quadrierten **Faktorladungen** ***einer*** Variablen x_i für ***alle*** Hauptkomponenten: $\sum_k a_{ik}^2$

Eigenwert vs. Kommunalität (II)

	HK1	HK2	Kommunalität
x_1	a_{11}^2	a_{12}^2	Σ
x_2	a_{21}^2	a_{22}^2	Σ
x_3	a_{31}^2	a_{32}^2	Σ
Eigenwert λ	Σ	Σ	

Beispiel

erklärte Varianz

HK	Eigenwert	% Varianz	% kum. Var.
1	4.80	28.2	28.2
2	2.68	15.8	44.0
3	2.04	12.0	55.9
4	1.21	7.1	63.1
5	0.96	5.6	68.7
6	0.92	5.4	74.2
7	0.74	4.3	78.5
8	0.67	3.9	82.4
9	0.59	3.5	85.9
10	0.50	2.9	88.8
11	0.45	2.7	91.5
12	0.42	2.5	93.9
13	0.38	2.2	96.2
14	0.26	1.5	97.7
15	0.18	1.1	98.7
16	0.15	0.9	99.6
17	0.07	0.4	100.0

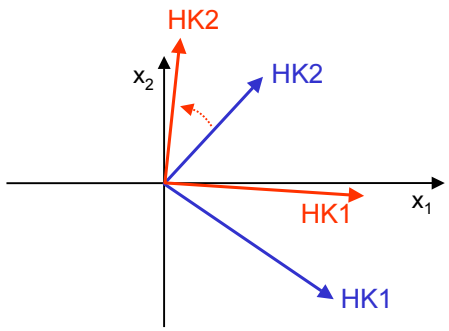
Kommunalität

Cl	0.811
NO3	0.783
HCO3	0.899
Ca	0.890
Mg	0.784
SO4	0.588
Fe	0.519
pH	0.809
Temp.	0.524
Na	0.687
Mn	0.582
NO2	0.374
K	0.594
NH4	0.427
Si	0.534
O2	0.622
CO2	0.296

Varimax-Rotation

Ziel: Achsen der Hauptkomponenten so rotieren, dass einzelne Variablen möglichst eindeutig einzelnen Hauptkomponenten zugeordnet werden können

=>: Gesamt-Varianz bleibt unverändert, geändert wird lediglich die Verteilung der Varianz auf die einzelnen Hauptkomponenten



Beispiel Faktorladungen

nicht rotiert

	1	2	3	4
Cl	0.75	0.09	0.49	0.06
NO3	0.21	-0.34	0.75	0.13
HCO3	0.75	-0.51	-0.27	-0.05
Ca	0.78	-0.52	-0.09	0.06
Mg	0.83	-0.32	-0.08	-0.01
SO4	0.72	0.01	0.19	0.19
Fe	0.25	0.37	-0.41	0.34
pH	0.42	-0.65	-0.45	-0.10
Tmp	0.62	0.25	-0.18	-0.32
Na	0.65	0.45	0.15	-0.17
Mn	0.26	0.59	-0.29	0.27
NO2	-0.02	0.02	-0.14	0.65
K	0.54	0.44	0.26	0.14
NH4	0.20	0.31	-0.18	0.27
Si	0.19	0.58	0.14	-0.45
O2	-0.43	-0.39	0.53	0.17
CO2	0.36	0.19	0.38	0.11

Varimax-Rotation

	1	2	3	4
Cl	0.31	0.83	-0.16	-0.01
NO3	0.15	0.41	-0.77	-0.01
HCO3	0.94	0.12	0.10	-0.01
Ca	0.91	0.26	-0.04	0.00
Mg	0.80	0.38	0.07	0.00
SO4	0.46	0.61	0.03	0.07
Fe	0.15	0.13	0.60	0.35
pH	0.86	-0.23	0.07	-0.07
Tmp	0.29	0.44	0.47	-0.15
Na	0.09	0.75	0.33	-0.12
Mn	-0.18	0.19	0.55	-0.47
NO2	0.02	0.00	-0.01	0.61
K	0.01	0.72	0.19	0.19
NH4	-0.13	0.12	0.03	0.63
Si	-0.30	0.53	0.32	-0.23
O2	-0.17	-0.16	-0.75	0.04
CO2	0.04	0.52	-0.09	0.11

