Multivariate Datenanalyse

MSc Psychologie WiSe 2021/22

Prof. Dr. Dirk Ostwald

(4) Hauptkomponentenanalyse

Modul A1/A3 Forschungsmethoden: Multivariate Verfahren

Datum	Einheit	Thema
15.10.2021	Einführung	(0) Einführung
15.10.2021	Grundlagen	(1) Vektoren
22.10.2021	Grundlagen	(2) Matrizen I
29.10.2021	Grundlagen	(3) Matrizen II
05.11.2021	Grundlagen	(4) Multivariate Normalverteilung
12.11.2021	**Achsentransformationen**	**(5) Hauptkomponentenanalyse**
19.11.2021	Achsentransformationen	(6) Faktoranalyse
26.11.2021	Maschinelles Lernen	(7) LDA und Optimierung
03.12.2021	Maschinelles Lernen	(8) Logistische Regression
10.12.2021	Maschinelles Lernen	(9) Support Vektor Maschinen
17.12.2021	Maschinelles Lernen	(10) Neuronale Netze
	Weihnachtspause	
07.01.2022	Frequentistische Inferenz	(11) T-Tests
14.01.2022	Frequentistische Inferenz	(12) Einfaktorielle Varianzanalyse
21.01.2022	Frequentistische Inferenz	(13) Multivariate Regression
28.01.2022	Frequentistische Inferenz	(14) Kanonische Korrelation
22.02.2022	Klausur	12 - 13 Uhr, G26-H1
Jul 2022	Klausurwiederholungstermin	

Überblick

- Hauptkomponentenanalyse heißt auf Englisch Principal Component Analysis (PCA).
- PCA ist eine Featureselektionsmethode.
  - "Features" sind die Komponenten multidimensionaler Zufallsvektoren.
  - Korrelierte Features repräsentieren redundate Information.
- PCA generiert ein korrelationsfreies Featureset durch lineare Featurekombination.
- PCA basiert auf
  - einer Eigenanalyse/Orthonormalzerlegung der Stichprobenkovarianzmatrix und
  - einer anschließenden Vektorkoordinatentransformation.
- Implementiert wird eine PCA oft mithilfe einer Singulärwertzerlegung.
- In der Psychologie dient PCA zum Beispiel
  - der Datenkompression beim Umgang mit neurophysiologischen Zeitseriendaten,
  - der Inspiration im Rahmen der "Exploratorischen Faktoranalyse".

Vektorkoordinatentransformation

Definition und Theorem

Singulärwertzerlegung

Datenkompression

Exploratorische Faktorenanalyse

Selbstkontrollfragen

Vektorkoordinatentransformation

Definition

Datenkompression

Singulärwertzerlegung

Exploratorische Faktorenanalyse

Selbstkontrollfragen

Vektorkoordinatentransformation

Im Folgenden wichtige Begriffe aus (1) Vektoren

Euklidischer Vektorraum. Das Tupel $((\mathbb{R}^m, +, \cdot), \langle\rangle)$ aus dem reellen Vektorraum $(\mathbb{R}^m, +, \cdot)$ und dem Skalarprodukt $\langle\rangle$ auf $\mathbb{R}^m$ heißt *reeller kanonischer Euklidischer Vektorraum*.

Basis. V sei ein Vektorraum und es sei $B \subseteq V$. Dann heißt B eine *Basis von* V, wenn die Vektoren in B linear unabhängig sind und die Vektoren in B den Vektorraum V aufspannen.

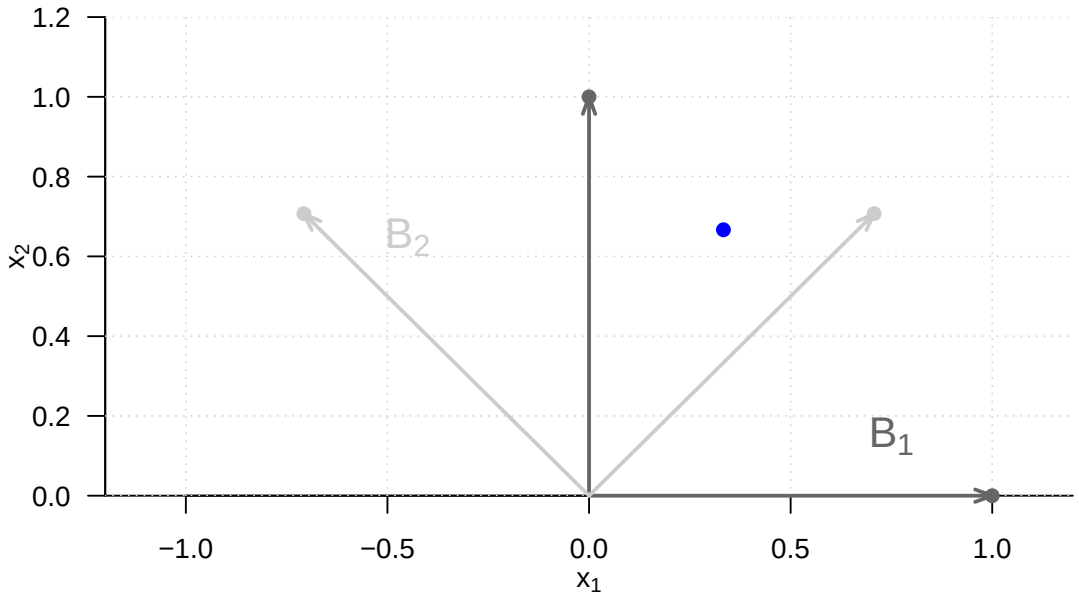
Basisdarstellung und Koordinaten. $B := \{b_1, ..., b_m\}$ sei eine Basis eines m-dimensionalen Vektorraumes V und es sei $x \in V$. Dann heißt die Linearkombination $x = \sum_{i=1}^{m} a_i b_i$ die *Darstellung von* x *bezüglich der Basis* B und die Koeffizienten $a_1, ..., a_m$ heißen die *Koordinaten von* x *bezüglich der Basis* B.

Orthonormalbasis von $\mathbb{R}^m$. Eine Menge von m Vektoren $q_1, ..., q_m \in \mathbb{R}^m$ heißt *Orthonormalbasis* von $\mathbb{R}^m$, wenn $q_1, ..., q_m$ jeweils die Länge 1 haben und wechselseitig orthogonal sind.

Im Folgenden wichtiger Begriff aus (2) Matrizen

Orthonormale Zerlegung einer symmetrischen Matrix. $S \in \mathbb{R}^{m \times m}$ sei eine symmetrische Matrix. Dann kannn S geschrieben werden als $S = Q\Lambda Q^T$, wobei $Q \in \mathbb{R}^{m \times m}$ eine orthogonale Matrix ist und $\Lambda \in \mathbb{R}^{m \times m}$ eine Diagonalmatrix ist. Dabei sind die Spalten von Q die Eigenvektoren von S und die Diagonalelemente von Λ sind die entsprechenden Eigenwerte.

Im Folgenden wichtige Intuition aus (1) Vektoren



- Im Rahmen von Hauptkomponentenranalyse werden wir daran interessiert sein, basierend auf den Koordinaten eines Vektors bezüglich einer Basis die Koordinaten desselben Vektors bezüglich einer anderen Basis zu berechnen.

Definition (Orthogonalprojektion)

x und q seien Vektoren im Euklidischen Vektorraum $\mathbb{R}^m$. Dann ist die *Orthogonalprojektion von x auf q* definiert als der Vektor

$$\tilde{x} = aq \text{ mit } a := \frac{q^T x}{q^T q}, \tag{1}$$

wobei der Skalar a *Projektionsfaktor* genannt wird.

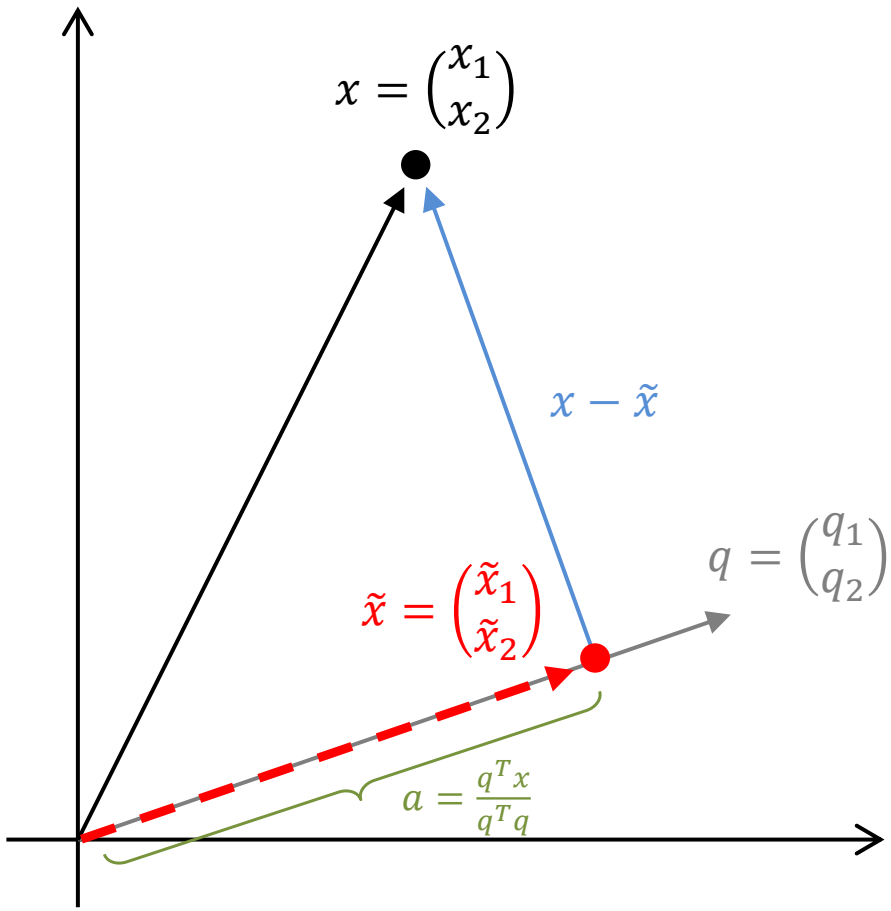
Bemerkungen

- Per definition ist $\tilde{x} = aq$ mit $a \in \mathbb{R}$ der Punkt in Richtung von q der x am nähesten ist.
- Diese minimierte Distanzeigenschaft impliziert die Orthogonalität von q und $x - \tilde{x}$.
- Die Formel von a folgt direkt aus der Orthogonalität von $x - \tilde{x}$ und q, da gilt

$$q^T(x - \tilde{x}) = 0 \Leftrightarrow q^T(x - aq) = 0 \Leftrightarrow q^T x - aq^T q = 0 \Leftrightarrow a = \frac{q^T x}{q^T q}.$$

- Wenn q die Länge $||q|| = \sqrt{q^T q} = 1$ hat, dann gilt $a = \frac{q^T x}{||q||^2} = q^T x$.

Orthogonalprojektion



Theorem (Vektorkoordinaten bezüglich einer Orthogonalbasis)

Es sei $x \in \mathbb{R}^m$ und es sei $B := \{q_1, \ldots, q_m\}$ eine Orthonormalbasis von von $\mathbb{R}^m$. Dann ergeben sich für $i = 1, \ldots, m$ die Koordinaten a_i in der Basisdarstellung von x bezüglich B als die Projektionsfaktoren

$$a_i = x^T q_i \tag{2}$$

in der Orthogonalprojektion von x auf q_i. Äquivalent ist die Basisdarstellung von x bezüglich B gegeben durch

$$x = \sum_{i=1}^{m} (x^T q_i) q_i . \tag{3}$$

Beweis

Für $i = 1, \ldots, m$ gilt

$$x = \sum_{j=1}^{m} a_j q_j \Leftrightarrow q_i^T x = q_i^T \sum_{j=1}^{m} a_j q_j \Leftrightarrow q_i^T x = \sum_{j=1}^{m} a_j q_i^T q_j \Leftrightarrow q_i^T x = a_i \Leftrightarrow a_i = x^T q_i . \tag{4}$$

□

Theorem (Vektorkoordinatentransformation)

