Multivariate Datenanalyse

MSc Psychologie WiSe 2022/23

Prof. Dr. Dirk Ostwald

(9) Hauptkomponentenanalyse

Datum	Einheit	Thema
14.10.2022	Grundlagen	(1) Einführung
21.10.2022	Grundlagen	(2) Vektoren
28.10.2022	Grundlagen	(3) Matrizen
04.11.2022	Grundlagen	(4) Eigenanalyse
11.11.2022	Grundlagen	(5) Multivariate Wahrscheinlichkeitstheorie
18.11.2022	Grundlagen	(6) Multivariate Normalverteilungen
25.11.2022	Frequentistische Inferenz	(7) Kanonische Korrelationsanalyse
02.12.2022	Frequentistische Inferenz	(8) T^2-Tests I
09.12.2022	Frequentistische Inferenz	(8) T^2-Tests II
16.12.2022	Latente Variablenmodelle	(9) Hauptkomponentenanalyse
	Weihnachtspause	
13.01.2023	Latente Variablenmodelle	(10) Lineare Normalverteilungsmodelle
20.01.2023	Latente Variablenmodelle	(11) Faktorenanalyse I
27.01.2023	Latente Variablenmodelle	(11) Faktorenanalyse II
Sommer 2023	Klausur	

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 Faktorenanalyse.

Vektorkoordinatentransformation

Definition und Theorem

Datenkompression

Selbstkontrollfragen

Vektorkoordinatentransformation

Definition und Theorem

Datenkompression

Selbstkontrollfragen

Vektorkoordinatentransformation

Euklidischer Vektorraum

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

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

- $B := \{b_1, \ldots, b_m\}$ sei eine Basis eines m-dimensionalen Vektorraumes V und es sei $v \in V$. Dann heißt die Linearkombination $v = \sum_{i=1}^{m} c_i b_i$ die *Darstellung von* v *bezüglich der Basis* B und die Koeffizienten $c_1, \ldots, c_m$ heißen die *Koordinaten von* v *bezüglich der Basis* B.

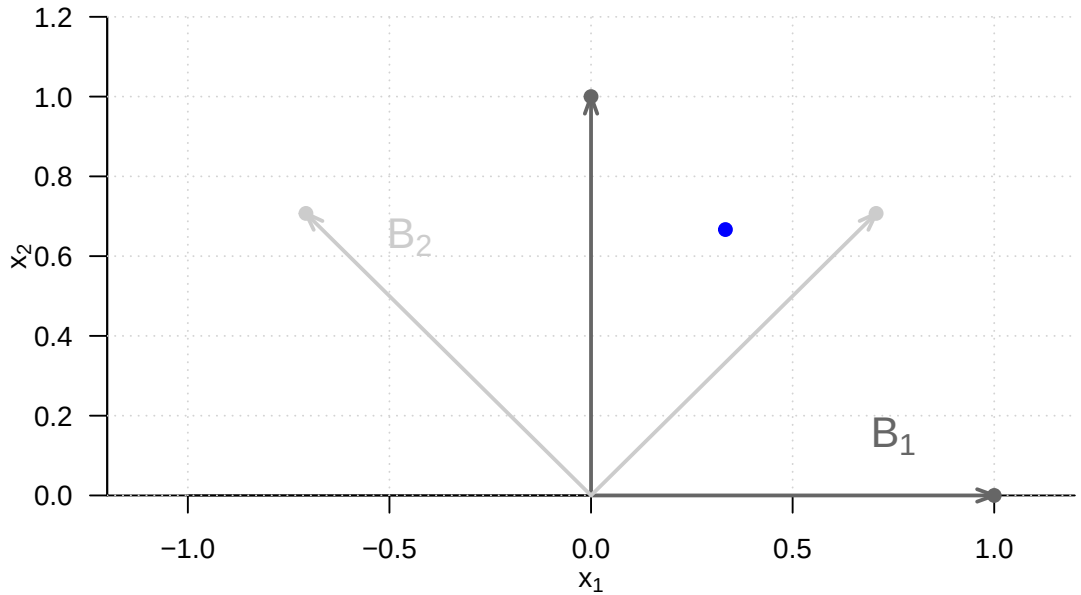
Orthonormalbasis von $\mathbb{R}^m$

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

Orthonormalzerlegung einer symmetrischen Matrix

- $S \in \mathbb{R}^{m \times m}$ sei eine symmetrische Matrix mit m verschiedenen Eigenwerten. 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.

Wiederholung von Begriffen aus (2) Vektoren, (3) Matrizen und (4) Eigenanalyse



Bei der Hauptkomponentenanalyse ist man daran interessiert, basierend auf den Koordinaten eines Vektors bezüglich einer Orthonormalbasis die Koordinaten desselben Vektors bezüglich einer anderen Basis zu berechnen. Mit der Orthogonalprojektion führen wir zunächst einen generellen Weg ein, die Koordinaten eines Vektors bezüglich einer Orthonormalbasis zu bestimmen. Wir geben dann ein Theorem an, um die Koordinaten eines Vektors bezüglich einer weiteren Orthonormalbasis zu bestimmen.

