Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(1) Matrizen

Motivation

Matrizen sind die Worte der Sprache der multivariaten Datenanalyse.

Vektoren sind spezielle Matrizen.

Matrizen können als Tabellen der Datenrepräsentation dienen.

Matrizen können lineare Abbildungen repräsentieren.

Matrizen können Vektorräume repräsentieren.

Ein sicherer Umgang mit Matrizen ist für
das Verständnis multivariater Verfahren unverzichtbar.

Definition

Operationen

Determinanten

Spur

Spezielle Matrizen

Selbstkontrollfragen

Definition

Operationen

Determinanten

Spur

Spezielle Matrizen

Selbstkontrollfragen

Definition (Matrix)

Eine Matrix ist eine rechteckige Anordnung von Zahlen, die wie folgt bezeichnet wird

$$A := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} := (a_{ij})_{1 \leq i \leq n,\, 1 \leq j \leq m}. \tag{1}$$

Bemerkungen

- Matrizen bestehen aus *Zeilen (rows)* und *Spalten (columns)*.
- Die Matrixeinträge a_{ij} werden mit einem *Zeilenindex* i und einem *Spaltenindex* j indiziert.
- Zum Beispiel gilt für $A := \begin{pmatrix} 2 & 7 & 5 & 2 \\ 8 & 2 & 5 & 6 \\ 6 & 4 & 0 & 9 \\ 9 & 2 & 1 & 2 \end{pmatrix}$, dass $a_{32} = 4$.

Definition

Bemerkungen (fortgeführt)

- Die *Größe* oder *Dimension* einer Matrix ergibt sich aus der Anzahl ihrer Zeilen $n \in \mathbb{N}$ und Spalten $m \in \mathbb{N}$.
- Matrizen mit $n = m$ heißen *quadratische Matrizen*.
- In der Folge benötigen wir nur Matrizen mit reellen Einträgen, also $a_{ij} \in \mathbb{R}\ \forall i = 1, ..., n, j = 1, ..., m$.
- Wir nennen die Matrizen mit reellen Einträge *reelle Matrizen*.
- Die Menge der reellen Matrizen mit n Zeilen und m Spalten bezeichnen wir mit $\mathbb{R}^{n \times m}$
- Aus dem Ausdruck $A \in \mathbb{R}^{2 \times 3}$ lesen wir ab, dass A eine reelle Matrix mit zwei Zeilen und drei Spalten ist.
- Wir identifizieren die Menge $\mathbb{R}^{1 \times 1}$ mit der Menge $\mathbb{R}$.
- Wir identifizieren die Menge $\mathbb{R}^{n \times 1}$ mit der Menge $\mathbb{R}^n$.
- Reelle Matrizen mit einer Spalte und n Zeilen sind also dasselbe wie n-dimensionale reelle Vektoren.

Definition

Operationen

Determinanten

Spur

Spezielle Matrizen

Selbstkontrollfragen

Matrixoperationen

Man kann mit Matrizen rechnen.

In der Folge betrachten wir folgende grundlegende Matrixoperationen

- Addition und Subtraktion von Matrizen gleicher Größe (Matrixaddition und Matrixsubtraktion)
- Multiplikation einer Matrix mit einem Skalar (Skalarmultiplikation)
- Vertauschen der Zeilen- und Spaltenanordnung (Matrixtransposition)
- Multiplikation einer Matrix mit einer passenden zweiten Matrix (Matrixmultiplikation)
- "Teilen" durch eine Matrix (Matrixinversion)

Definition (Matrixaddition)

Es seien $A, B \in \mathbb{R}^{n \times m}$. Dann ist die *Addition* von A und B definiert als die Abbildung

$$+ : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \to \mathbb{R}^{n \times m},\ (A, B) \mapsto +(A, B) := A + B \tag{2}$$

mit

$$\begin{aligned} A + B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix} \\ &:= \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2m} + b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \cdots & a_{nm} + b_{nm} \end{pmatrix}. \end{aligned} \tag{3}$$

Bemerkungen

- Nur Matrizen identischer Größe können miteinander addiert werden.
- Die Addition zweier gleich großer Matrizen ist elementweise definiert.

Definition (Matrixsubtraktion)

Es seien $A, B \in \mathbb{R}^{n \times m}$. Dann ist die *Subtraktion* von A und B definiert als die Abbildung

$$- : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \to \mathbb{R}^{n \times m},\ (A, B) \mapsto -(A, B) := A - B \tag{4}$$

mit

$$\begin{aligned} A - B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} - \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix} \\ &:= \begin{pmatrix} a_{11} - b_{11} & a_{12} - b_{12} & \cdots & a_{1m} - b_{1m} \\ a_{21} - b_{21} & a_{22} - b_{22} & \cdots & a_{2m} - b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} - b_{n1} & a_{n2} - b_{n2} & \cdots & a_{nm} - b_{nm} \end{pmatrix}. \end{aligned} \tag{5}$$

Bemerkungen

- Nur Matrizen identischer Größe können voneinander subtrahiert werden.
- Die Subktration zweier gleich großer Matrizen ist elementweise definiert.