$B_v := \{v_1, \ldots, v_m\}$ und $B_w := \{w_1, \ldots, w_m\}$ seien zwei Orthonormalbasen eines Vektorraums. $A \in \mathbb{R}^{m \times m}$ sei die Matrix, die durch die spaltenweise Konkatenation der Koordinaten der Vektoren in B_w in der Basisdarstellung bezüglich der Basis B_v ergibt. Dann können die Koordinaten $x_i, i = 1, \ldots, m$ eines Vektors x bezüglich der Basis B_v in die Koordinaten $\tilde{x}_1, \ldots, \tilde{x}_m$ des Vektors bezüglich der Basis B_w durch

$$\tilde{x} = A^T x \tag{5}$$

transformiert werden. Analog können die Koordinaten $\tilde{y}_1, \ldots, \tilde{y}_m$ des Vektors hinsichtlich der Basis B_w in die Koordinaten $y_1, \ldots, y_m$ des Vektors hinsichtlich B_v durch

$$x = A\tilde{x}. \tag{6}$$

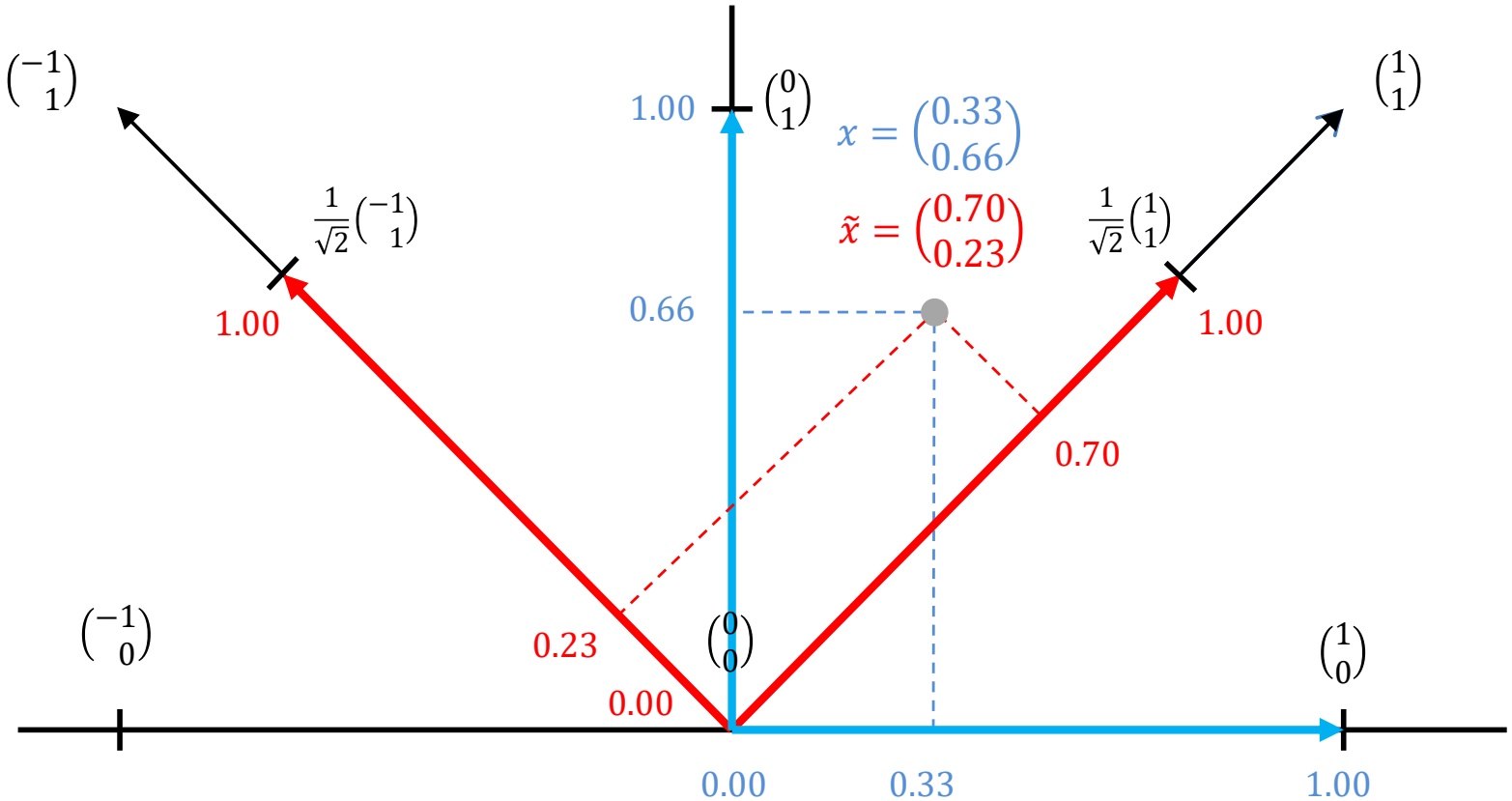
transformiert werden.

Bemerkungen

- Das Theorem erlaubt die Berechnung von Vektorkoordinaten bezüglich einer anderen Orthonormalbasis.
- Für die Berechnung muss zunächst die Matrix A gebildet und dann (nur) entsprechend multipliziert werden.
- Wir verzichten auf einen Beweis und demonstrieren das Theorem an einem Beispiel.

Ein Vektor wird hier als fester Punkt in $\mathbb{R}^m$ betrachtet; die Komponenten (Zahlen) des Vektors werden dagegen nur als Koordinaten bezüglich einer spezifischen Basis interpretiert!

Beispiel



Man beachte, dass x and $\tilde{x}$ am selben Ort in $\mathbb{R}^2$ liegen!

Beispiel

Wir nehmen an, dass wir die Koordinaten von $x = (1/3, 2/3)^T \in \mathbb{R}^2$ hinsichtlich der kanonischen Orthonormalbasis $B_v := \{e_1, e_2\}$ in die Koordinaten bezüglich der Basis

$$B_w := \left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\} \tag{7}$$

transformieren wollen. Die Basisdarstellungen der in Vektoren B_w bezüglich der Basisvektoren in B_v sind

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = a_{11} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + a_{21} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = a_{12} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + a_{22} \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{8}$$

Die Projektionsfaktoren der Orthogonalprojektionen der Vektoren in B_w auf die Vektoren in B_v sind

$$a_{11} = \frac{1}{\sqrt{2}}, a_{21} = \frac{1}{\sqrt{2}}, a_{12} = -\frac{1}{\sqrt{2}}, a_{22} = \frac{1}{\sqrt{2}}. \tag{9}$$

Die Transformationsmatrix $A \in \mathbb{R}^{m \times m}$ in obigem Theorem ergibt sich also zu

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{10}$$

Die Vektorkoordinatentransformation von $x \in \mathbb{R}^2$ ergibt sich also zu

$$\tilde{x} = A^T x = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix} \approx \begin{pmatrix} 0.70 \\ 0.23 \end{pmatrix}. \tag{11}$$

Vektorkoordinatentransformation

Definition und Theorem

Singulärwertzerlegung

Datenkompression

Exploratorische Faktorenanalyse

Selbstkontrollfragen

Definition (Hauptkomponentenanalyse)

$\mathbb{C}(X)$ sei die Kovarianzmatrix eines m-dimensionalen Zufallsvektors X. Dann heißt die orthonormale Zerlegung

$$\mathbb{C}(X) = Q\Lambda Q^T, \tag{12}$$

wobei

- $Q \in \mathbb{R}^{m \times m}$ die Matrix der spaltenweisen Konkatenation der Eigenvektoren von $\mathbb{C}(X)$ ist und
- $\Lambda \in \mathbb{R}^{m \times m}$ die Diagonalmatrix der zugehörigen Eigenwerte bezeichnen,

die *Hauptkomponentenanalyse* von $\mathbb{C}(X)$ und die Spalten von Q heißen die *Hauptkomponenten* von $\mathbb{C}(X)$. Der m-dimensionale Zufallsvektor

$$\tilde{X} = Q^T X \tag{13}$$

heißt *PCA-transformierter Zufallsvektor*.

Bemerkungen

- Man spricht auch von der Hauptkomponentenanalyse/den Hauptkomponenten von X.

Theorem (Hauptkomponentenanalyse)

$\mathbb{C}(X) \in \mathbb{R}^{m \times m}$ sei die Kovarianzmatrix eines m-dimensionalen Zufallsvektors X, es sei $\mathbb{E}(X) = 0_m$ und es sei

$$\mathbb{C}(X) = Q \Lambda Q^T, \tag{14}$$

die Hauptkomponentenanalyse von $\mathbb{C}(X)$. Dann gelten

(1) Die Spalten von Q bilden eine Orthonormalbasis von $\mathbb{R}^m$.

(2) Multiplikation mit Q^T transformiert die kanonischen Koordinaten von X in Koordinaten bezüglich der Hauptkomponenten von $\mathbb{C}(X)$.

(3) Die Kovarianzmatrix des PCA-transformierten Zufallsvektors ist die Diagonalmatrix Λ.

(4) In dem Koordinatensystem, das von den Hauptkomponenten von $\mathbb{C}(X)$ aufgespannt wird, gilt

$$\mathbb{V}(\tilde{X}_i) = \lambda_i \text{ für } i = 1, \ldots, m \text{ und } \mathbb{C}(\tilde{X}_i, \tilde{X}_j) = 0 \text{ für } i \neq j, 1 \leq i, j \leq m. \tag{15}$$

Beweis

(1) Mit dem Theorem zu den Eigenschaften von Basen aus Einheit (1) Vektoren gilt, dass jede Menge von m linear unabhängigen Vektoren Basis eines m-dimensionalen Vektorraums ist. Die Spalten $q_1, \ldots q_m \in \mathbb{R}^m$ von Q sind m orthonormale Vektoren und damit insbesondere auch linear unabhängig, denn für $i = 1, \ldots, m$ gilt

$$\begin{aligned} a_1 q_1 + a_2 q_2 + \cdots + a_m q_m &= 0_m \\ \Leftrightarrow (a_1 q_1 + a_2 q_2 + \cdots + a_m q_m)^T &= 0_m^T \\ \Leftrightarrow (a_1 q_1 + a_2 q_2 + \cdots + a_m q_m)^T q_i &= 0_m^T q_i \\ \Leftrightarrow \sum_{j=1}^{m} a_j q_j^T q_i &= 0 \\ \Leftrightarrow a_i &= 0. \end{aligned} \tag{16}$$

Es ist also $a_i = 0$ für $i = 1, \ldots, m$ und die einzige Repräsentation des Nullelements 0_m durch eine Linearkombination der Spalten von Q ist die triviale Repräsentation. Die Spalten von Q sind also m unabhängige Vektoren und damit eine Basis von $\mathbb{R}^m$. Da die Spalten von Q auch orthonormal sind, bilden sie eine Orthonormalbasis von $\mathbb{R}^m$.

Beweis (fortgeführt)

(2) Wir betrachten das Theorem zur Vektorkoordinatentransformation aus dieser Einheit und setzen $B_v := \{e_1, ..., e_m\}$ und $B_w := \{q_1, ...q_m\}$ mit den Spalten $q_1, ..., q_m \in \mathbb{R}^m$ von Q. Dann gilt, dass $Q \in \mathbb{R}^{m \times m}$ die Matrix ist, die sich durch die spaltenweise Konkatenation der Koordinaten der Vektoren in B_w in der Basisdarstellung bezüglich der Basis B_v ergibt, denn für $i = 1, ..., m$ gilt, dass die Basisdarstellung von q_i bezüglich der kanonischen Basis B_v gegeben ist durch

$$q_i = \sum_{j=1}^{m} (q_i^T e_j) e_j = \sum_{j=1}^{m} q_{i_j} e_j = q_i. \tag{17}$$

Äquivalent ist natürlich jeder Vektor $q \in \mathbb{R}^m$ schon immer identisch mit der Basisdarstellung von q bezüglich der kanonischen Basis. Damit folgt aber mit Theorem zur Vektorkoordinatentransformation direkt, dass der PCA-transfomierte Zufallsvektor

$$\tilde{X} = Q^T X \tag{18}$$

aus den Koordinaten des Vektors bezüglich der Hauptkomponenten von $\mathbb{C}(X)$ besteht.

(3) Wir erinnern zunächst daran, dass die inverse Matrix einer orthogonalen Matrix Q durch Q^T gegeben ist (vgl. Einheit (2) Definition symmetrischer, diagonaler, und orthogonaler Matrize). Mit $QQ^T = Q^T Q = I_m$ gilt dann, dass

$$\mathbb{C}(X) = Q\Lambda Q^T \Leftrightarrow Q^T \mathbb{C}(X) Q = Q^T Q \Lambda Q^T Q \Leftrightarrow Q^T \mathbb{C}(X) Q = \Lambda. \tag{19}$$

Beweis (fortgeführt)

Weiterhin gilt, dass mit $\mathbb{E}(X) = 0_m$ die Kovarianzmatrix von X gegeben ist durch (vgl. Einheit (3) Definition der Kovarianzmatrix)

$$\mathbb{C}(X) = \mathbb{E}\left((X - \mathbb{E}(X))\,(X - \mathbb{E}(X))^T\right) = \mathbb{E}\left(XX^T\right). \tag{20}$$

Damit ergibt sich für die Kovarianzmatrix des PCA-transformierte Vektors $\tilde{X} = Q^T X$ aber, dass

$$\begin{aligned}
\mathbb{C}(\tilde{X}) &= \mathbb{E}\left(\left(\tilde{X} - \mathbb{E}(\tilde{X})\right)\left(\tilde{X} - \mathbb{E}(\tilde{X})\right)^T\right) \\
&= \mathbb{E}\left(\left(Q^T X - \mathbb{E}(Q^T X)\right)\left(Q^T X - \mathbb{E}(Q^T X)\right)^T\right) \\
&= \mathbb{E}\left(\left(Q^T X - Q^T \mathbb{E}(X)\right)\left(Q^T X - Q^T \mathbb{E}(X)\right)^T\right) \\
&= \mathbb{E}\left((Q^T X)(Q^T X)^T\right) \\
&= Q^T \mathbb{E}\left(XX^T\right) Q \\
&= Q^T \mathbb{C}(X) Q \\
&= \Lambda.
\end{aligned} \tag{21}$$

(4) Die Koordinaten von $\tilde{X}$ entsprechen den Koordinaten von X in dem Koordinatensystem, dass von den Hauptkomponenten $q_1, \ldots, q_m$ von $\mathbb{C}(X)$ aufgespannt wird. Mit $\mathbb{C}(\tilde{X}) = \Lambda$ folgt Aussage (4) dann direkt mit der Definition der Kovarianzmatrix in Einheit (3). □

Definition und Theorem

Bemerkungen

- Die Eigenwerte $\lambda_1, ..., \lambda_m$ von $\mathbb{C}(X)$ sind die Varianzen von $\tilde{X}_1, ..., \tilde{X}_m$.

- Bei Annahme von $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ mit zugehörigen Eigenvektoren $q_1, ..., q_m$ gilt

$$\mathbb{V}(\tilde{X}_1) > \mathbb{V}(\tilde{X}_2) > \cdots > \mathbb{V}(\tilde{X}_m) \Leftrightarrow \mathbb{V}(q_1^T X) > \mathbb{V}(q_2^T X) > \cdots > \mathbb{V}(q_m^T X) \tag{22}$$

- Die paarweise nicht-identischen Kovarianzen der Komponenten von $\tilde{X}$ sind Null.

  $\Rightarrow$ Die Komponenten von $\tilde{X}$ sind unkorreliert.

  $\Rightarrow$ Die Komponenten von $\tilde{X}$ repräsentieren keine redundante Information.

- $q_1^T X$ maximiert die Varianz der unkorrelierten Linearkombinationen der Komponenten von X.