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.

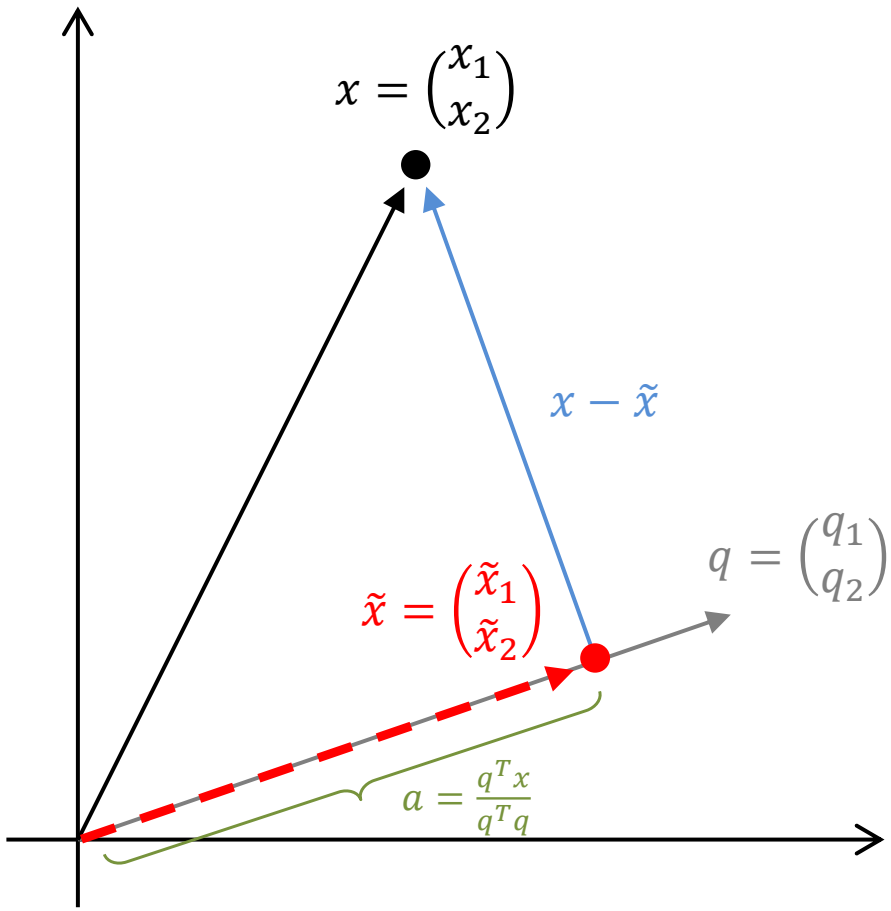
Bemerkungen

- 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 $\mathbb{R}^m$. Dann ergeben sich für $i = 1, \ldots, m$ die Koordinaten c_i in der Basisdarstellung von x bezüglich B als die Projektionsfaktoren

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

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

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

Bemerkung

- Hinsichtlich der kanonischen Basis von $\mathbb{R}^m$ ergibt sich offenbar $c_i = x^T e_i = x_i$ für $i = 1, \ldots, m$.

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 $\tilde{x}_1, \ldots, \tilde{x}_m$ des Vektors hinsichtlich der Basis B_w in die Koordinaten $x_1, \ldots, x_m$ des Vektors hinsichtlich B_v durch