Faktorwerte, Ladungen

- ***Berechnung*** der Hauptkomponenten:

  Multiplikation der Variablen mit den ***Koeffizienten*** c_k: zur Berechnung der ***Faktorwerte*** HK:

$$HK_1 = c_1 \cdot x_1 + c_2 \cdot x_2 + \ldots$$

- ***Interpretation*** der Hauptkomponenten:

  Bestimmung der ***Ladung*** einer Variablen auf eine Hauptkomponente = Korrelation (Pearson-***r***) zwischen Variable und Hauptkomponente

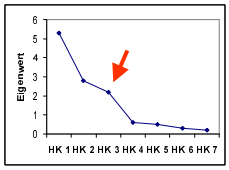
Auswahl der Hauptkomponenten

Hintergrund:

- wenn die Zahl der Hauptkomponenten (HK) = Zahl der Variablen, so wird 100% der Varianz durch die Hauptkomponenten reproduziert
- i.d.R. interessieren aber nur die wichtigsten HK

Übliche Kriterien der Auswahl der Hauptkomponenten:

1. **Kaiser-(Gutmann-)Kriterium**: Auswahl der HK mit Eigenwert > 1.
   Begründung: sind die Daten z-normiert, beträgt die Varianz der einzelnen Variablen jeweils 1, und ein Eigenwert = 1 entspricht einer Varianz = 1 => Auswahl von HK, die einen größeren Teil der Varianz des Datensatzes erklären als jede einzelne Variable
1. **Scree-Plot**: Suche nach deutlichen „Knicks“



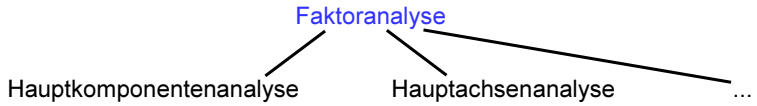
Hauptkomponentenanalyse vs. Hauptachsenanalyse

(akademische Unterscheidung)

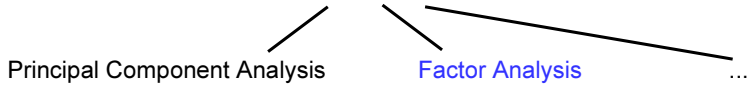
	Hauptkomponentenanalyse	**Hauptachsenanalyse**
Grundgedanke:	Reproduktion der gesamten Varianz	Varianz kann nicht vollständig reproduziert werden
Ziel:	Identifizierung der Einflussfaktoren	Bestimmung unbekannter Einflussfaktoren, die die Restvarianz erklären

Begriffsklärung

Deutsche Unterscheidung:



Englische Unterscheidung:



Zusammenfassung: Hauptkomponentenanalyse

Prinzip: Zusammenfassung verschiedener Variablen zu wenigen synthetischen Linearkombinationen (= ***Hauptkomponenten***), die zusammen einen möglichst großen Anteil der Varianz erklären

Faktorladung: Korrelation einer Variablen mit einer Hauptkomponente

Faktorwert: Wert der Hauptkomponente für einzelne Variable (Proben)

Eigenwert: Beitrag einer Hauptkomponente zur Erklärung der Varianz aller Variablen

Kommunalität: Anteil der Varianz einer Variablen, der durch die Faktoranalyse erklärt (repräsentiert) wird

Aufgabe

Voraussetzung: (log-)normalverteilte und z-transformierte Daten

1. Erstellen Sie eine Korrelationsmatrix (Produkt-Moment-Korrelation) mit allen Ihnen sinnvoll erscheinenden Variablen.
2. Führen Sie eine Hauptkomponentenanalyse mit den auf der Basis der Korrelationsmatrix ausgewählten Variablen durch.
3. Bewerten und gegeb. optimieren Sie Ihre Lösung (Kommunalitäten, Eigenwerte).
4. Vergleichen Sie die Faktorladungen der unrotierten und der Varimax-rotierten Lösung, und ordnen Sie den Hauptkomponenten griffige Bezeichnungen zu.
5. Berechnen und interpretieren Sie die Hauptkomponenten-Werte für die einzelnen Messstellen.