Definition (Hauptkomponentenanalyse eines Datensatzes)

$Y \in \mathbb{R}^{m \times n}$ sei ein Datensatz aus n unabhängigen Realisierungen eines m-dimensionalen Zufallsvektors und es sei $C \in \mathbb{R}^{m \times m}$ die Stichprobenkovarianzmatrix des Datensatzes. Dann heißt die Orthonormalzerlegung

$$C = Q \Lambda Q^T \tag{23}$$

wobei

- $Q \in \mathbb{R}^{m \times m}$ die spaltenweise Konkatenation der Eigenvektoren von C ist und
- $\Lambda \in \mathbb{R}^{m \times m}$ die Diagonalmatrix der zugehörigen Eigenwerten bezeichnen,

die *Hauptkomponentenanalyse von* C und die Spalten von Q heißen die *Hauptkomponenten von* C. Der $m \times n$-dimensionale Datensatz

$$\tilde{Y} = Q^T Y \tag{24}$$

heißt *PCA-transformierter Datensatz*.

Bemerkungen

- Man spricht auch von der Hauptkomponentenanalyse/den Hauptkomponenten des Datensatzes Y.

Theorem (Hauptkomponentenanalyse eines Datensatzes)

$C \in \mathbb{R}^{m \times m}$ sei die Stichprobenkovarianzmatrix eines Datensatzes $Y \in \mathbb{R}^{m \times n}$ und es sei

$$C = Q \Lambda Q^T, \tag{25}$$

die Hauptkomponentenanalyse von C. Dann gelten

(1) Die Spalten von Q bilden eine Orthonormalbasis von $\mathbb{R}^m$.

(2) Multiplikation mit Q^T transformiert die kanonischen Koordinaten der Spalten von Y in Koordinaten bezüglich der Hauptkomponenten von C.

(3) Die Stichprobenkovarianzmatrix des PCA-transformierten Datensatzes ist die Diagonalmatrix Λ.

(4) In dem Koordinatensystem, das von den Hauptkomponenten von C aufgespannt wird, gilt

$$S^2(\tilde{Y}_i) = \lambda_i \text{ für } i = 1, \ldots, m \text{ und } C(\tilde{Y}_i, \tilde{Y}_j) = 0 \text{ für } i \neq j, 1 \leq i, j \leq m. \tag{26}$$

wobei $S^2(\tilde{Y}_i)$ die Stichprobenvarianz der $i$ten Komponente des Datensatzes und $C(\tilde{Y}_i, \tilde{Y}_j)$ die Stichprobenkovarianz der $i$ten und $j$ten Komponente des Datensatzes bezeichnen.

Bemerkungen

- Der Beweis ergibt sich in Analogie zum Beweis Theorems zur Hauptkomponentenanalyse
- Wir verzichten aus eine Ausformulierung des Beweises.

Hauptkomponentenanalyse eines simulierten Datensatzes

Datensatzgeneration

```
# R Pakete
library(matrixcalc)                          # Matrix Paket (is.positive.definite())
library(MASS)                                # Multivariate Normalverteilung (mvrnorm())

# Simulationsparameter
set.seed(1)                                  # Reproduzierbare Randomisierung
m     = 5                                    # Datenpunktdimension
n     = 20                                   # Anzahl Realisierung
mu    = rep(0,m)                             # Erwartungswertparameter
Sigma = matrix(runif(m^2), nrow = m)         # zufällige Matrix
Sigma = 0.5*(Sigma+t(Sigma))                 # symmetrische Matrix
Sigma = Sigma + m*diag(m)                    # positiv definite Matrix
print(is.positive.definite(Sigma))           # Positiv-Definitheits Check
```

```
> [1] TRUE
```

```
# Datensatzgeneration
Y     = t(mvrnorm(n,mu,Sigma))
```

Hauptkomponentenanalyse eines simulierten Datensatzes

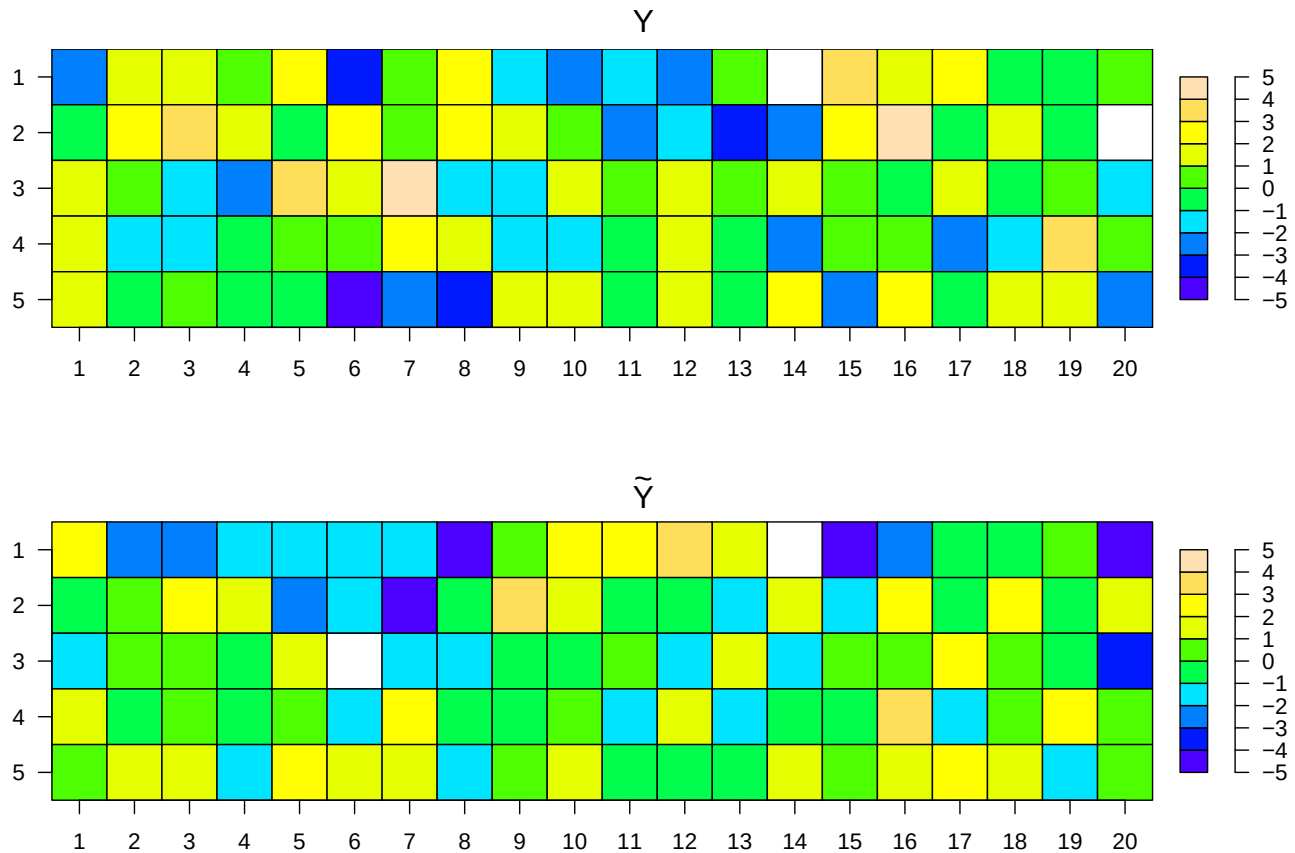
Hauptkomponentenanalyse durch Eigenanalyse

```
# Hauptkomponentenanalyse durch Eigenanalyse
I_n     = diag(n)                                    # Einheitsmatrix I_n
J_n     = matrix(rep(1,n^2), nrow = n)               # 1_{nn}
C       = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y)) # Stichprobenkovarianzmatrix
D       = diag(1/sqrt(diag(C)))                      # Kov-Korr-Transformationsmatrix
R       = D %*% C %*% D                              # Stichprobenkorrelationsmatrix
EA      = eigen(C)                                   # Eigenanalyse von C
lambda  = EA$values                                  # Eigenwerte von C
Q       = EA$vectors                                 # Eigenvektoren von C
Y_tilde = t(Q) %*% Y                                 # Transformierter Datensatz

# Stichproben- und Korrelationsmatrix des transformierten Datensatzes
C_tilde = (1/(n-1))*(Y_tilde %*% (I_n-(1/n)*J_n) %*% t(Y_tilde))
D_tilde = diag(1/sqrt(diag(C_tilde)))
R_tilde = D_tilde %*% C_tilde %*% D_tilde
```

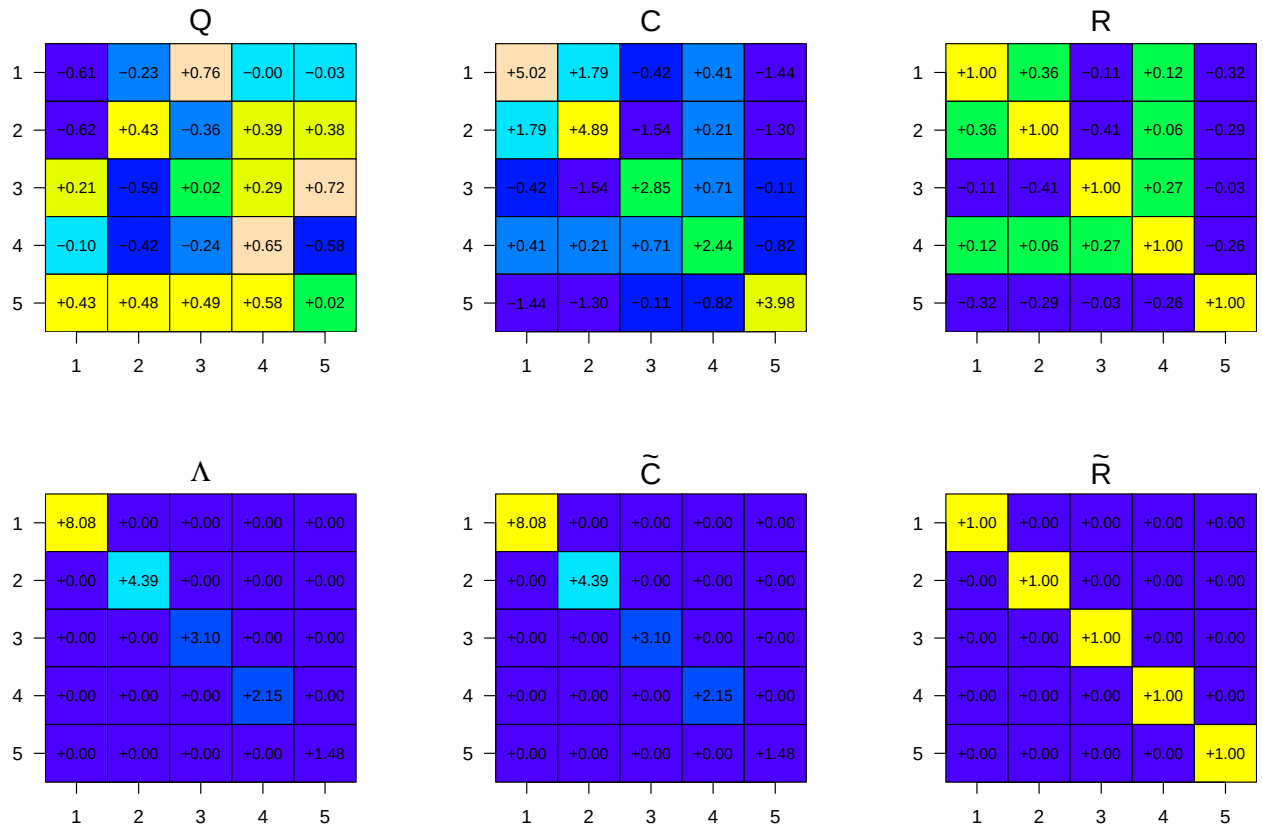
Hauptkomponentenanalyse eines simulierten Datensatzes

Hauptkomponentenanalyse durch Eigenanalyse



Hauptkomponentenanalyse eines simulierten Datensatzes

Hauptkomponentenanalyse durch Eigenanalyse



Vektorkoordinatentransformation

Definition

Singulärwertzerlegung

Datenkompression

Exploratorische Faktorenanalyse

Selbstkontrollfragen

Definition (Singulärwertzerlegung)

$X \in \mathbb{R}^{m \times n}$ sei eine Matrix. Dann heißt die Zerlegung

$$X = USV^T, \tag{27}$$

wobei $U \in \mathbb{R}^{m \times m}$ eine orthogonale matrix ist, $S \in \mathbb{R}^{m \times n}$ eine Diagonalmatrix ist und $V \in \mathbb{R}^{n \times n}$ eine orthogonale Matrix ist, *Singulärwertzerlegung (Singular Value Decomposition (SVD))* von X. Die Diagonalelemente von S heißen die *Singulärwerte* von X.

Bemerkungen

- Die Existenz der Singulärwertzerlegung folgt aus dem Spektralsatz der Linearen Algebra.
- Singulärwertzerlegungen können in R mit svd() berechnet werden.

Theorem (Singulärwertzerlegung und Eigenanalyse)

$X \in \mathbb{R}^{m \times n}$ sei eine Matrix und

$$X = USV^T \tag{28}$$

sei ihre Singulärwertzerlegung. Dann gilt:

- Die Spalten von U sind die Eigenvektoren von XX^T,
- die Spalten von V sind die Eigenvektoren von X^TX und
- die entsprechenden Singulärwerte sind die Quadratwurzeln der zugehörigen Eigenwerte.

Bemerkung

- Singulärwertzerlegung und Eigenanalyse sind eng verwandt.