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

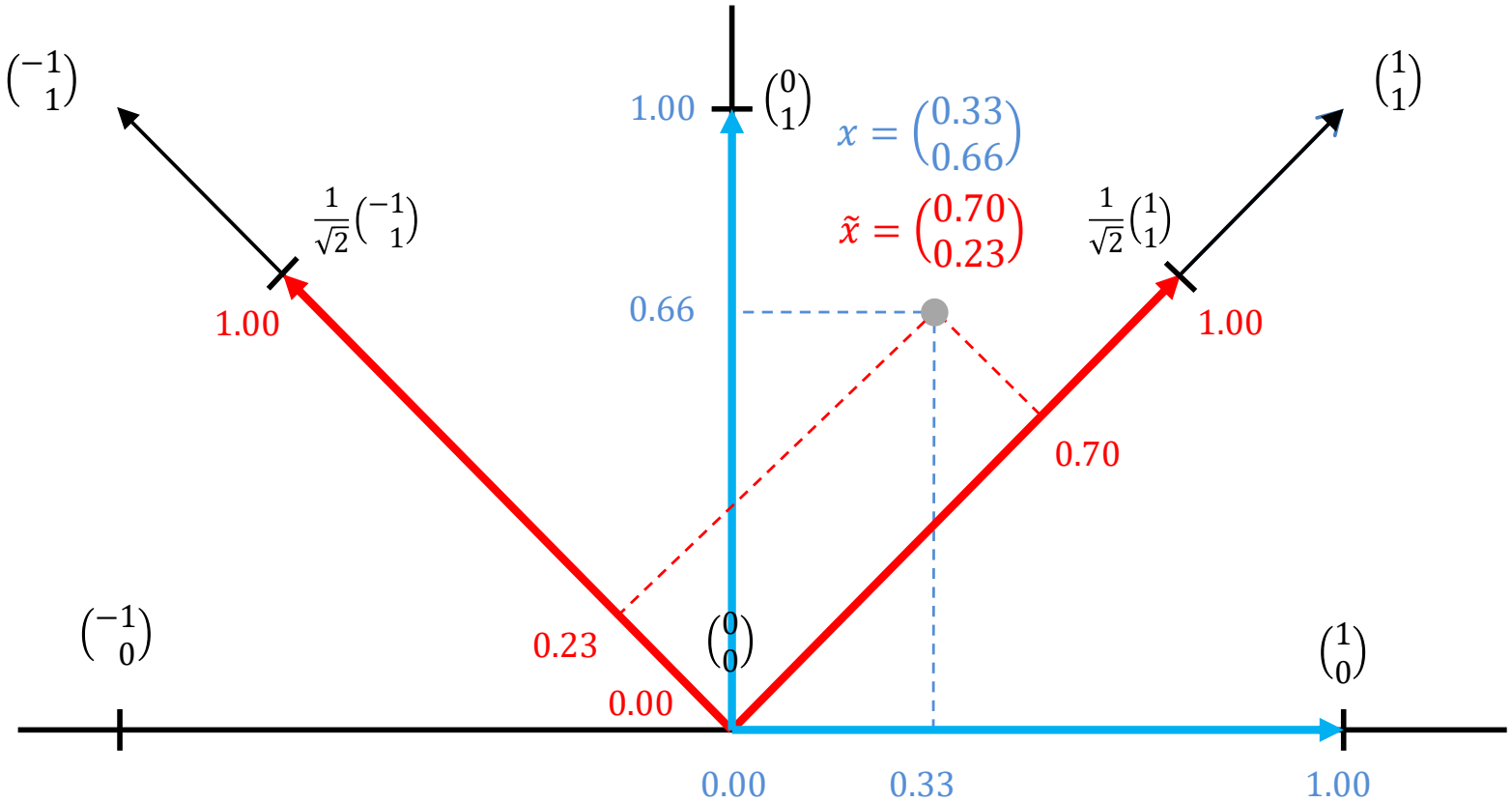
transformiert werden.

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 und $\tilde{x}$ 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

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

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

Selbstkontrollfragen

Definition (Hauptkomponentenanalyse)

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

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

wobei

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

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

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

heißt *PCA-transformierter Zufallsvektor*.

Bemerkungen

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

Theorem (Hauptkomponentenanalyse)

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

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

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

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

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

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

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

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

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

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

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

(3) Wir erinnern zunächst daran, dass die inverse Matrix einer orthogonalen Matrix Q durch Q^T gegeben ist. Mit $QQ^T = Q^T Q = I_m$ gilt dann, dass

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

Beweis (fortgeführt)

Weiterhin gilt, dass mit $\mathbb{E}(y) = 0_m$ die Kovarianzmatrix von y gegeben ist durch

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

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

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

(4) Die Koordinaten von $\tilde{y}$ entsprechen den Koordinaten von y in dem Koordinatensystem, dass von den Hauptkomponenten $q_1, \ldots, q_m$ von $\mathbb{C}(y)$ aufgespannt wird. Mit $\mathbb{C}(\tilde{y}) = \Lambda$ folgt Aussage (4) dann direkt.

□

Definition und Theorem

Bemerkungen

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

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

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

- Die paarweise nicht-identischen Kovarianzen der Komponenten von $\tilde{y}$ sind Null.
  - $\Rightarrow$ Die Komponenten von $\tilde{y}$ sind unkorreliert.
  - $\Rightarrow$ Die Komponenten von $\tilde{y}$ repräsentieren keine redundante Information.