Beispiel

Es seien $A, B \in \mathbb{R}^{2\times 3}$ definiert als

$$A := \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \text{ und } B := \begin{pmatrix} 4 & 1 & 0 \\ -4 & 2 & 0 \end{pmatrix}. \tag{6}$$

Da A und B gleich groß sind, können wir sie addieren

$$\begin{aligned} C = A + B &= \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} + \begin{pmatrix} 4 & 1 & 0 \\ -4 & 2 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2+4 & -3+1 & 0+0 \\ 1-4 & 6+2 & 5+0 \end{pmatrix} \\ &= \begin{pmatrix} 6 & -2 & 0 \\ -3 & 8 & 5 \end{pmatrix} \end{aligned} \tag{7}$$

und voneinander subtrahieren

$$\begin{aligned} D = A - B &= \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} - \begin{pmatrix} 4 & 1 & 0 \\ -4 & 2 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 2-4 & -3-1 & 0-0 \\ 1+4 & 6-2 & 5-0 \end{pmatrix} \\ &= \begin{pmatrix} -2 & -4 & 0 \\ 5 & 4 & 5 \end{pmatrix}. \end{aligned} \tag{8}$$

Definition (Skalarmultiplikation)

Es sei $c \in \mathbb{R}$ ein Skalar und $A \in \mathbb{R}^{n \times m}$. Dann ist die *Skalarmultiplikation* von c und A definiert als die Abbildung

$$\cdot : \mathbb{R} \times \mathbb{R}^{n \times m} \to \mathbb{R}^{n \times m},\ (c, A) \mapsto \cdot(c, A) := cA \tag{9}$$

mit

$$cA = c \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} := \begin{pmatrix} ca_{11} & ca_{12} & \cdots & ca_{1m} \\ ca_{21} & ca_{22} & \cdots & ca_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{n1} & ca_{n2} & \cdots & ca_{nm} \end{pmatrix}. \tag{10}$$

Bemerkungen

- Die Skalarmultiplikation ist elementweise definiert.

Beispiel

Es seiein $c := -3$ und $A \in \mathbb{R}^{4\times 3}$ definiert als

$$A := \begin{pmatrix} 3 & 1 & 1 \\ 5 & 2 & 5 \\ 2 & 7 & 1 \\ 3 & 4 & 2 \end{pmatrix}. \tag{11}$$

Dann ergibt sich

$$B := cA = -3 \begin{pmatrix} 3 & 1 & 1 \\ 5 & 2 & 5 \\ 2 & 7 & 1 \\ 3 & 4 & 2 \end{pmatrix} = \begin{pmatrix} -3\cdot 3 & -3\cdot 1 & -3\cdot 1 \\ -3\cdot 5 & -3\cdot 2 & -3\cdot 5 \\ -3\cdot 2 & -3\cdot 7 & -3\cdot 1 \\ -3\cdot 3 & -3\cdot 4 & -3\cdot 2 \end{pmatrix} = \begin{pmatrix} -9 & -3 & -3 \\ -15 & -6 & -15 \\ -6 & -21 & -3 \\ -9 & -12 & -6 \end{pmatrix}. \tag{12}$$

Theorem (Vektorraum $\mathbb{R}^{n\times m}$)

Das Tripel $(\mathbb{R}^{n\times m}, +, \cdot)$ mit der oben definierten Matrixaddition und Skalarmultiplikation ist ein Vektorraum. Insbesondere gelten also für $A, B, C \in \mathbb{R}^{n\times m}$ und $r, s, t \in \mathbb{R}$ folgende Rechenregeln:

(1) Kommutativität der Addition	$A + B = B + A$
(2) Assoziativität der Addition	$(A + B) + C = A + (B + C)$
(3) Existenz eines neutralen Elements der Addition	$\exists 0 \in \mathbb{R}^{n\times m}$ mit $A + 0 = 0 + A = A$.
(4) Existenz inverser Elemente der Addition	$\forall A \exists -A$ mit $A + (-A) = 0$.
(5) Existenz eines neutralen Elements der Skalarmultiplikation	$\exists 1 \in \mathbb{R}$ mit $1 \cdot A = A$.
(6) Assoziativität der Skalarmultiplikation	$r \cdot (s \cdot t) = (r \cdot s) \cdot t$.
(7) Distributivität hinsichtlich der Matrixaddition	$r \cdot (A + B) = r \cdot A + r \cdot B$.
(8) Distributivität hinsichtlich der Skalaraddition	$(r + s) \cdot A = r \cdot A + s \cdot A$.

Bemerkungen

- Wir verzichten auf einen Beweis.
- Der Beweis ergibt sich mit dem elementweisen Charakter von $+, -, \cdot$ und den Rechenregeln in $(\mathbb{R}, +, \cdot)$.
- Das neutrale Element der Addition heißt *Nullmatrix*; wir schreiben $0_{nm} := (0)_{1\leq i\leq n, 1\leq j\leq m}$ mit $0 \in \mathbb{R}$.
- Die inversen Elemente der Addition sind durch $-A := (-a_{ij})_{1\leq i\leq n, 1\leq j\leq m}$ gegeben.
- Das neutrale Element der Skalarmultiplikation ist $1 \in \mathbb{R}$.