Beweis

Wir halten zunächst fest, dass mit

$$\left(XX^T\right)^T = XX^T \text{ and } \left(X^TX\right)^T = X^TX, \tag{29}$$

XX^T und X^TX symmetrische Matrizen sind und somit Orthornomalzerlegungen haben. Wir halten weiterhin fest, dass mit der Definition der Singulärwertzerlegung gelten, dass sowohl

$$XX^T = USV^T \left(U\Sigma V^T\right)^T = USV^TVS^TU^T = USSU^T = U\Lambda U^T \tag{30}$$

als auch

$$X^TX = \left(USV^T\right)^T USV^T = VS^TUUS^TV^T = V\Lambda V^T \tag{31}$$

ist, wobei wir $\Lambda := SS$ definiert haben. Weil das Produkt von Diagonalmatrizen wieder eine Diagonalmatrix ist, ist Λ eine Diagonalmatrix und per Definition sind U und V orthogonale Matrizen. Wir haben also XX^T und X^TX in Form der Orthonormalzerlegungen

$$XX^T = U\Lambda U^T \text{ and } X^TX = V\Lambda V^T \tag{32}$$

geschrieben und damit ist alles gezeigt.

□

Theorem (Datenhauptkomponentenanalyse durch Singulärwertzerlegung)

$Y \in \mathbb{R}^{m \times n}$ sei ein Datensatz von n unabhängigen Realisierungen eines m-dimensionalen Zufallsvektors. Weiterhin sei

$$\frac{1}{\sqrt{n-1}} Y_c = USV^T \tag{33}$$

die Singulärwertzerlegung des skalierten und zentrierten Datenmatrix. Dann sind die Spalten von U die Eigenvektoren der Stichprobenkovarianzmatrix des Datensatzes Y und die quadrierten Singulärwerte sind die zugehörigen Eigenwerte.

Beweis

Nach Definition der Singulärwertzerlegung sind die Spalten von U die Eigenvektoren von

$$\frac{1}{\sqrt{n-1}} Y_c \frac{1}{\sqrt{n-1}} Y_c^T = \frac{1}{n-1} Y_c Y_c^T =: C, \tag{34}$$

und damit identisch mit den Eigenvektoren der Stichprobenkovarianzmatrix. Weil die Singulärwerte die Quadratwurzeln der zugehörigen Eigenwerte sind, sind ihre quadrierten Werte identisch zu den zugehörigen Eigenwerten.

□

Eine PCA kann durch Eigenanalyse der Stichprobenkovarianzmatrix berechnet werden

Eine PCA kann auch durch Singulärwertzerlegung der skaliert-zentrierten Datenmatrix berechnet werden

Hauptkomponentenanalyse eines simulierten Datensatzes

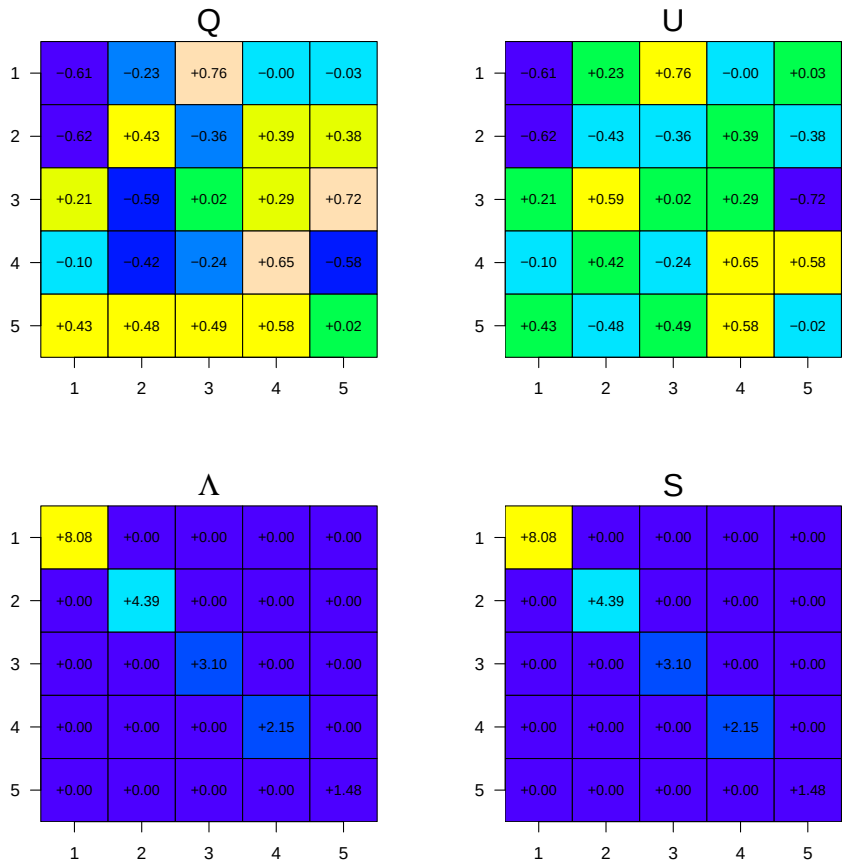
Hauptkomponentenanalyse durch Singulärwertzerlegung

```
# Hauptkomponentenanalyse durch Singulärwertzerlegung
Y_sc    = (1/sqrt(n-1))*(Y-as.matrix(rowMeans(Y)) %*% rep(1,n)) # skaliert-zentrierter Datensatz
SWD     = svd(Y_sc)                                              # Singulärwertzerlegung
U       = SWD$u                                                  # Eigenvektoren von C
s       = SWD$d^2                                                # Eigenwerte von C
Y_tilde = t(U) %*% Y                                             # Transformierter Datensatz

# Stichproben- und Korrelationsmatrix des transformierten Datensatzes
C_tilde = (1/(n-1))*(Y_tilde %*% (I_n-(1/n)*J_n) %*% t(Y_tilde))
D_tilde = diag(1/sqrt(diag(C_tilde)))
R_tilde = D_tilde %*% C_tilde %*% D_tilde
```

Hauptkomponentenanalyse eines simulierten Datensatzes

Hauptkomponentenanalyse durch Singulärwertzerlegung



Vektorkoordinatentransformation

Definition

Singulärwertzerlegung

Datenkompression

Exploratorische Faktorenanalyse

Selbstkontrollfragen

Überblick

- Datenkompression entspricht einer Reduktion der Dimension m von Daten.
- Im Rahmen der prädiktiven Modellierung wird PCA zur *Dimensionsreduktion* eingesetzt.
- Ziel ist es hier, dem Undersampling hochdimensionaler Datenräume entgegen zu wirken.
- Dimensionsreduktion entspricht dem Verwerfen von $k < m$ Komponenten von $\tilde{Y}$.
- Wahl von Komponenten mit hohen Eigenwerten hält den *Datenrekonstruktionsfehler* klein.

Definition (Dimensionsreduzierter Datensatz)

$Y \in \mathbb{R}^{m \times n}$ sei ein Datensatz,

$$C = \frac{1}{n-1}\left(Y\left(I_n - \frac{1}{n}1_{nn}\right)Y^T\right) \in \mathbb{R}^{m \times m} \tag{35}$$

sei die zugehörige Stichprobenkovarianzmatrix,

$$C = Q\Lambda Q^T \tag{36}$$

sei die Hauptkomponentenanalyse von C und es gelte $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ für die Diagonalelemente von Λ. Schließlich sei für $k \leq m$ Q_k die Matrix, die aus Q durch Streichen der Spalten $k+1, \ldots, m$ entsteht. Dann heißt

$$\tilde{Y}_k = Q_k^T Y \in \mathbb{R}^{k \times n} \tag{37}$$

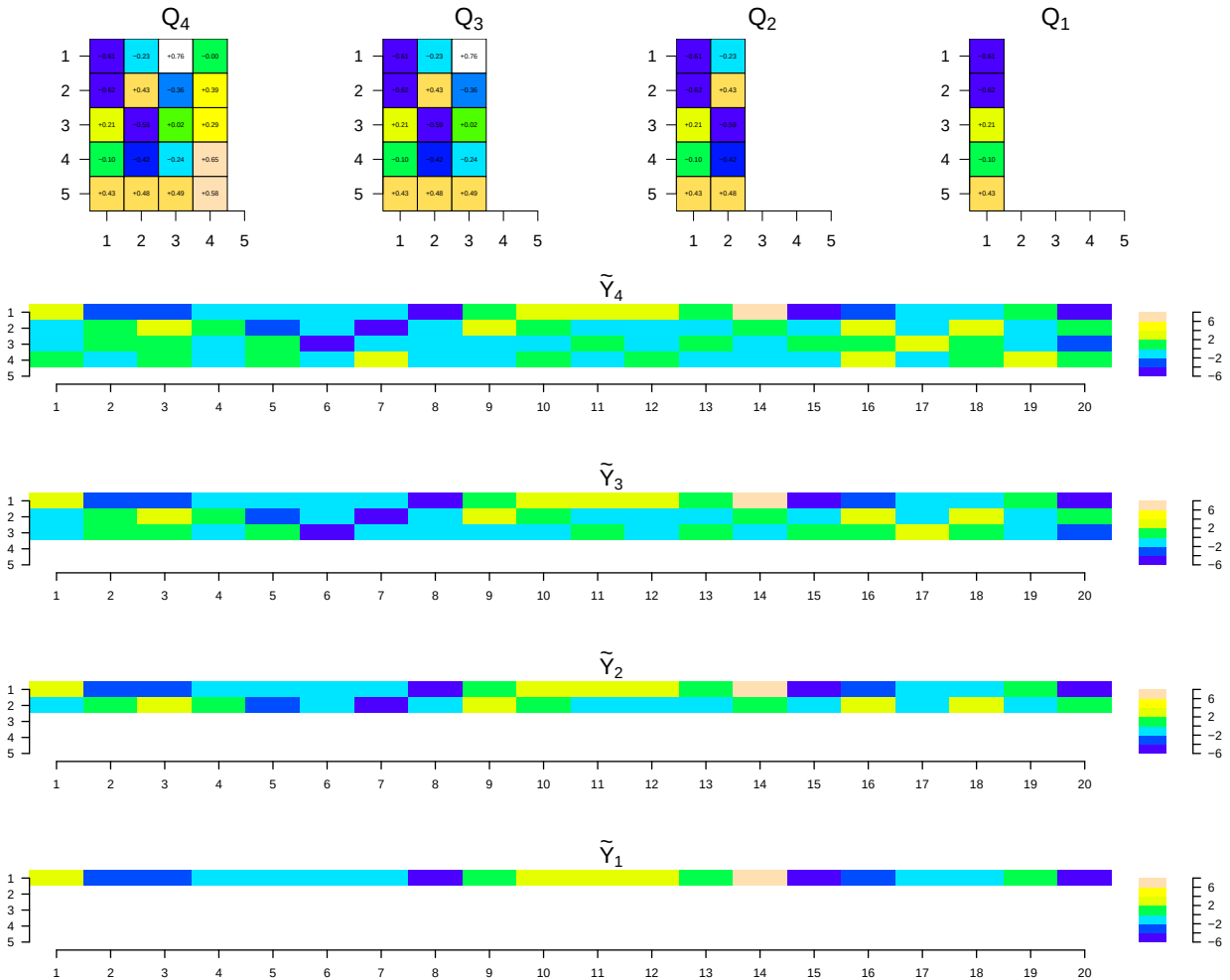
PCA-dimensionsreduzierter Datensatz.

Bemerkung

- $\tilde{Y}_k = Q_k^T Y$ entspricht einer $(k \times n) = (k \times m) \cdot (m \times n)$ Matrixmultiplikation
- $\tilde{Y}_k$ ist der Datensatz, der aus $\tilde{Y}$ durch Streichen der $(k+1)$-ten bis m-ten Zeile ensteht.

Dimensionalitätsreduktion eines simulierten Datensatzes

Dimensionsreduzierte Datensätze



Definition (Rekonstruierter Datensatz, Datenrekonstruktionsfehler)

$Y \in \mathbb{R}^{m \times n}$ sei ein Datensatz und für $k \leq m$ sei

$$\tilde{Y}_k = Q_k^T Y \in \mathbb{R}^{k \times n} \tag{38}$$

ein PCA-dimensionsreduzierter Datensatz. Dann heißt

$$Y_k = Q_k \tilde{Y}_k \in \mathbb{R}^{m \times n} \tag{39}$$

rekonstruierter Datensatz und

$$e = \|\text{vec}(Y - Y_k)\| \geq 0 \tag{40}$$

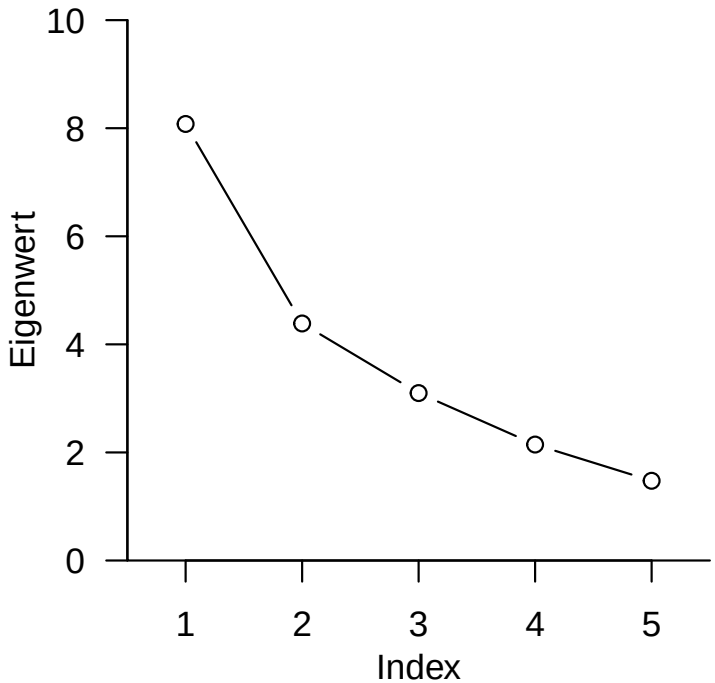
heißt *Datenrekonstruktionsfehler*.