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

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
```

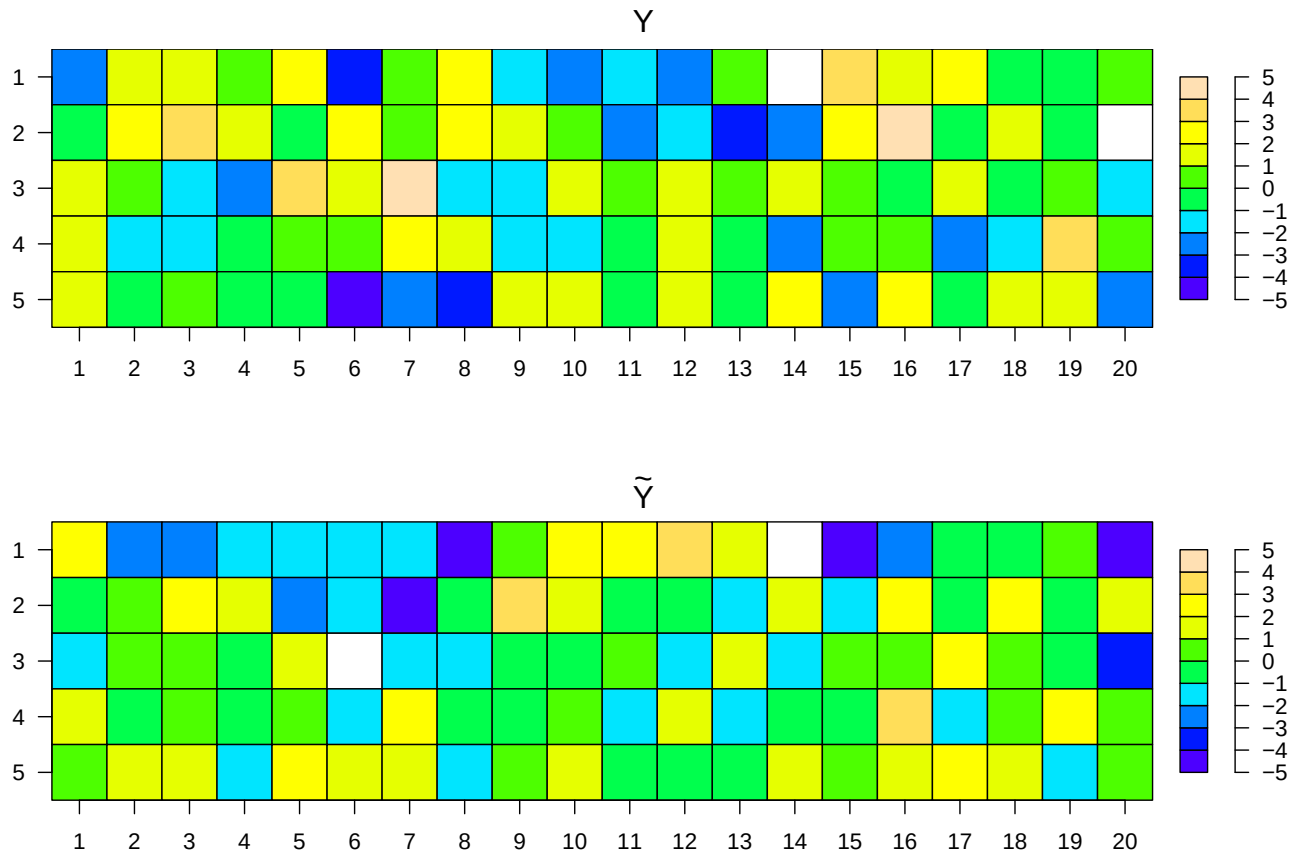
```
# Datensatzgeneration
Y     = t(mvrnorm(n,mu,Sigma))
```

Hauptkomponentenanalyse eines simulierten Datensatzes

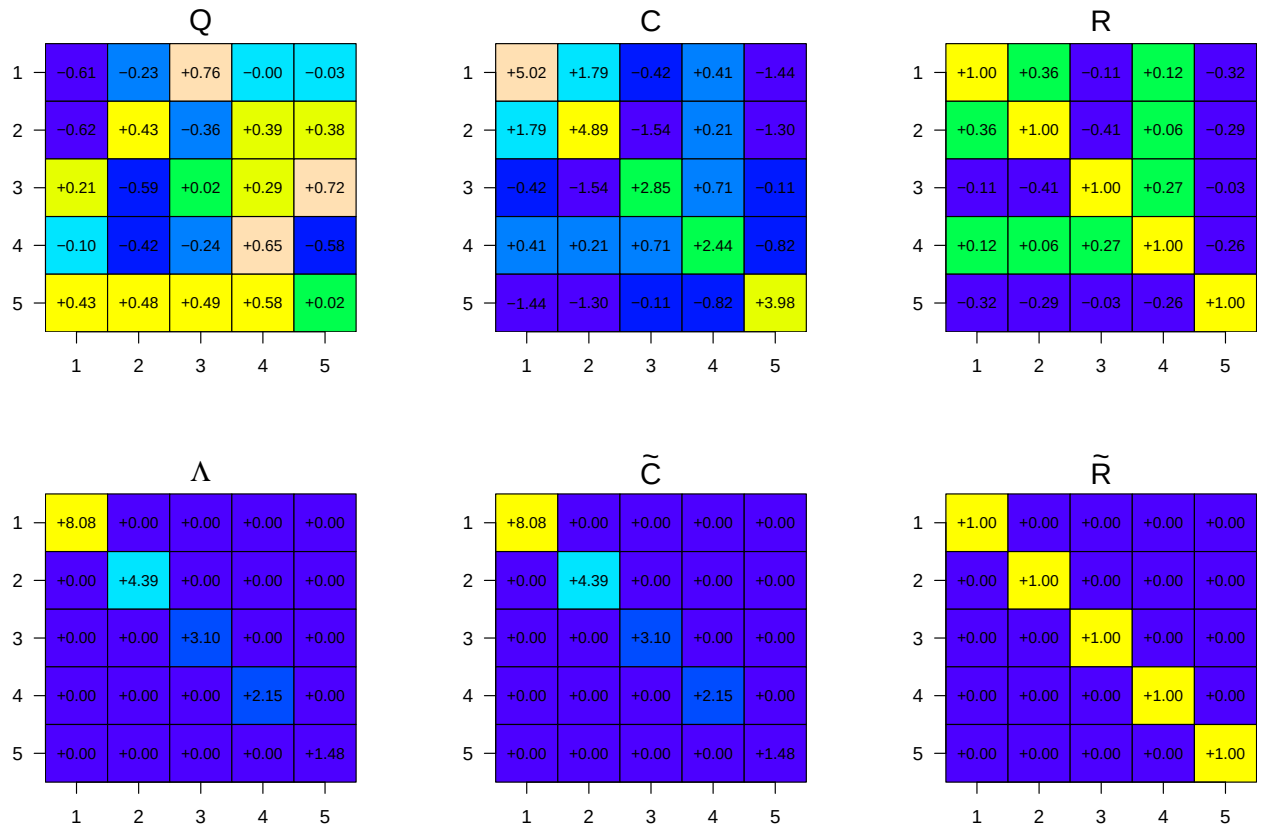
```
# Hauptkomponentenanalyse durch Eigenanalyse
I_n     = diag(n)                                       # Einheitsmatrix I_n
J_n     = matrix(rep(1,n^2), nrow = n)                  # 1_{nn}
C       = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))    # Stichprobenkovarianzmatrix
D       = diag(1/sqrt(diag(C)))                         # 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 eines simulierten Datensatzes