Definition (Matrixtransposition)

Es sei $A \in \mathbb{R}^{n \times m}$. Dann ist die *Transposition* von A definiert als die Abbildung

$$.^T : \mathbb{R}^{n \times m} \to \mathbb{R}^{m \times n},\ A \mapsto .^T(A) := A^T \tag{13}$$

mit

$$A^T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}^T := \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{mn} \end{pmatrix} \tag{14}$$

Bemerkungen

- Die Matrixtransposition "vertauscht" Zeilen und Spalten.
- Für $A \in \mathbb{R}^{n \times m}$ gilt immer $A^T \in \mathbb{R}^{m \times n}$.
- Für $A \in \mathbb{R}^{1 \times 1}$ gilt immer $A^T = A$.
- Es gilt $\left(A^T\right)^T = A$.
- Es gilt $(a_{ii})_{1 \leq i \leq \min(n,m)} = (a_{ii})^T_{1 \leq i \leq \min(n,m)}$
- Matrixelemente auf der Hauptdiagonalen einer Matrix bleiben bei Transposition also unberührt.

Beispiel

Es sei $A \in \mathbb{R}^{2\times 3}$ definiert durch

$$A := \begin{pmatrix} 2 & 3 & 0 \\ 1 & 6 & 5 \end{pmatrix}, \tag{15}$$

Dann gilt $A^T \in \mathbb{R}^{3\times 2}$ und speziell

$$A^T := \begin{pmatrix} 2 & 1 \\ 3 & 6 \\ 0 & 5 \end{pmatrix}. \tag{16}$$

Weiterhin gilt offenbar $\min(m, n) = 2$ und folglich

$$(a_{11}) = (a_{11})^T \text{ und } (a_{22}) = (a_{22})^T. \tag{17}$$

Definition (Matrixmultiplikation)

Es seien $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{m \times k}$. Dann ist die *Matrixmultiplikation* von A und B definiert als die Abbildung

$$\cdot : \mathbb{R}^{n \times m} \times \mathbb{R}^{m \times k} \to \mathbb{R}^{n \times k},\ (A, B) \mapsto \cdot(A, B) := AB \tag{18}$$

mit

$$\begin{aligned} AB &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ b_{21} & b_{22} & \cdots & b_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mk} \end{pmatrix} \\ &:= \begin{pmatrix} \sum_{i=1}^{m} a_{1i} b_{i1} & \sum_{i=1}^{m} a_{1i} b_{i2} & \cdots & \sum_{i=1}^{m} a_{1i} b_{ik} \\ \sum_{i=1}^{m} a_{2i} b_{i1} & \sum_{i=1}^{m} a_{2i} b_{i2} & \cdots & \sum_{i=1}^{m} a_{2i} b_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{m} a_{ni} b_{i1} & \sum_{i=1}^{m} a_{ni} b_{i2} & \cdots & \sum_{i=1}^{m} a_{ni} b_{ik} \end{pmatrix} \\ &= \left(\sum_{i=1}^{m} a_{ji} b_{il} \right)_{1 \le j \le n, 1 \le l \le k} \end{aligned} \tag{19}$$

Operationen

Bemerkungen

- Das Matrixprodukt AB ist nur dann definiert, wenn A genau so viele Spalten hat wie B Zeilen.
- Informell gilt für die beteiligten Matrixgrößen immer $(n \times m)(m \times k) = (n \times k)$.
- In AB ist $(AB)_{ij}$ die Summe der multiplizierten $i$ten Zeilen von A und $j$ten Spalten von B.
- Zum Berechnen von $(AB)_{ij}$ für $1 \leq i \leq n, 1 \leq j \leq k$ geht man also wie folgt vor:
  1. Man legt in Gedanken die Tranposition der $i$ten Zeile von A über die $j$te Spalte von B.
  2. Weil A genau m Spalten hat und B genau m Zeilen hat, gibt es zu jedem Element der Zeile aus A ein korrespondierendes Element in der Spalte von B.
  3. Man multipliziert die korrespondierenden Elemente miteinander.
  4. Die Summe dieser Produkte ist dann der Eintrag mit Index ij in AB.
- Die Multiplikation von Matrizen ist im Allgemeinen nicht kommutativ (also meist $AB \neq BA$).

Beispiel

$A \in \mathbb{R}^{2\times 3}$ und $B \in \mathbb{R}^{3\times 2}$ seien definiert als

$$A := \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \text{ und } B := \begin{pmatrix} 4 & 2 \\ -1 & 0 \\ 1 & 3 \end{pmatrix}. \tag{20}$$

Wir wollen $C := AB$ und $D := BA$ berechnen.

Mit $n = 2, m = 3$ und $k = 2$ wissen wir schon, dass $C \in \mathbb{R}^{2\times 2}$ und $D \in \mathbb{R}^{3\times 3}$, weil

$$(2 \times 3)(3 \times 2) = (2 \times 2) \tag{21}$$

und

$$(3 \times 2)(2 \times 3) = (3 \times 3) \tag{22}$$

Es gilt hier also sicher $AB \neq BA$.