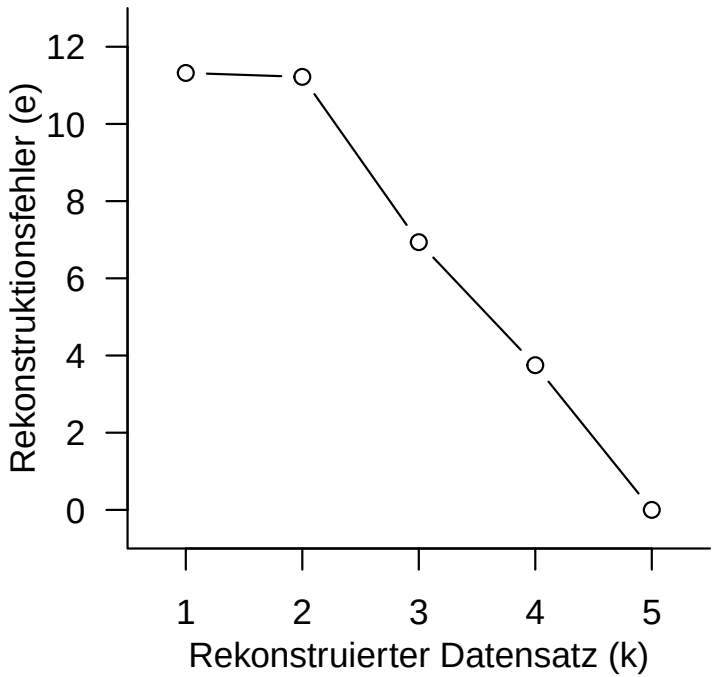
Bemerkungen

- $Y_k = Q_k \tilde{Y}_k$ entspricht einer $(m \times n) = (m \times k) \cdot (k \times n)$ Matrixmultiplikation
- Für $M \in \mathbb{R}^{m \times n}$ ist $\text{vec}(M) \in \mathbb{R}^{mn}$ der Vektor, der durch Stapeln der Spalten von M entsteht.
- Für $k = m$ gilt $Q\tilde{Y}_k = QQ^T Y = Y$ und damit $e = 0$.

Datensatzrekonstruktion und -rekonstruktionsfehler eines simulierten Datensatzes

Eigenwerte (“Scree-Plot”) und Rekonstruktionsfehler



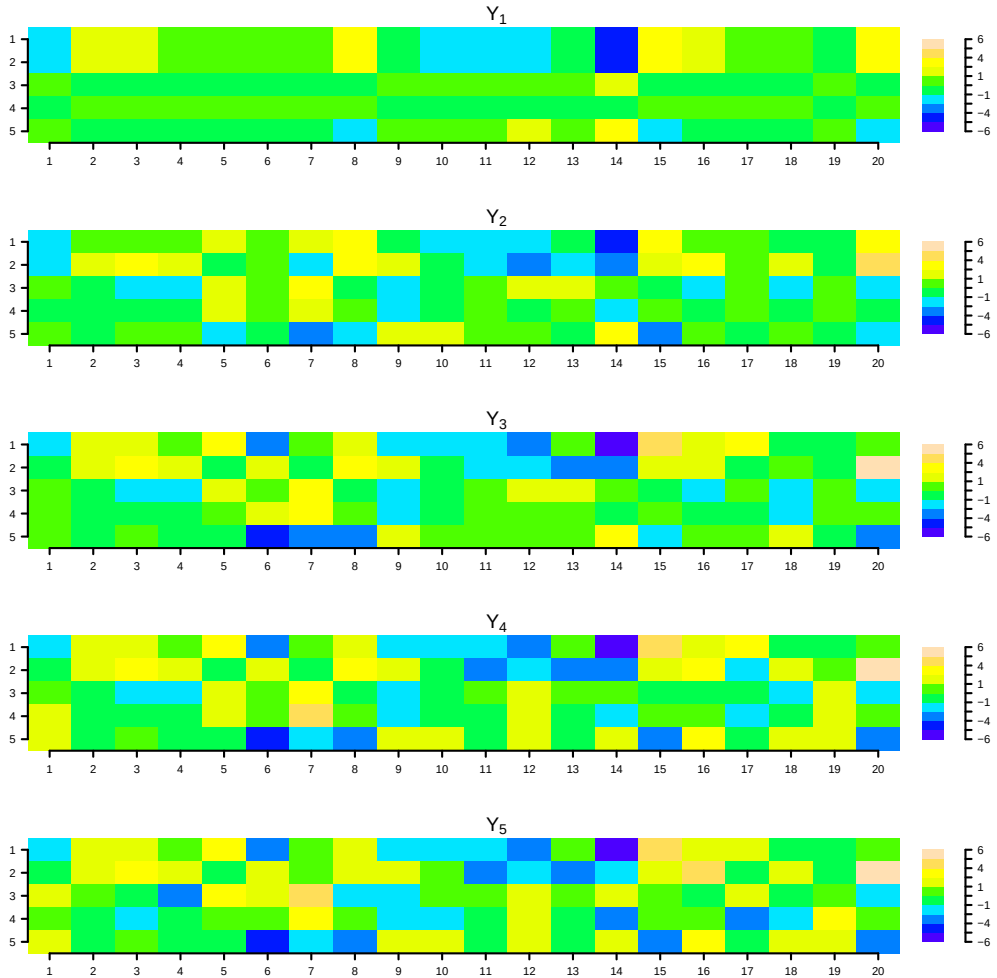


Datensatzrekonstruktion und -rekonstruktionsfehler eines simulierten Datensatzes

Scree (engl.) Schutthalde

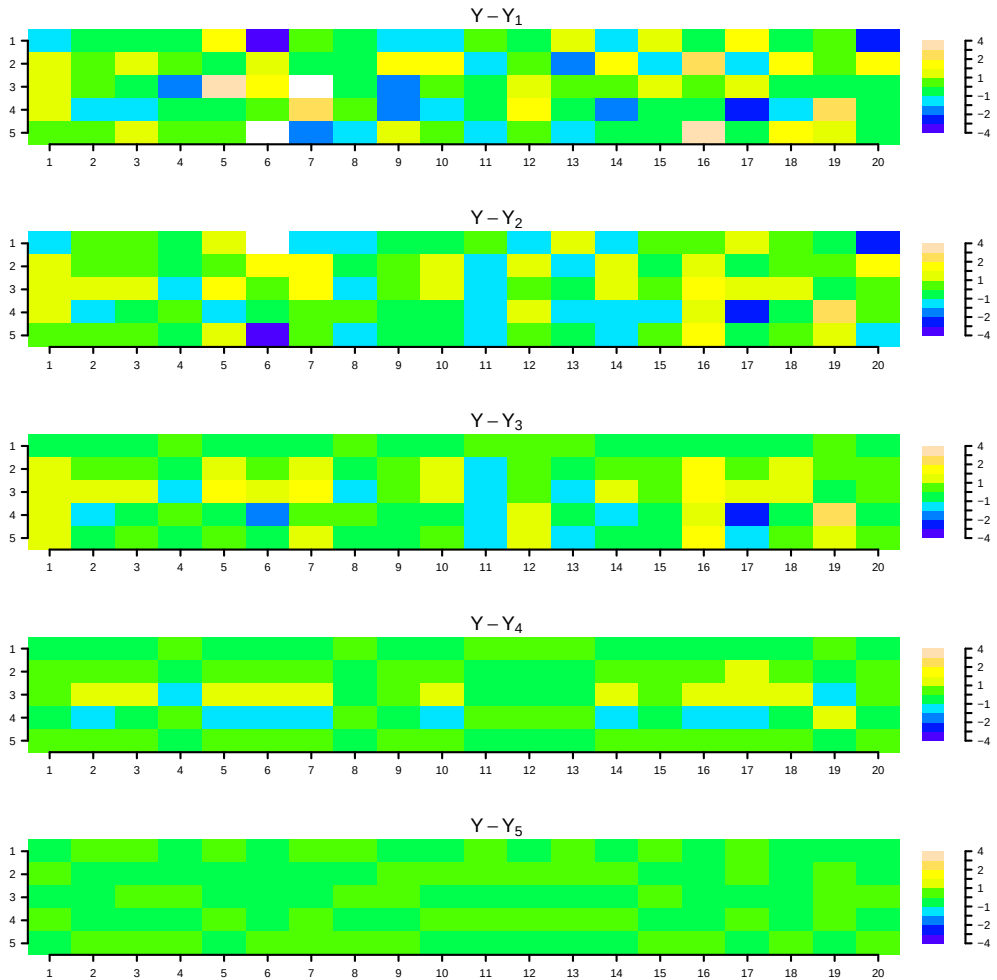
Datensatzrekonstruktion und -rekonstruktionsfehler eines simulierten Datensatzes

Rekonstruierte Datensätze



Datensatzrekonstruktion und -rekonstruktionsfehler eines simulierten Datensatzes

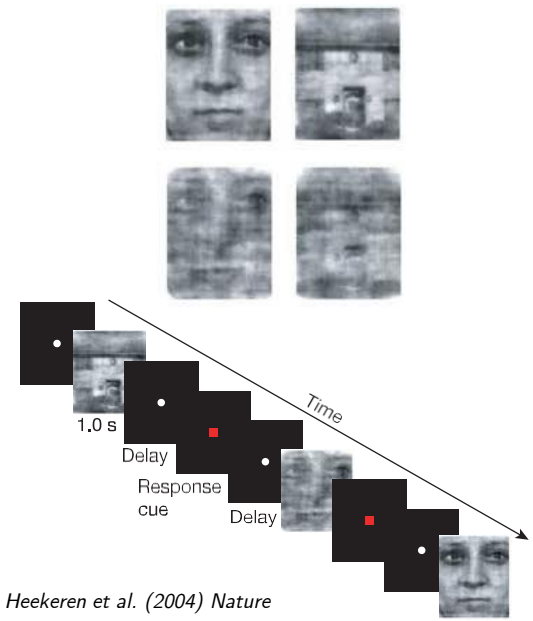
Originaldatensatz minus rekonstruierter Datensatz



Featureselektion bei EEG Daten

Wie werden visuelle Stimuli im Gehirn verarbeitet?

Wie entscheiden Menschen, ob sie ein Haus oder ein Gesicht wahrnehmen?



Heekeren et al. (2004) Nature

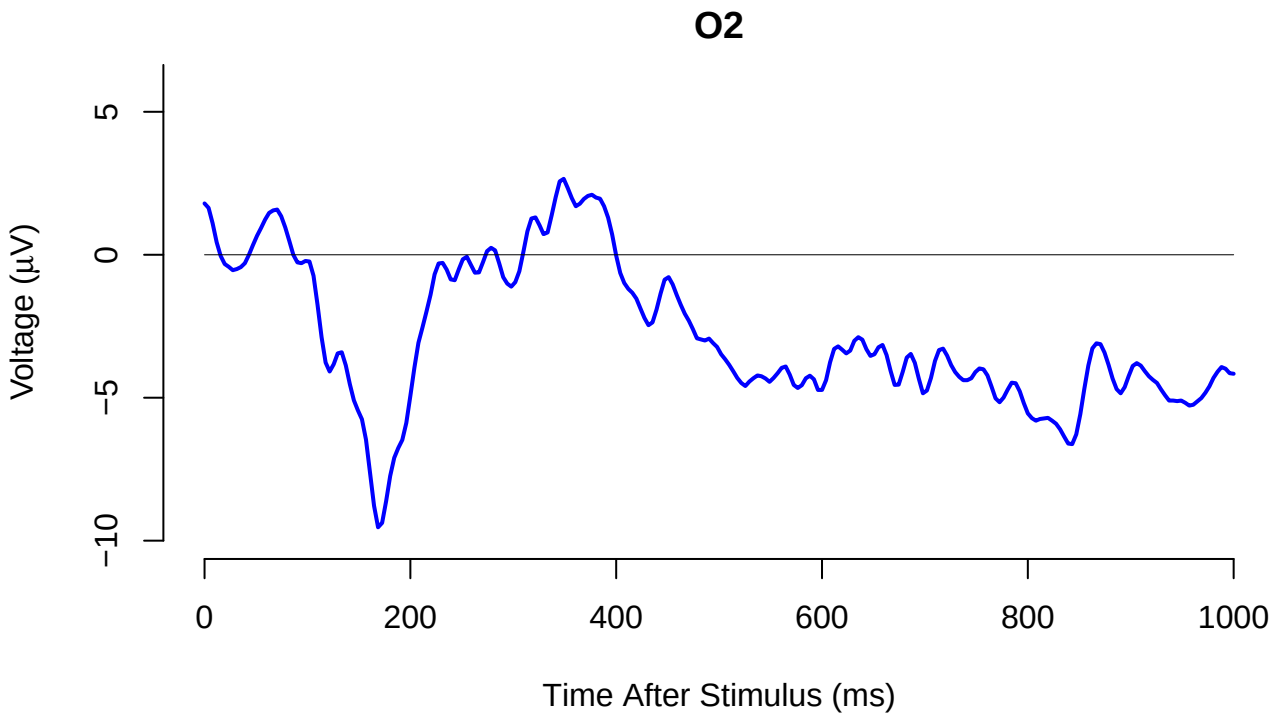
→ Allgemeine Psychologie, Biologische Psychologie, Kognitive Neurowissenschaften

Featureselektion bei EEG Daten

- In der prädiktiven EEG Analyse werden mentale Zustände aus Raumzeitserien dekodiert.
- Dabei wird ein Klassifikationsalgorithmus anhand von Single-Trial-Daten trainiert.
- Oft ist die Datendimension (Elektroden $\times$ Perieventsamples) sehr viel höher als die Trialanzahl.
- Wegen des Curse of Dimensionality performen Classifier auf Rohdatensätzen nicht optimal.
- Im Rahmen der Featureselektion können redundante Features durch PCA entfernt werden.
- Featureselektion entspricht einer Dimensionsreduktion und kann Klassifikationsraten erhöhen.
- Wir visualisieren das Vorgehen exemplarisch für eine Single-Trial-Elektrodenraum-Zeitserie.