Vektorkoordinatentransformation

Definition und Theorem

Datenkompression

Selbstkontrollfragen

Überblick

- Datenkompression entspricht einer Reduktion der Dimension m von Daten.
- In der probabilistischen Modellierung wird PCA manchmal zur *Dimensionsreduktion* eingesetzt.
- Ziel ist es hier, dem Undersampling hochdimensionaler Datenräume entgegen zu wirken.
- Dimensionsreduktion entspricht dem Verwerfen von $k < m$ Hauptkomponenten von $\tilde{Y}$.
- Ziel der Auswahl von Hauptkomponenten mit hohen Eigenwerten ist es dabei, den *Datenrekonstruktionsfehler* möglichst klein zu halten.
- Im Rahmen der Exploratorischen Faktorenanalyse werden die nicht verworfenen Hauptkomponenten zu "psychologischen Faktoren" erhoben und sich an ihrer Interpretation versucht.

Definition (Dimensionsreduzierter Datensatz)

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

$$C = Q \Lambda Q^T \tag{27}$$

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

$$\tilde{Y}_k = Q_k^T Y \in \mathbb{R}^{k \times n} \tag{28}$$

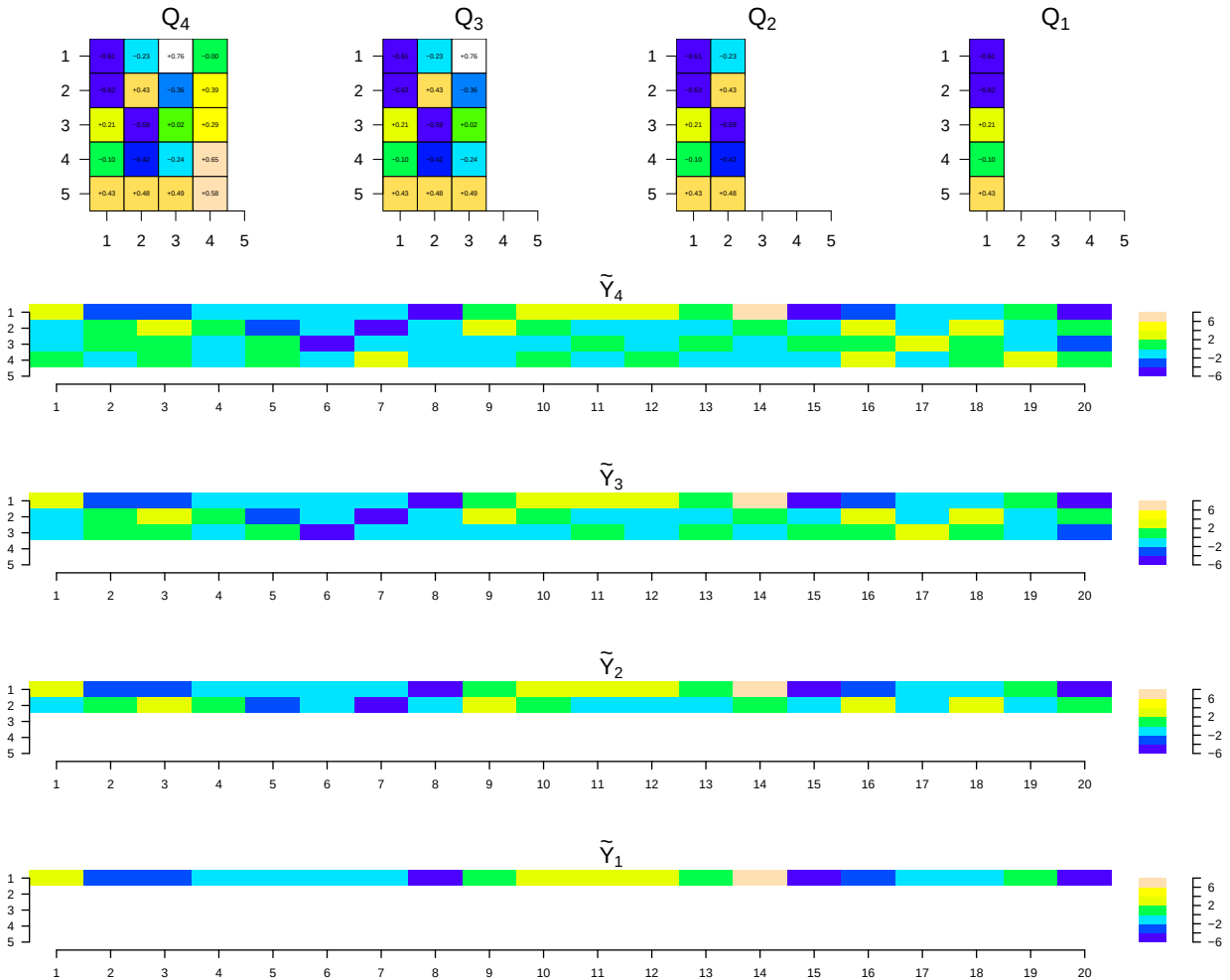
dimensionsreduzierter Datensatz.

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

$$\tilde{Y}_k = Q_k^T Y \in \mathbb{R}^{k \times n} \tag{29}$$

ein dimensionsreduzierter Datensatz. Dann heißt

$$Y_k = Q_k \tilde{Y}_k \in \mathbb{R}^{m \times n} \tag{30}$$

rekonstruierter Datensatz und

$$e = \|\text{vec}(Y - Y_k)\| \geq 0 \tag{31}$$

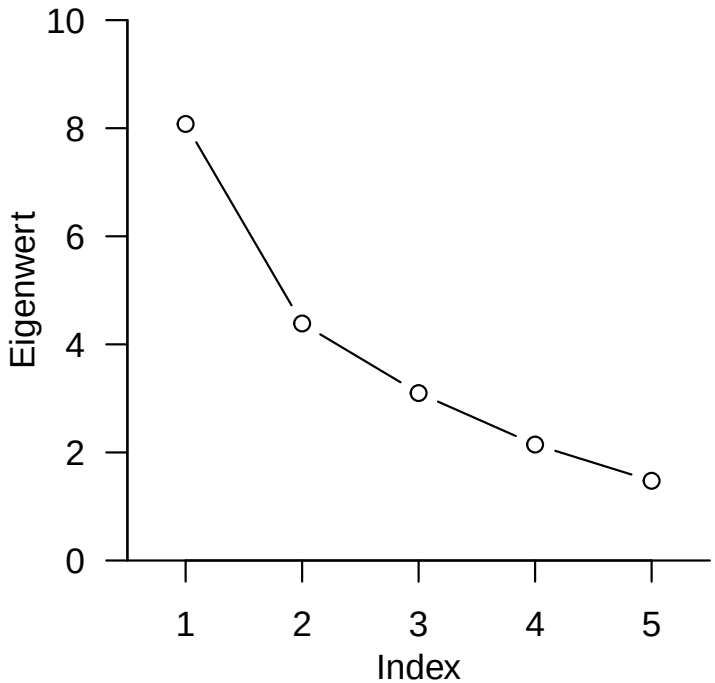
heißt *Datenrekonstruktionsfehler*.