Beispiel (fortgeführt)

Es ergibt sich zum einen

$$
\begin{aligned}
C &= AB \\
&= \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \begin{pmatrix} 4 & 2 \\ -1 & 0 \\ 1 & 3 \end{pmatrix} \\
&= \begin{pmatrix} 2 \cdot 4 + (-3) \cdot (-1) + 0 \cdot 1 & 2 \cdot 2 + (-3) \cdot 0 + 0 \cdot 3 \\ 1 \cdot 4 + 6 \cdot (-1) + 5 \cdot 1 & 1 \cdot 2 + 6 \cdot 0 + 5 \cdot 3 \end{pmatrix} \\
&= \begin{pmatrix} 8 + 3 + 0 & 4 + 0 + 0 \\ 4 - 6 + 5 & 2 + 0 + 15 \end{pmatrix} \\
&= \begin{pmatrix} 11 & 4 \\ 3 & 17 \end{pmatrix}.
\end{aligned}
\tag{23}
$$

Beispiel (fortgeführt)

Es ergibt sich zum anderen

$$
\begin{aligned}
D &= BA \\
&= \begin{pmatrix} 4 & 2 \\ -1 & 0 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 2 & -3 & 0 \\ 1 & 6 & 5 \end{pmatrix} \\
&= \begin{pmatrix} 4 \cdot 2 + 2 \cdot 1 & 4 \cdot (-3) + 2 \cdot 6 & 4 \cdot 0 + 2 \cdot 5 \\ (-1) \cdot 2 + 0 \cdot 1 & (-1) \cdot (-3) + 0 \cdot 6 & (-1) \cdot 0 + 0 \cdot 5 \\ 1 \cdot 2 + 3 \cdot 1 & 1 \cdot (-3) + 3 \cdot 6 & 1 \cdot 0 + 3 \cdot 5 \end{pmatrix} \\
&= \begin{pmatrix} 8 + 2 & -12 + 12 & 0 + 5 \\ -2 + 0 & 3 + 0 & 0 + 0 \\ 2 + 3 & -3 + 18 & 0 + 15 \end{pmatrix} \\
&= \begin{pmatrix} 10 & 0 & 10 \\ -2 & 3 & 0 \\ 5 & 15 & 15 \end{pmatrix}
\end{aligned}
\tag{24}
$$

Theorem (Matrixmultiplikation und Skalarprodukt)

Es seien $x, y \in \mathbb{R}^n$. Dann gilt

$$\langle x, y \rangle = x^T y. \tag{25}$$

Weiterhin seien für $A \in \mathbb{R}^{n \times m}$ für $i = 1, \ldots, n$

$$\bar{a}_i := (a_{ji})_{1 \leq j \leq m} \in \mathbb{R}^m \tag{26}$$

die Spalten von A^T und für $B \in \mathbb{R}^{m \times k}$ für $i = 1, \ldots, k$

$$\bar{b}_j := (b_{ij})_{1 \leq j \leq m} \in \mathbb{R}^m \tag{27}$$

die Spalten von B, also

$$A^T = \begin{pmatrix} \bar{a}_1 & \bar{a}_2 & \cdots & \bar{a}_n \end{pmatrix} \in \mathbb{R}^{m \times n} \text{ und } B = \begin{pmatrix} \bar{b}_1 & \bar{b}_2 & \cdots & \bar{b}_k \end{pmatrix} \in \mathbb{R}^{m \times k}. \tag{28}$$

Dann gilt

$$AB = \left(\langle \bar{a}_i, \bar{b}_j \rangle \right)_{1 \leq i \leq n, 1 \leq j \leq k} \tag{29}$$

Bemerkungen

- Der Eintrag $(AB)_{ij}$ enstpricht dem Skalarprodukt von $i$ter Spalte von A^T und $j$ter Spalte von B.
- Die erste Aussage folgt mit der Identifikation von $\mathbb{R}^n = \mathbb{R}^{n \times 1}$
- Wir verzichten auf einen ausführlichen Beweis.

Theorem (Matrixmultiplikation und Transposition)

Es seien $A \in \mathbb{R}^{m \times n}$ und $B \in \mathbb{R}^{n \times k}$. Dann gilt

$$(AB)^T = B^T A^T. \tag{30}$$

Beweis

$$\begin{aligned}
(AB)^T &= \left(\left(\sum_{i=1}^{m} a_{ji} b_{il} \right)_{1 \leq j \leq n, 1 \leq l \leq k} \right)^T \\
&= \left(\sum_{i=1}^{m} a_{ij} b_{li} \right)_{1 \leq i \leq k, 1 \leq j \leq n} \\
&= \left(\sum_{i=1}^{m} b_{li} a_{ij} \right)_{1 \leq j \leq k, 1 \leq l \leq n} \\
&= B^T A^T
\end{aligned} \tag{31}$$