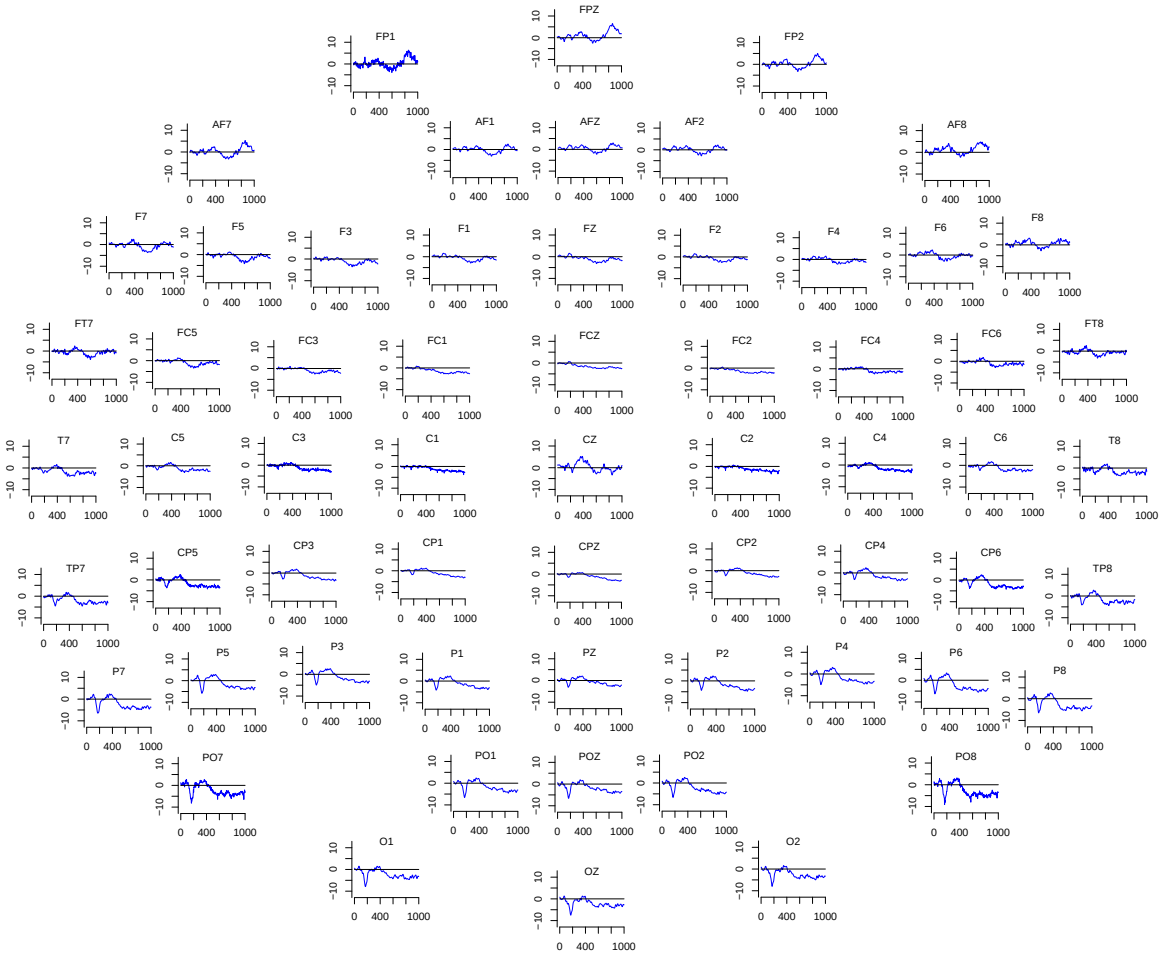
Featureselektion bei EEG Daten

Single-Trial Evoziertes Potential einer Elektrode



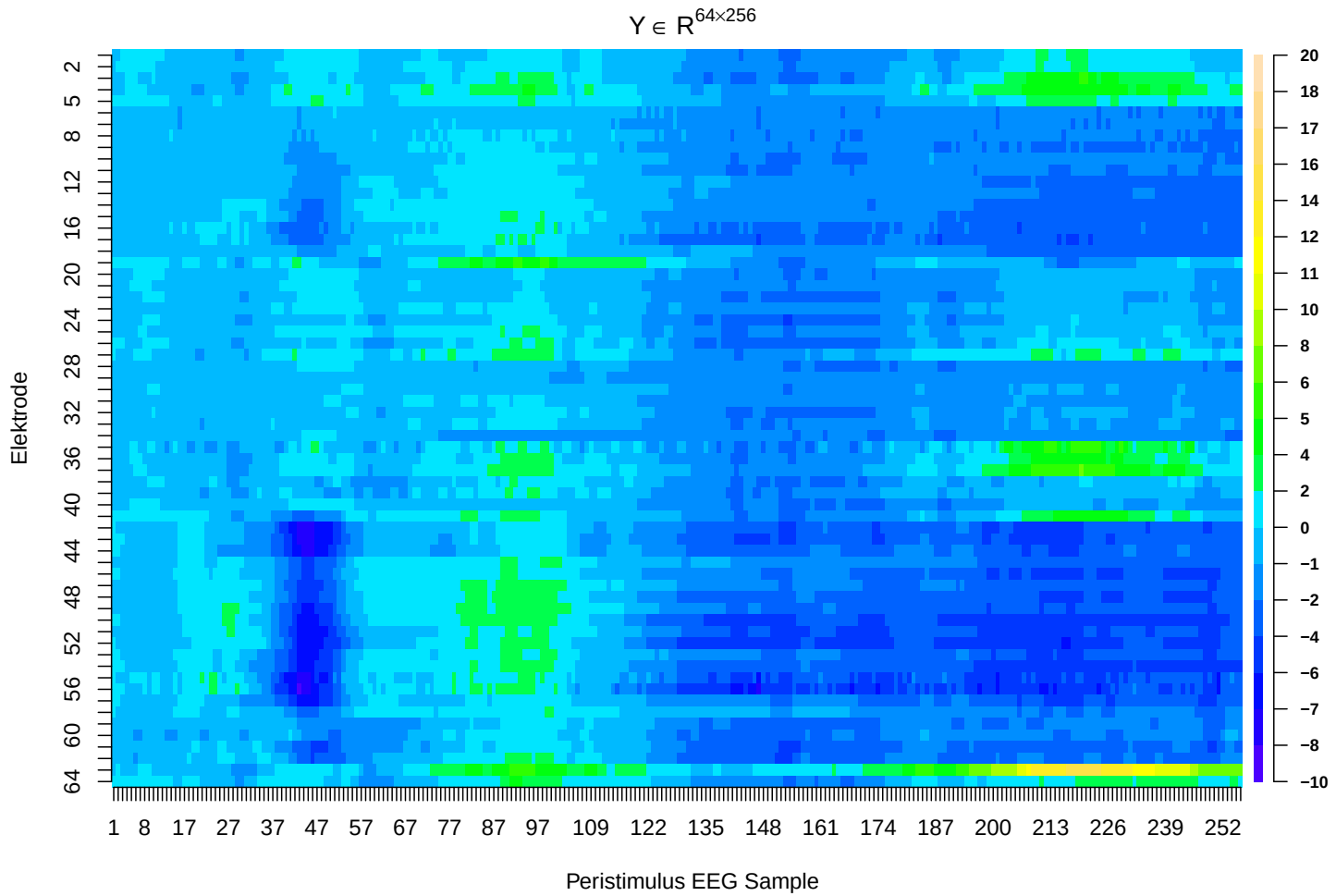
Featureselektion bei EEG Daten

Single-Trial Evoziertes Potential aller Elektroden



Featureselektion bei EEG Daten

Single-Trial Evoziertes Potential aller Elektroden in Matrixform



Featureselektion bei EEG Daten

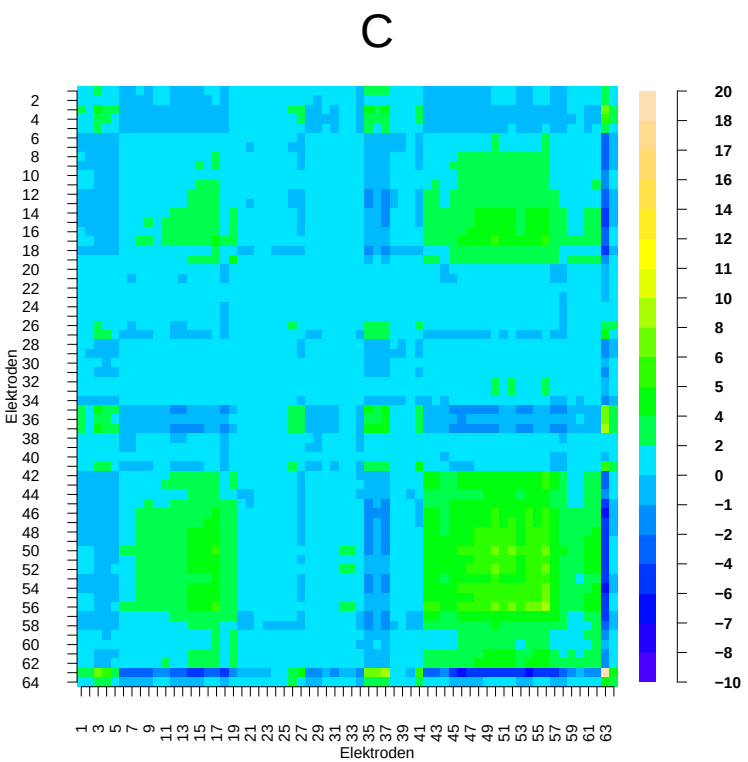
Hauptkomponentenanalyse

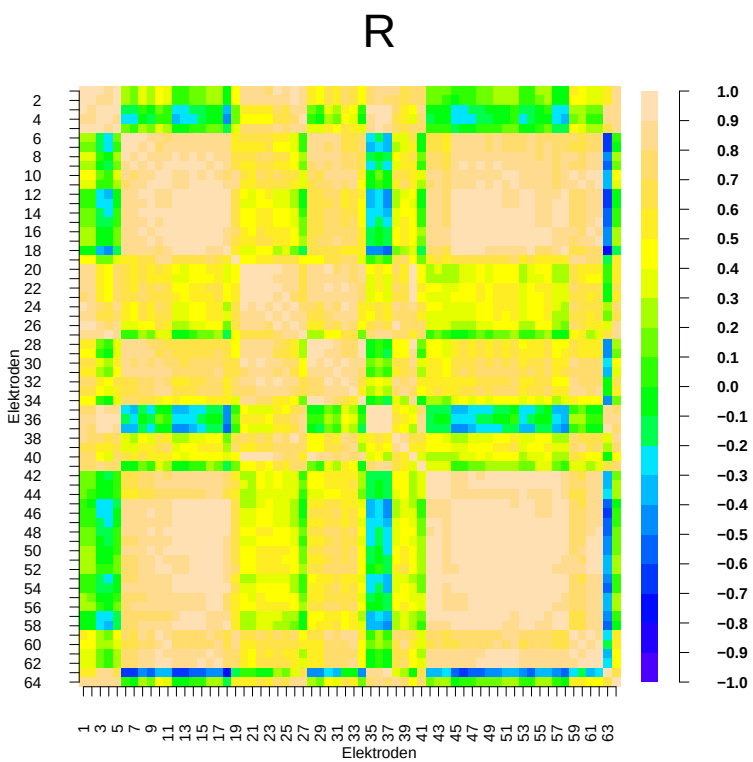
```
# Laden der Daten
Y       = as.matrix(readRDS(file.path(getwd(), "4_Daten", "eeg.csv")))

# Hauptkomponentenanalyse durch Eigenanalyse
m       = nrow(Y)                                    # Datendimension (Anzahl Elektroden)
n       = ncol(Y)                                    # Datenpunktanzahl (Anzahl Time-Bins)
I_n     = diag(n)                                    # Einheitsmatrix I_n
J_n     = matrix(rep(1,n^2), nrow = n)               # 1_{nn}
C       = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y)) # Stichprobenkovarianzmatrix
D       = diag(1/sqrt(diag(C)))                      # Kov-Korr-Transformationsmatrix
R       = D %*% C %*% D                              # Stichprobenkorrelationsmatrix
EA      = eigen(C)                                   # Eigenanalyse von C
lambda  = EA$values                                  # Eigenwerte von C
Q       = EA$vectors                                 # Eigenvektoren von C
Y_tilde = t(Q) %*% Y                                 # Transformierter Datensatz
```

Featureselektion bei EEG Daten

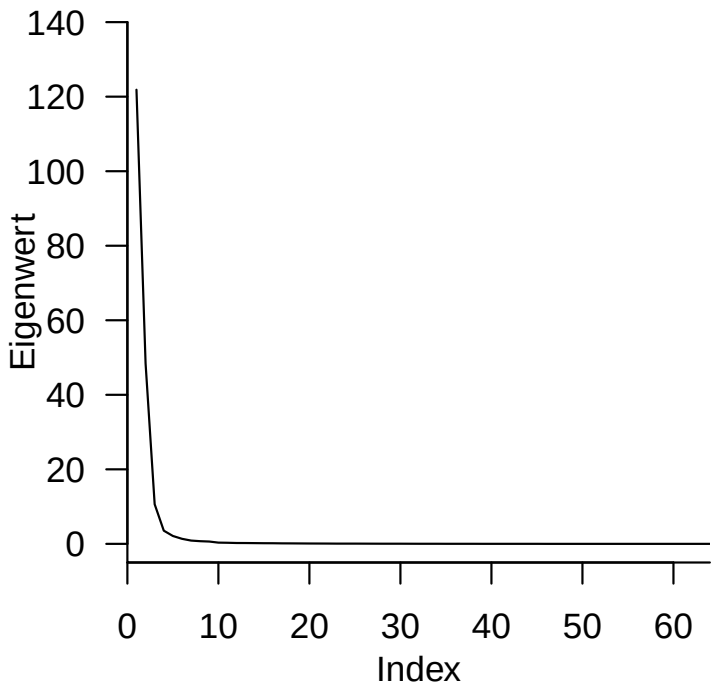
Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix