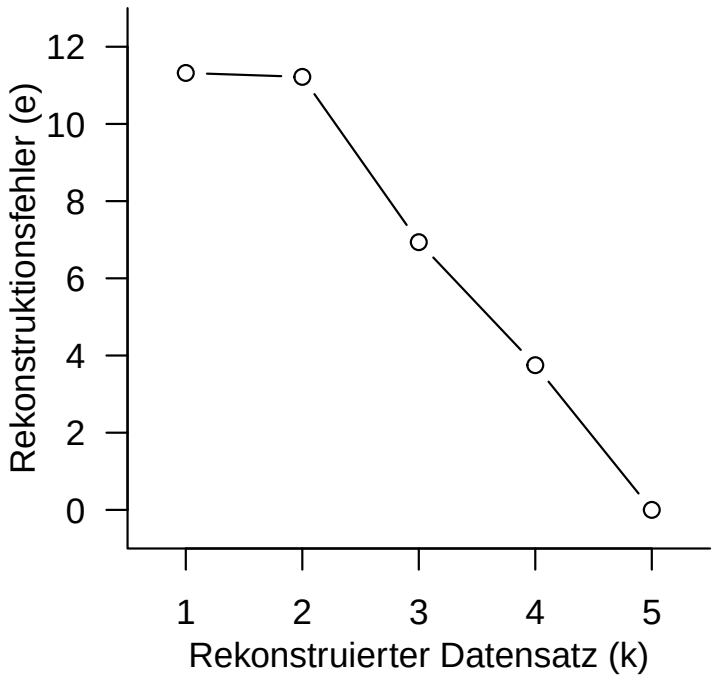
Bemerkungen

- $Y_k = Q_k \tilde{Y}_k$ 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





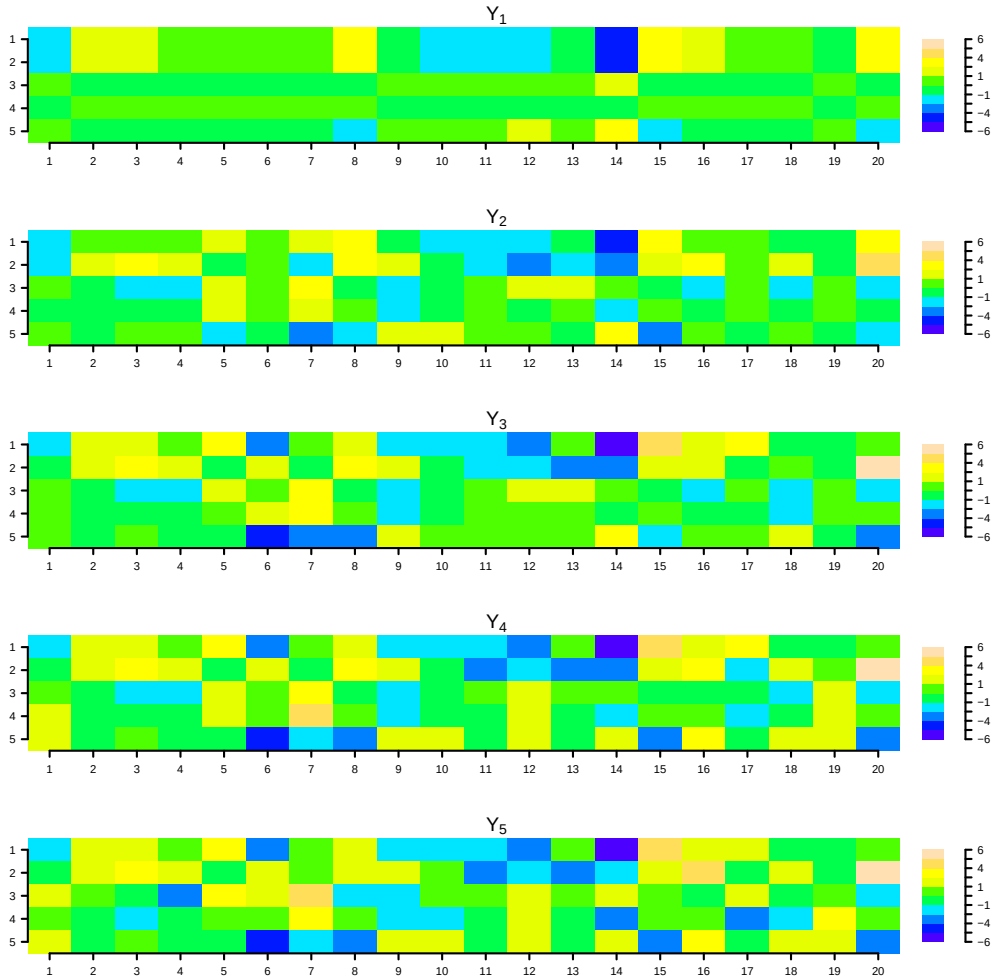
Datensatzrekonstruktion und -rekonstruktionsfehler eines simulierten Datensatzes