□

Motivation für Begriff der Inversen einer quadratischen Matrix

- Es seien $A \in \mathbb{R}^{n \times n}, x \in \mathbb{R}^n$ und $b \in \mathbb{R}^n$, A und b seien als bekannt vorausgesetzt, x sei unbekannt.
- Zum Beispiel sei $A := \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ und $b := \begin{pmatrix} 5 \\ 11 \end{pmatrix}$
- In diesem Fall gilt $Ax = b \Leftrightarrow \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 11 \end{pmatrix} \Leftrightarrow \begin{matrix} 1x_1 + 2x_2 & = 5 \\ 3x_1 + 4x_2 & = 11 \end{matrix}$
- Wir haben also ein *lineares Gleichungssystem (LGS)* mit zwei Gleichungen und zwei Unbekannten.
- Wir stellen uns vor, dass wissen möchten, für welche(s) x das LGS erfüllt ist.
- Wären $A = a \in \mathbb{R}$, $x \in \mathbb{R}$ und $b \in \mathbb{R}$, also $ax = b$ gegeben so würden mit dem *multiplikativem Inversen* von a multiplizieren, also dem Wert, der mit a multipliziert 1 ergibt und durch $a^{-1} = \frac{1}{a}$ gegeben ist.
- Dann würde nämlich gelten $ax = b \Leftrightarrow a^{-1}ax = a^{-1}b \Leftrightarrow 1 \cdot x = a^{-1}b \Leftrightarrow x = \frac{b}{a}$
- Konkret etwa $2x = 6 \Leftrightarrow 2^{-1}2x = 2^{-1}6 \Leftrightarrow \frac{1}{2}2x = \frac{1}{2}6 \Leftrightarrow x = 3$.
- Analog möchte mit dem *multiplikativen Inversen* A^{-1} von A multiplizieren können, sodass "$A^{-1}A = 1$".
- Dann hätte man nämlich $Ax = b \Leftrightarrow A^{-1}Ax = A^{-1}b \Leftrightarrow x = A^{-1}b$
- Die Idee des multiplikativen Inversen wird im folgenden als *Inverse eine quadratischen Matrix* formalisiert.

Definition (Einheitsmatrix)

Die Matrix

$$I_n := (a_{ij})_{1 \leq i \leq n, 1 \leq j \leq n} \in \mathbb{R}^{n \times n} := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{32}$$

mit $a_{ij} = 1$ für $i = j$ und $a_{ij} = 0$ für $i \neq j$ heißt *n-dimensionale Einheitsmatrix*.

- I_n wird in R mit dem Befehl `diag(n)` erzeugt.

Theorem (Neutrales Element der Matrixmultiplikation)

I_n ist das neutrale Element der Matrixmultiplikation, d.h. es gilt für $A \in \mathbb{R}^{n \times m}$, dass

$$I_n A = A \text{ und } A I_m = A. \tag{33}$$

Beweis

Es sei $B = (b_{ij}) = I_n A \in \mathbb{R}^{n \times m}$. Dann gilt für alle $1 \leq i \leq n$ und alle $1 \leq j \leq n$

$$d_{ij} = 0 \cdot a_{1j} + 0 \cdot a_{2j} + \cdots + 0 \cdot a_{i-1,j} + 1 \cdot a_{ij} + + \cdots + 0 \cdot a_{i+1,j} + 0 \cdot a_{nj} = a_{ij} \tag{34}$$

und analog für $A I_m$. □

Definition (Invertierbare Matrix und inverse Matrix)

Eine quadratische Matrix $A \in \mathbb{R}^{n \times n}$ heißt *invertierbar*, wenn es eine quadratische Matrix $A^{-1} \in \mathbb{R}^{n \times n}$ gibt, so dass

$$A^{-1}A = AA^{-1} = I_n \tag{35}$$

ist. Die Matrix A^{-1} heißt die *inverse Matrix von* A.

Bemerkungen

- Invertierbarkeit und inverse Matrizen beziehen sich nur auf quadratische Matrizen.
- Inverse Matrizen heißen auch einfach *Inverse*.
- Quadratische Matrizen können, müssen aber nicht invertierbar sein.
- Nicht invertierbare Matrizen nennt man *singuläre* Matrizen
- Für $A = a \in \mathbb{R}^{1 \times 1}$ gilt $A^{-1} = \frac{1}{a}$.
- Die Definition sagt nur aus, was eine inverse Matrix ist, nicht wie man sie berechnet.

Beispiel für eine invertierbare Matrix

Die Matrix $A = \begin{pmatrix} 2.0 & 1.0 \\ 3.0 & 4.0 \end{pmatrix}$ ist invertierbar mit inverser Matrix $A^{-1} = \begin{pmatrix} 0.8 & -0.2 \\ -0.6 & 0.4 \end{pmatrix}$, denn

$$\begin{pmatrix} 2.0 & 1.0 \\ 3.0 & 4.0 \end{pmatrix} \begin{pmatrix} 0.8 & -0.2 \\ -0.6 & 0.4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0.8 & -0.2 \\ -0.6 & 0.4 \end{pmatrix} \begin{pmatrix} 2.0 & 1.0 \\ 3.0 & 4.0 \end{pmatrix}, \tag{36}$$

wovon man sich durch Nachrechnen überzeugt.

Beispiel für eine nicht-invertierbare Matrix

Die Matrix $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ ist nicht invertierbar, denn wäre B invertierbar, dann gäbe es $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ mit

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{37}$$

Das würde aber bedeuten, dass $0 = 1$ in $\mathbb{R}$ und das ist ein Widerspruch. Also kann B nicht invertierbar sein.

Berechnen inverser Matrizen

- 2×2 bis etwa 5×5 Matrizen kann man prinzipiell per Hand invertieren.
- Dazu lernt man im BSc Mathematik verschiedene Verfahren.
- Wir verzichten auf eine Einführung in die Matrizeninvertierung per Hand.
- Ein kurzes (30 min) Erklärvideo findet sich hier.
- In der Anwendung werden Matrizen standardmäßig numerisch invertiert.
- Matrixinversion ist ein weites Feld in der numerischen Mathematik.
- Es gibt sehr viele Algorithmen zur Invertierung invertierbarer Matrizen.
- In **R** berechnet man inverse Matrizen für gewöhnlich mit `solve`.