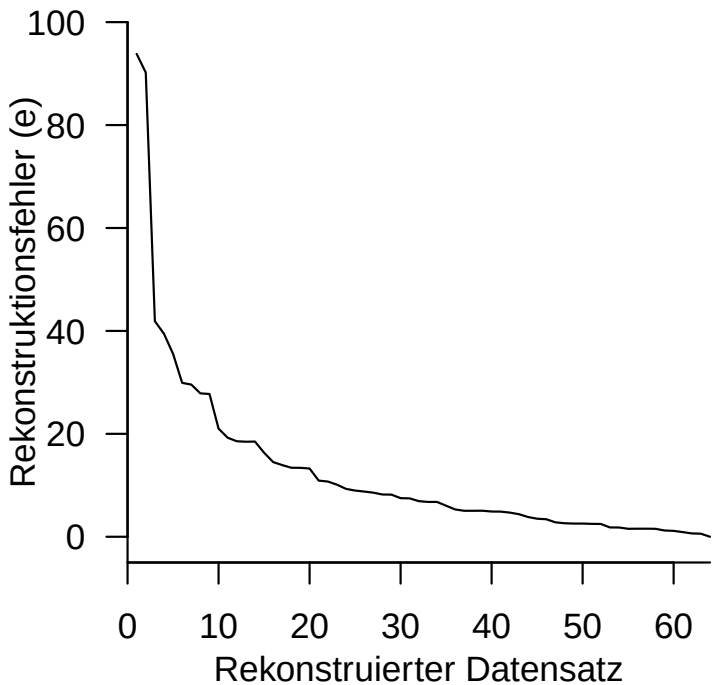




Featureselektion bei EEG Daten

Eigenwerte ("Scree (engl. Schutthalde)-Plot") und Rekonstruktionsfehler





Vektorkoordinatentransformation

Definition

Singulärwertzerlegung

Datenkompression

Exploratorische Faktorenanalyse

Selbstkontrollfragen

Überblick

Faktorenanalyse (oder Faktoranalyse) ist ein Sammelbegriff für viele datenanalytische Verfahren.

Prinzipiell entsprechen "Faktoren" latenten Zufallsvariablen in probabilistischen Modellen.

Latente Zufallsvariablen sind nur indirekt beobachtbar.

In der Psychologie dienen latente Zufallsvariablen oft der Modellierung von "Konstrukten".

Wir betrachten probabilistische Modelle mit latenten Zufallsvariablen in (5) Faktorenanalyse.

"Exploratorische Faktorenanalyse (EFA)" ist eine intuitive Vorform des Latent-Variable-Modellings.

Im Wesentlichen entspricht EFA einer speziellen Interpretation einer Hauptkomponentenanalyse.

Wir erläutern EFA hier anhand der Sedimentationshypothese der Persönlichkeitspsychologie.

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

- Alle wichtigen Persönlichkeitseigenschaften sind durch Adjektive repräsentiert.
- Persönlichkeitsadjektive haben sich analog zu Persönlichkeitseigenschaften entwickelt
- Persönlichkeitsadjektive decken die relevanten individuellen Differenzen ab.
- $\Rightarrow$ Big-Five der Persönlichkeitspsychologie (OCEAN-Modell)

Datengeneration

- 30 Proband:innen schätzen eine Person hinsichtlich des Zutreffens von Adjektiven ein.
- Neunstufige Skala (1: trifft überhaupt nicht zu, 9: trifft voll zu)
- Hochkorrelierte Ajdektive beschreiben eine "übergeordnete Persönlichkeitseigenschaften".

Rudolf & Buse (2020) Multvariate Verfahren Kapitel 9

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

Datensatz (Proband:innen 1 bis 15)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
angriffslustig	4	9	5	8	7	2	4	3	1	4	2	6	4	8	7
penibel	6	3	1	1	3	4	3	5	4	7	6	4	8	7	8
streitbar	3	8	4	6	6	3	3	3	1	3	2	4	3	7	7
kämpferisch	4	6	2	8	7	3	4	4	5	6	5	7	5	8	8
grimmig	5	4	3	4	4	5	5	4	3	4	2	3	5	7	6
gründlich	5	2	2	2	3	4	5	5	4	6	6	5	6	6	7
akkurat	5	2	1	1	3	4	4	5	4	6	6	5	7	5	5
gewissenhaft	1	2	3	2	4	3	6	4	4	5	6	2	5	4	4
kleinlich	5	1	1	1	3	1	2	1	2	3	6	1	1	7	8
übergenau	6	1	1	1	3	4	3	5	4	7	6	1	1	7	8
herausfordernd	4	1	2	8	7	3	4	4	1	1	2	3	5	7	6
hitzig	5	4	3	4	4	5	6	4	3	4	2	3	5	7	6

Rudolf & Buse (2020) Multvariate Verfahren Kapitel 9

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

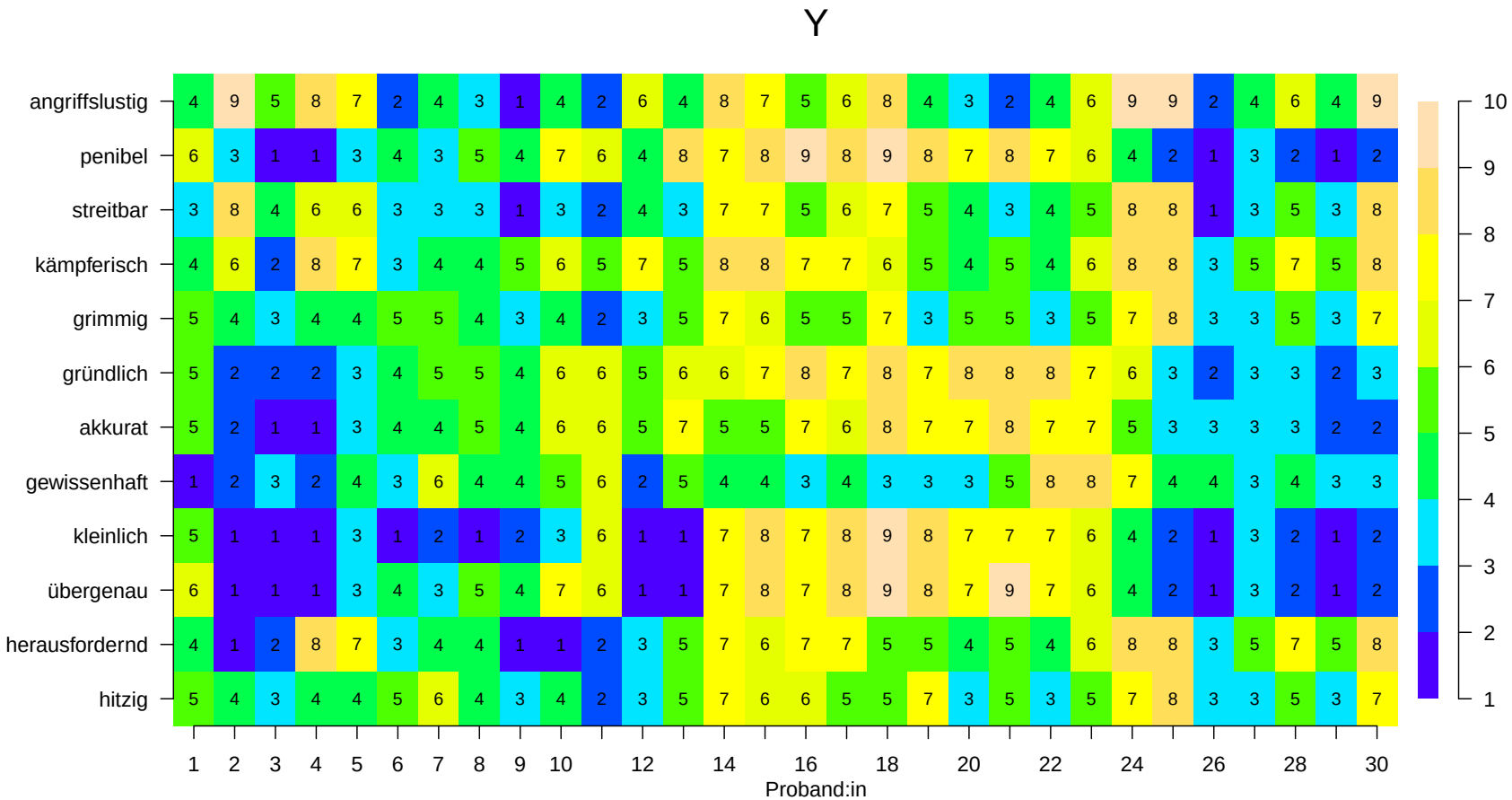
Datensatz (Proband:innen 16 bis 30)

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
angriffslustig	5	6	8	4	3	2	4	6	9	9	2	4	6	4	9
penibel	9	8	9	8	7	8	7	6	4	2	1	3	2	1	2
streitbar	5	6	7	5	4	3	4	5	8	8	1	3	5	3	8
kämpferisch	7	7	6	5	4	5	4	6	8	8	3	5	7	5	8
grimmig	5	5	7	3	5	5	3	5	7	8	3	3	5	3	7
gründlich	8	7	8	7	8	8	8	7	6	3	2	3	3	2	3
akkurat	7	6	8	7	7	8	7	7	5	3	3	3	3	2	2
gewissenhaft	3	4	3	3	3	5	8	8	7	4	4	3	4	3	3
kleinlich	7	8	9	8	7	7	7	6	4	2	1	3	2	1	2
übergenau	7	8	9	8	7	9	7	6	4	2	1	3	2	1	2
herausfordernd	7	7	5	5	4	5	4	6	8	8	3	5	7	5	8
hitzig	6	5	5	7	3	5	3	5	7	8	3	3	5	3	7

Rudolf & Buse (2020) Multvariate Verfahren Kapitel 9

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

Datensatz



Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

```
# Hauptkomponentenanalyse durch Eigenanalyse
m       = nrow(Y)                                  # Variablenanzahl (Anzahl Adjektive)
n       = ncol(Y)                                  # Datenpunktanzahl (Anzahl Proband:innen)
I_n     = diag(n)                                  # Einheitsmatrix I_n
J_n     = matrix(rep(1,n^2), nrow = n)             # 1_{nn}
C       = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))  # Stichprobenkovarianzmatrix
D       = diag(1/sqrt(diag(C)))                    # Kov-Korr-Transformationsmatrix
R       = D %*% C %*% D                            # Stichprobenkorrelationsmatrix
EA      = eigen(C)                                 # Eigenanalyse von C
lambda  = EA$values                                # Eigenwerte von C
Q       = EA$vectors                               # Eigenvektoren von C
Y_tilde = t(Q) %*% Y                               # Transformierter Datensatz

# Stichproben- und Korrelationsmatrix des transformierten Datensatzes
C_tilde = (1/(n-1))*(Y_tilde %*% (I_n-(1/n)*J_n) %*% t(Y_tilde))
D_tilde = diag(1/sqrt(diag(C_tilde)))
R_tilde = D_tilde %*% C_tilde %*% D_tilde
```

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

Stichprobenkovarianzmatrix und Stichprobenkorrelationsmatrix

C

	angriffslustig	penibel	streitbar	kämpferisch	grimmig	gründlich	akkurat	gewissenhaft	kleinlich	übergenau	herausfordernd	hitzig
angriffslustig	+5.9	−0.9	+4.8	+3.2	+2.3	−0.8	−1.4	−0.4	+0.3	−1.1	+3.2	+2.1
penibel	−0.9	+7.1	+0.2	+0.3	+0.8	+5.3	+5.1	+0.9	+6.1	+6.4	+0.0	+0.8
streitbar	+4.8	+0.2	+4.4	+2.7	+2.2	+0.1	−0.5	−0.3	+1.4	+0.3	+2.9	+2.1
kämpferisch	+3.2	+0.3	+2.7	+3.0	+1.4	+0.2	−0.2	−0.0	+0.9	+0.2	+2.6	+1.4
grimmig	+2.3	+0.8	+2.2	+1.4	+2.3	+0.7	+0.4	+0.1	+1.0	+0.8	+2.1	+1.9
gründlich	−0.8	+5.3	+0.1	+0.2	+0.7	+4.6	+4.2	+1.4	+5.0	+5.2	+0.4	+0.7
akkurat	−1.4	+5.1	−0.5	−0.2	+0.4	+4.2	+4.3	+1.4	+4.3	+4.6	−0.2	+0.4
gewissenhaft	−0.4	+0.9	−0.3	−0.0	+0.1	+1.4	+1.4	+2.8	+1.1	+1.1	+0.2	+0.1
kleinlich	+0.3	+6.1	+1.4	+0.9	+1.0	+5.0	+4.3	+1.1	+8.0	+7.2	+1.4	+1.1
übergenau	−1.1	+6.4	+0.3	+0.2	+0.8	+5.2	+4.6	+1.1	+7.2	+7.8	+0.3	+0.9
herausfordernd	+3.2	+0.0	+2.9	+2.6	+2.1	+0.4	−0.2	+0.2	+1.4	+0.3	+4.8	+2.2
hitzig	+2.1	+0.8	+2.1	+1.4	+1.9	+0.7	+0.4	+0.1	+1.1	+0.9	+2.2	+2.4