Scree (engl.) Schutthalde

Wikipedia

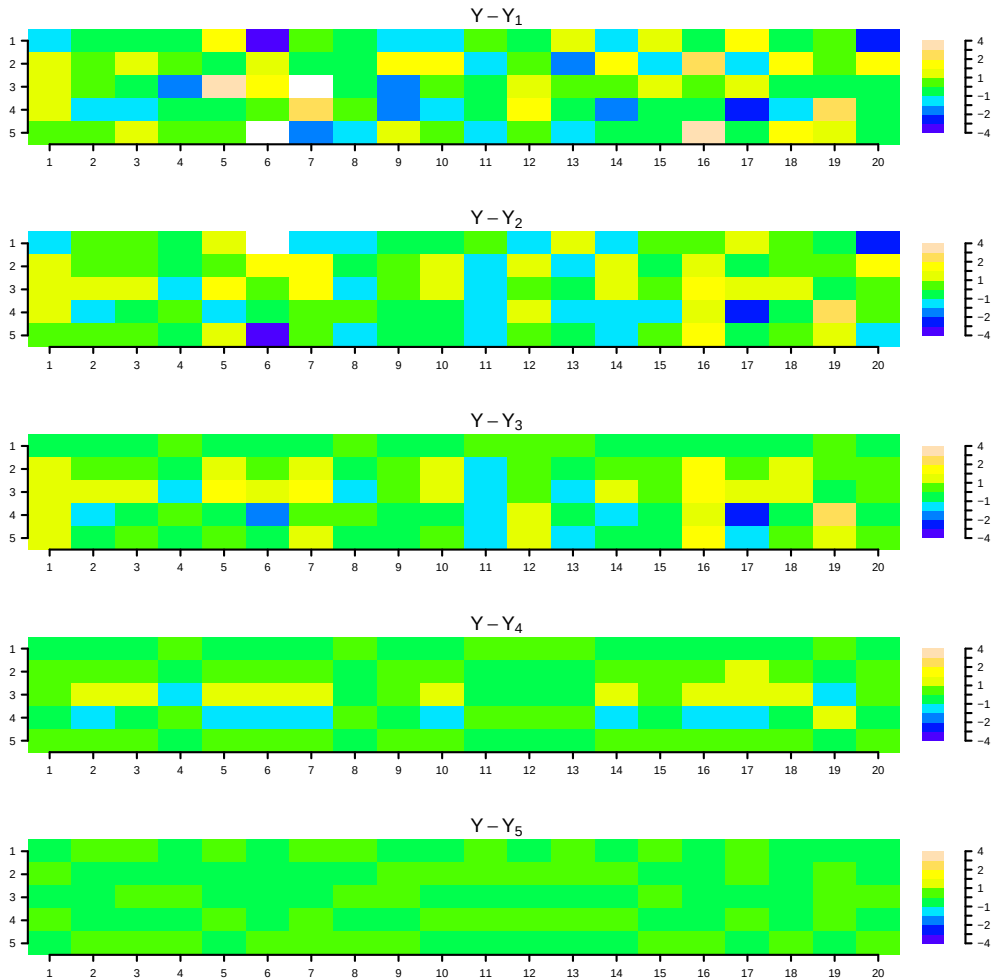
Datensatzrekonstruktion und -rekonstruktionsfehler eines simulierten Datensatzes

Rekonstruierte Datensätze



Datensatzrekonstruktion und -rekonstruktionsfehler eines simulierten Datensatzes

Originaldatensatz minus rekonstruierter Datensatz



Vektorkoordinatentransformation

Definition und Theorem

Datenkompression

Selbstkontrollfragen

Vektorkoordinatentransformation

Definition und Theorem

Datenkompression

Selbstkontrollfragen

Selbstkontrollfragen

1. Definieren Sie den Begriff Orthogonalprojektion.
2. Geben Sie das Theorem zu Vektorkoordinaten bezüglich einer Orthogonalbasis wieder.
3. Geben Sie das Vektorkoordinatentransformationstheorem wieder.
4. Erläutern Sie das Vektorkoordinatentransformationstheorem.
5. Geben Sie die Definition einer Hauptkomponentenanalyse wieder.
6. Geben Sie das Theorem zur Hauptkomponentenanalyse wieder.
7. Geben Sie die Definition der Hauptkomponentenanalyse eines Datensatzes wieder.
8. Geben Sie das Theorem zur Hauptkomponentenanalyse eines Datensatzes wieder.
9. Erläutern Sie das Prinzip der Datenkompression durch Hauptkomponentenanalyse
10. Definieren Sie den Begriff des PCA-dimensionreduzierten Datensatzes.
11. Definieren Sie den Begriff des PCA-rekonstruierten Datensatzes.
12. Definieren Sie den Begriff des PCA-Rekonstruktionsfehlers.