Definition

Operationen

Determinanten

Spur

Spezielle Matrizen

Selbstkontrollfragen

Definition (Determinante)

Für $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$ mit $n > 1$ sei $A_{ij} \in \mathbb{R}^{n-1 \times n-1}$ die Matrix, die aus A durch Entfernen der $i$ten Zeile und der $j$ten Spalte entsteht. Dann heißt die Zahl

$$|A| := a_{11} \qquad \text{für } n = 1 \tag{38}$$

$$|A| := \sum_{j=1}^{n} a_{1j}(-1)^{1+j}|A_{1j}| \text{ für } n > 1 \tag{39}$$

die *Determinante von* A.

Bemerkungen

- Für

$$A := \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \tag{40}$$

  ergeben sich zum Beispiel

$$A_{11} = \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix}, A_{12} = \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix}, A_{21} = \begin{pmatrix} 2 & 3 \\ 8 & 9 \end{pmatrix}, A_{22} = \begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix} \tag{41}$$

- Determinanten sind nichtlineare Abbildungen der Form $|\cdot| : \mathbb{R}^{n \times n} \to \mathbb{R}, A \mapsto |A|$

Theorem (Determinanten von 2×2 und 3×3 Matrizen)

(1) Es sei $A = (a_{ij})_{1 \leq i,j \leq 2} \in \mathbb{R}^{2 \times 2}$. Dann gilt

$$|A| = a_{11}a_{22} - a_{12}a_{21}. \tag{42}$$

(2) Es sei $A = (a_{ij})_{1 \leq i,j \leq 3} \in \mathbb{R}^{3 \times 3}$. Dann gilt

$$|A| = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}. \tag{43}$$

Bemerkungen

- Für 2×2 und 3×3 Matrizen (und nur für diese) gilt die *Sarrusche Merkregel*

  "Summe der Produkte auf den Diagonalen minus Summe der Produkte auf den Gegendiagonalen"

- Bei 3×3 Matrizen bezieht sich die Merkregel auf das Schema

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & | & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & | & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & | & a_{31} & a_{32} \end{pmatrix} \tag{44}$$

Beweis

Für $A \in \mathbb{R}^{2\times 2}$ gilt nach Definition

$$\begin{aligned} |A| &= \sum_{j=1}^{n} a_{1j}(-1)^{1+j}|A_{1j}| \\ &= a_{11}(-1)^{1+1}|A_{11}| + a_{12}(-1)^{1+2}|A_{12}| \\ &= a_{11}|(a_{22})| - a_{12}|(a_{21})| \\ &= a_{11}a_{22} - a_{12}a_{21} \end{aligned} \tag{45}$$

Für $A \in \mathbb{R}^{3\times 3}$ gilt nach Definition und mit der Formel für Determinanten von 2×2 Matrizen

$$\begin{aligned} |A| &= \sum_{j=1}^{n} a_{1j}(-1)^{1+j}|A_{1j}| \\ &= a_{11}(-1)^{1+1}|A_{1j}| + a_{12}(-1)^{1+2}|A_{12}| + a_{13}(-1)^{1+3}|A_{13}| \\ &= a_{11}|A_{11}| - a_{12}|A_{12}| + a_{13}|A_{13}| \\ &= a_{11}\left|\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}\right| - a_{12}\left|\begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix}\right| + a_{13}\left|\begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}\right| \\ &= a_{11}\left(a_{22}a_{33} - a_{23}a_{32}\right) - a_{12}\left(a_{21}a_{33} - a_{23}a_{31}\right) + a_{13}\left(a_{21}a_{32} - a_{22}a_{31}\right) \\ &= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} \\ &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}. \end{aligned} \tag{46}$$

Beispiel 1

Es seien

$$A := \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \text{ und } B := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{47}$$

Dann ergeben sich

$$|A| = 2 \cdot 4 - 1 \cdot 3 = 8 - 3 = 5. \tag{48}$$

und

$$|B| = 1 \cdot 0 - 0 \cdot 0 = 0 - 0 = 0. \tag{49}$$

Beispiel 2

Es sei

$$C := \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{50}$$

Dann ergibt sich

$$|C| = 2 \cdot 1 \cdot 3 + 0 \cdot 0 \cdot 0 + 0 \cdot 0 \cdot 0 - 0 \cdot 0 \cdot 3 - 0 \cdot 0 \cdot 0 - 0 \cdot 1 \cdot 0 = 2 \cdot 1 \cdot 3 = 6 \tag{51}$$

Theorem (Rechenregeln für Determinanten)

(Determinantenmultiplikationssatz.) Für $A, B \in \mathbb{R}^{n \times n}$ gilt

$$|AB| = |A||B|. \tag{52}$$

(Transposition.) Für $A \in \mathbb{R}^{n \times n}$ gilt

$$|A| = \left|A^T\right|. \tag{53}$$

(Inversion.) Für eine invertierbare Matrix $A \in \mathbb{R}^{n \times n}$ gilt

$$\left|A^{-1}\right| = \frac{1}{|A|} \tag{54}$$

(Dreiecksmatrizen.) Für Matrizen $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$ mit $a_{ij} = 0$ für $i > j$ oder $a_{ij} = 0$ für $j > i$ gilt

$$|A| = \prod_{i=1}^{n} a_{ii} \tag{55}$$

Bemerkungen

- Wir verzichten auf einen Beweis.
- Bei Dreiecksmatrizen sind alle Elemente unterhalb ($i > j$) oder oberhalb ($j > i$) der Diagonalen 0
- Bei I_n sind alle nicht-diagonalen Elemente 0 und alle diagonalen Elemente 1, also folgt $|I_n| = 1$.