R

	angriffslustig	penibel	streitbar	kämpferisch	grimmig	gründlich	akkurat	gewissenhaft	kleinlich	übergenau	herausfordernd	hitzig
angriffslustig	+1.0	−0.1	+0.9	+0.8	+0.6	−0.2	−0.3	−0.1	+0.0	−0.2	+0.6	+0.5
penibel	−0.1	+1.0	+0.0	+0.1	+0.2	+0.9	+0.9	+0.2	+0.8	+0.9	+0.0	+0.2
streitbar	+0.9	+0.0	+1.0	+0.7	+0.7	+0.0	−0.1	−0.1	+0.2	+0.1	+0.6	+0.6
kämpferisch	+0.8	+0.1	+0.7	+1.0	+0.6	+0.1	−0.0	−0.0	+0.2	+0.0	+0.7	+0.5
grimmig	+0.6	+0.2	+0.7	+0.6	+1.0	+0.2	+0.1	+0.0	+0.2	+0.2	+0.6	+0.8
gründlich	−0.2	+0.9	+0.0	+0.1	+0.2	+1.0	+1.0	+0.4	+0.8	+0.9	+0.1	+0.2
akkurat	−0.3	+0.9	−0.1	−0.0	+0.1	+1.0	+1.0	+0.4	+0.7	+0.8	−0.0	+0.1
gewissenhaft	−0.1	+0.2	−0.1	−0.0	+0.0	+0.4	+0.4	+1.0	+0.2	+0.2	+0.1	+0.1
kleinlich	+0.0	+0.8	+0.2	+0.2	+0.2	+0.8	+0.7	+0.2	+1.0	+0.9	+0.2	+0.2
übergenau	−0.2	+0.9	+0.1	+0.0	+0.2	+0.9	+0.8	+0.2	+0.9	+1.0	+0.0	+0.2
herausfordernd	+0.6	+0.0	+0.6	+0.7	+0.6	+0.1	−0.0	+0.1	+0.2	+0.0	+1.0	+0.6
hitzig	+0.5	+0.2	+0.6	+0.5	+0.8	+0.2	+0.1	+0.1	+0.2	+0.2	+0.6	+1.0

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

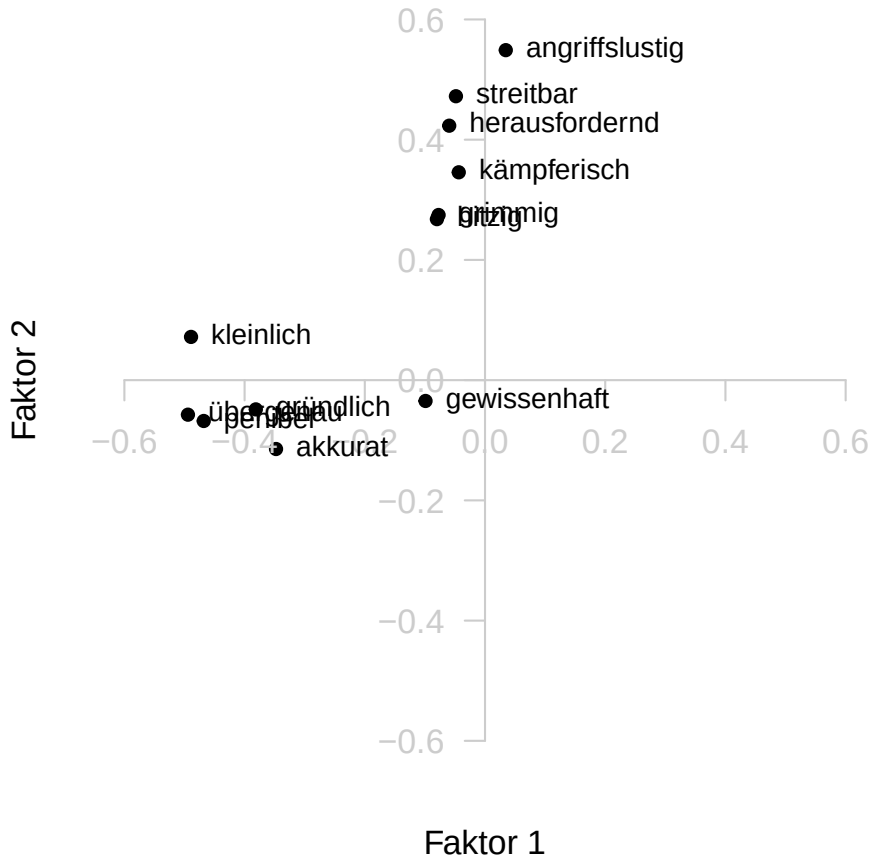
Ladungen (= Komponenten) der Faktoren (= Eigenvektoren) mit den höchsten beiden Eigenwerten

```
# "Ladungen" der "Faktoren" mit den höchsten beiden Eigenwerten
L = Q[, 1:2]
```

	Faktor 1	Faktor 2
angriffslustig	0.035	0.549
penibel	-0.468	-0.068
streitbar	-0.048	0.472
kämpferisch	-0.044	0.346
grimmig	-0.077	0.275
gründlich	-0.381	-0.049
akkurat	-0.348	-0.115
gewissenhaft	-0.099	-0.035
kleinlich	-0.489	0.072
übergenau	-0.494	-0.057
herausfordernd	-0.060	0.423
hitzig	-0.080	0.268

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

Ladungen (Komponenten) der Faktoren (Eigenvektoren) mit den höchsten beiden Eigenwerten

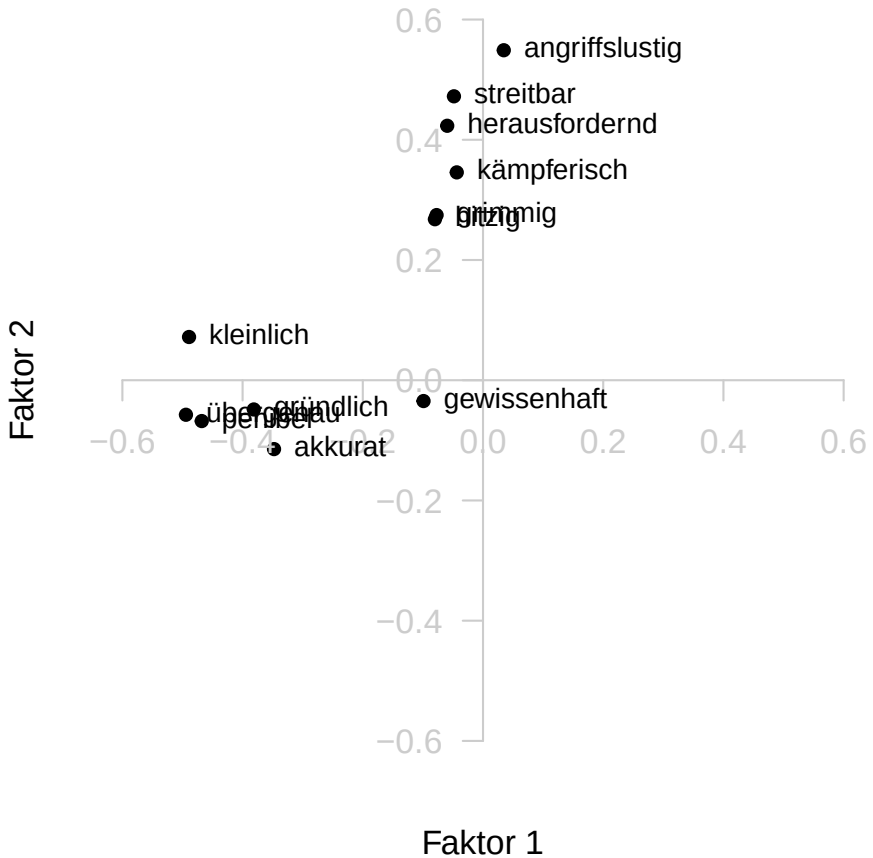


Angriffslustig, streitbar, herausfordernd, kämpferisch, hitzig ≈ Hohe Werte Faktor 1, Niedrige Werte Faktor 2

Kleinlich, akkurat, übergenau, gründlich, gewissenhaft, penibel ≈ Niedrige Werte Faktor 1, Hohe Werte Faktor 2

Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

Ladungen (Komponenten) der Faktoren (Eigenvektoren) mit den höchsten beiden Eigenwerten

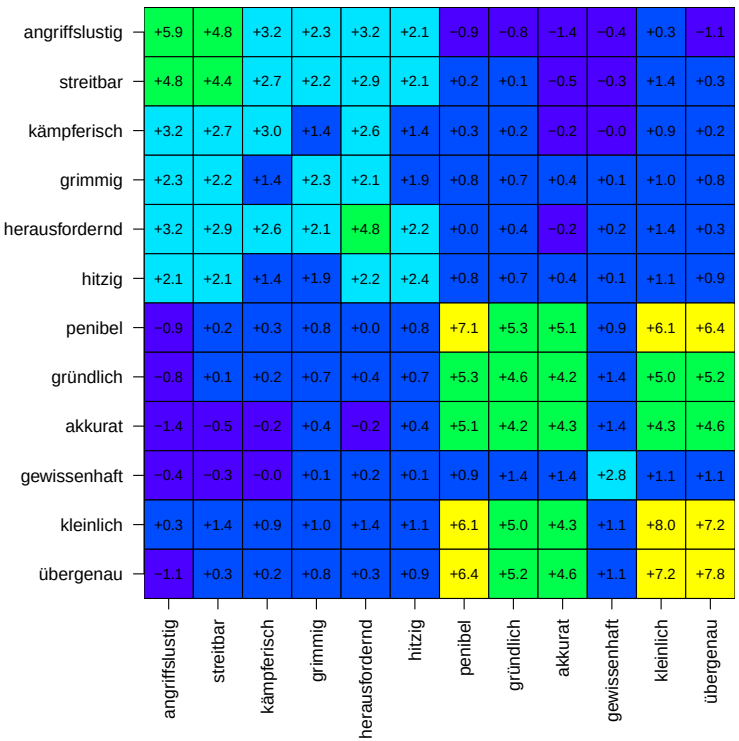


Faktor 1 = Aggressivität und Faktor 2 = Perfektionismus als Persönlichkeitseigenschaften

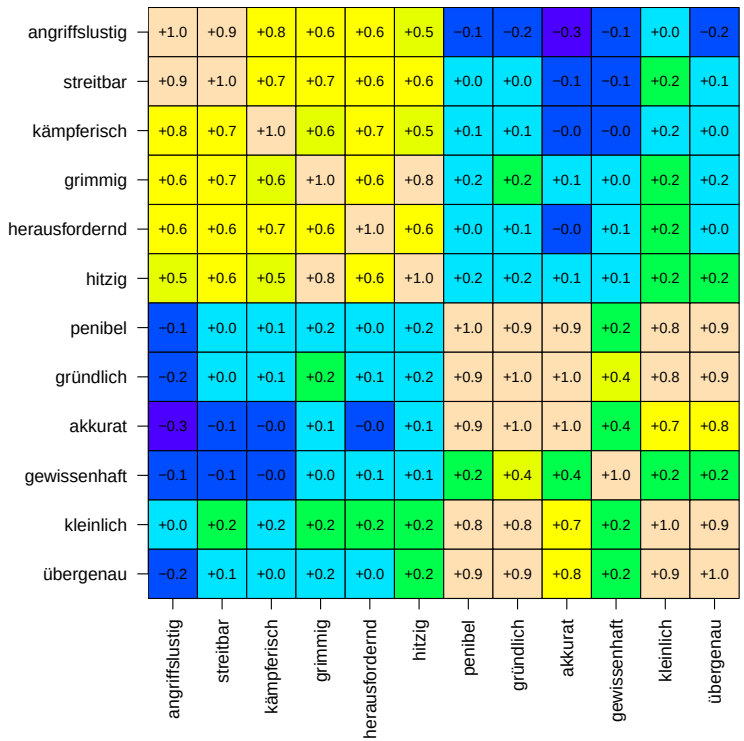
Sedimentationshypothese (Lexikalischer Ansatz) der Persönlichkeitspsychologie

. . . es ginge auch direkter . . .

C



R



Vektorkoordinatentransformation

Definition

Singulärwertzerlegung

Datenkompression

Exploratorische Faktorenanalyse

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie den Begriff “Featureselektion”.
2. Definieren Sie den Begriff Orthogonalprojektion.
3. Geben Sie das Theorem zu Vektorkoordinaten bezüglich einer Orthogonalbasis wieder.
4. Geben Sie das Vektorkoordinatentransformationstheorem wieder.
5. Erläutern Sie das Vektorkoordinatentransformationstheorem.
6. Geben Sie die Definition einer Hauptkomponentenanalyse wieder.
7. Geben Sie das Theorem zur Hauptkomponentenanalyse wieder.
8. Geben Sie die Definition der Hauptkomponentenanalyse eines Datensatzes wieder.
9. Geben Sie das Theorem zur Hauptkomponentenanalyse eines Datensatzes wieder.
10. Schreiben Sie R Code zur Implementation einer Hauptkomponentenanalyse durch Eigenanalyse.
11. Geben Sie die Definition einer Singulärwertzerlegung wieder.
12. Geben Sie das Theorem zum Zusammenhang von Singulärwertzerlegung und Eigenanalyse wieder.

Selbstkontrollfragen

13. Geben Sie das Theorem zur Datenhauptkomponentenanalyse durch Singulärwertzerlegung wieder.
14. Schreiben Sie R Code zur Implementation einer Hauptkomponentenanalyse durch Singulärwertzerlegung.
15. Erläutern Sie das Prinzip der Datenkompression durch Hauptkomponentenanalyse
16. Definieren Sie den Begriff des PCA-dimensionreduzierten Datensatzes.
17. Definieren Sie den Begriff des (PCA)-rekonstruierten Datensatzes.
18. Definieren Sie den Begriff des (PCA)-Rekonstruktionsfehlers.
19. Erläutern Sie die Idee eines Scree-Plots.
20. Erläutern Sie ein Beispiel zur Datendimensionreduktion in der Analyse von EEG Daten.
21. Erläutern Sie den Begriff "Exploratorische Faktorenanalyse".
22. Erläutern Sie die Idee der Sedimenationshypothese/des lexikalischen Ansatzes der Persönlichkeitspsychologie.
23. Erläutern Sie, wie man mithilfe einer Hauptkomponentenanalyse und eines Datensatzes von Einschätzungen einer ihnen bekannten Person durch Proband:innen hinsichtlich einer Menge von Adjektiven zur Identifikation von latenten Persönlichkeitseigenschaften gelangen kann.