Theorem (Invertierbarkeit und Determinante)

$A \in \mathbb{R}^{n \times n}$ ist dann und nur dann invertierbar, wenn gilt, dass $|A| \neq 0$. Es gilt also

$$A \text{ ist invertierbar} \Leftrightarrow |A| \neq 0 \text{ und } A \text{ ist nicht invertierbar} \Leftrightarrow |A| = 0. \tag{56}$$

Beweisandeutung

Wir zeigen lediglich, dass aus der Invertierbarkeit von A folgt, dass $|A|$ nicht null sein kann. Nehmen wir also an, dass A invertierbar ist. Dann gibt es eine Matrix B mit $AB = I_n$ und mit dem Determinantenmultiplikationssatz folgt

$$|AB| = |A||B| = |I_n| = 1. \tag{57}$$

Also kann $|A| = 0$ nicht gelten, denn sonst wäre $0 = 1$.

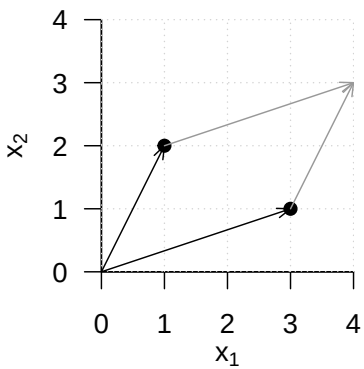
□

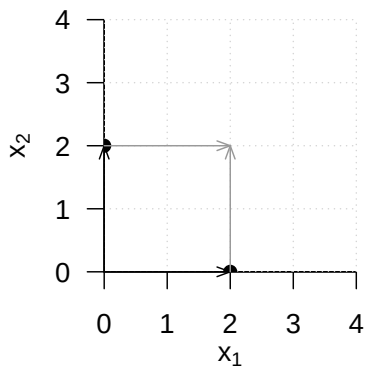
Visuelle Intuition

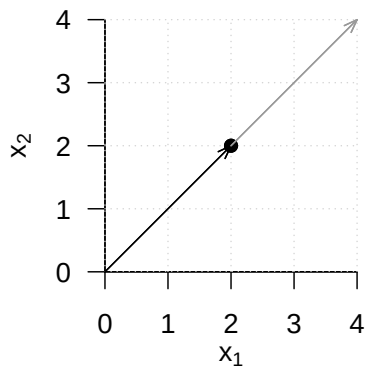
$a_1, ..., a_n \in \mathbb{R}^n$ seien die Spalten von $A \in \mathbb{R}^{n \times n}$.

$\Rightarrow |A|$ enstpricht dem signierten Volumen des von $a_1, ..., a_n \in \mathbb{R}^n$ aufgespannten Parallelotops.

$$A_1 = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad A_3 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$







$$|A_1| = 3 \cdot 2 - 1 \cdot 1 = 5 \qquad |A_2| = 2 \cdot 2 - 0 \cdot 0 = 4 \qquad |A_3| = 2 \cdot 2 - 2 \cdot 2 = 0$$

Definition

Operationen

Determinanten

Spur

Spezielle Matrizen

Selbstkontrollfragen

Definition (Spur)

csei eine quadratische Matrix. Dann ist die *Spur von A* definiert als die Zahl

$$\text{tr}(A) = \sum_{i=1}^{n} a_{ii}. \tag{58}$$

Bemerkungen

- Die Spur einer quadratischen Matrix ist die Summe ihrer Diagonalelemente.
- tr ist die Abkürzung für *trace*.

Beispiel

Es sei

$$A := \begin{pmatrix} 3 & 2 & 1 \\ 8 & 1 & 7 \\ 4 & 5 & 2 \end{pmatrix} \tag{59}$$

Dann gilt

$$\text{tr}(A) = 3 + 1 + 2 = 6. \tag{60}$$

Theorem (Eigenschaften der Spur)

(1) (Invarianz bei Transposition) Es sei $A \in \mathbb{R}^{n \times n}$. Dann gilt

$$\mathsf{tr}\left(A^T\right) = \mathsf{tr}(A) \tag{61}$$

(2) (Linearität) Es seien $A, B \in \mathbb{R}^{n \times n}$ $c, d \in \mathbb{R}$. Dann gilt

$$\mathsf{tr}(cA + dB) = c\,\mathsf{tr}(A) + d\,\mathsf{tr}(B) \tag{62}$$

(3) (Reihenfolgeninvarianz) Es sei $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{m \times n}$. Dann gilt

$$\mathsf{tr}(AB) = \mathsf{tr}(BA). \tag{63}$$

(4) (Zyklische Permutationsinvarianz) Es seien $A, B, C \in \mathbb{R}^{n \times n}$. Dann gilt

$$\mathsf{tr}(ABC) = \mathsf{tr}(CAB) = \mathsf{tr}(BCA). \tag{64}$$

Bemerkung

- Eigenschaft (4) kann auf beliebig viele und multiplikationskonforme Matrizen erweitert werden.

Spur

Beweis

(1) Da die Diagonalelemente von A^T und A identisch sind, sind auch ihre Summen identisch.

(2) Es gilt

$$\text{tr}(cA + dB) = \sum_{i=1}^{n} ca_{ii} + db_{ii} = c\sum_{i=1}^{n} a_{ii} + d\sum_{i=1}^{n} b_{ii} = c\,\text{tr}(A) + d\,\text{tr}(B). \tag{65}$$

(3) Wir halten zunächst fest, dass nach Definition der Matrixmultiplikation, die Diagonalelemente von $AB \in \mathbb{R}^{n\times n}$ gegeben sind durch

$$(AB)_{ii} = \sum_{j=1}^{m} a_{ij}b_{ji} \text{ für } i = 1, ..., n. \tag{66}$$

und die Diagonalelemente von $BA \in \mathbb{R}^{m\times m}$ gegeben sind durch

$$(BA)_{jj} = \sum_{i=1}^{m} b_{ji}a_{ij} \text{ für } j = 1, ..., m. \tag{67}$$

Dann aber gilt mit den Eigenschaften von Summen und der Kommutativität der Skalarmultiplikation

$$\text{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji}a_{ij} = \text{tr}(BA). \tag{68}$$

(4) Mit der Assoziativität der Matrixmultiplikation und Eigenschaft (4) gelten

$$\text{tr}(ABC) = \text{tr}((AB)C) = \text{tr}(C(AB)) = \text{tr}(CAB) = \text{tr}((CA)B) = \text{tr}(B(CA)) = \text{tr}(BCA). \tag{69}$$

Definition

Operationen

Spur

Determinanten

Spezielle Matrizen

Selbstkontrollfragen

Definition (Einheitsmatrizen und Einheitsvektoren)

- Wir bezeichnen die *Einheitsmatrix* mit

$$I_n := (i_{jk})_{1 \leq i \leq n, 1 \leq j \leq n} \in \mathbb{R}^{n \times n} \text{ mit } i_{jk} = 1 \text{ für } j = k \text{ und } i_{jk} = 0 \text{ für } j \neq k \tag{70}$$

- Wir bezeichnen die *Einheitsvektoren* $e_i, i = 1, \ldots, n$ mit

$$e_i := (e_{i_j})_{1 \leq j \leq n} \in \mathbb{R}^n \text{ mit } e_{i_j} = 1 \text{ für } i = j \text{ und } e_{i_j} = 0 \text{ für } i \neq j \tag{71}$$

Bemerkungen

- I_n besteht nur aus Nullen und Diagonalelementen gleich Eins.
- $e_i, i = 1, \ldots., n$ besteht nur aus Nullen und einer Eins in der $i$ten Komponente.
- Es gilt

$$I_n = \begin{pmatrix} e_1 & \cdots & e_n \end{pmatrix} \tag{72}$$

- Es gelten weiterhin zum Beispiel für $1 \leq i, j \leq n$

$$e_i^T e_j = 0 \text{ für } i \neq j, e_i^T e_i = 1 \text{ und } e_i^T v = v^T e_i = v_i \text{ für } v \in \mathbb{R}^n. \tag{73}$$

Definition (Nullmatrizen, Einsmatrizen)

- Wir bezeichnen *Nullmatrizen* mit

$$0_{nm} := (0)_{1 \leq i \leq n, 1 \leq j \leq m} \in \mathbb{R}^{n \times m} \text{ und } 0_n := (0)_{1 \leq i \leq n} \in \mathbb{R}^n \tag{74}$$

- Wir bezeichnen den *Einsmatrizen* mit

$$1_{nm} := (1)_{1 \leq i \leq n, 1 \leq j \leq m} \in \mathbb{R}^{n \times m} \text{ und } 1_n := (1)_{1 \leq i \leq n} \in \mathbb{R}^n \tag{75}$$

Bemerkungen

- 0_{nm} und 0_n bestehen nur aus Nullen.
- 1_{nm} und 1_n bestehen nur aus Nullen.
- Es gelten zum Beispiel

$$0_n 0_n^T = 0_{nn} \text{ und } 1_n 1_n^T = 1_{nn}. \tag{76}$$

Definition (Diagonalmatrix)

Eine Matrix $D \in \mathbb{R}^{n \times m}$ heißt *Diagonalmatrix*, wenn $d_{ij} = 0$ für $1 \leq i \leq n, 1 \leq j \leq m$ mit $i \neq j$.

Bemerkungen

- Eine Diagonalmatrix $D \in \mathbb{R}^{n \times n}$ mit Diagonalelementen $d_1, ..., d_n$ schreibt man auch als

$$D = \text{diag}(d_1, ..., d_n). \tag{77}$$

- Diagonalmatrizen haben viele "gute" Eigenschaften.
- Zum Beispiel überzeugt man sich leicht davon, dass Multiplikation einer Matrix A von links mit einer Diagonalmatrix D der Multiplikation der Zeilen der Matrix A mit den entsprechenden Diagonaleinträgen von D entspricht. Die entsprechende Multiplikation von rechts entspricht der Multiplikation der Spalten von A mit entsprechenden Diagonaleinträgen von D.
- Eine weitere wichtige Eigenschaft ist

$$D := \text{diag}(d_1, ..., d_n) \Rightarrow |D| = \prod_{i=1}^{n} d_i \tag{78}$$

- Für Beweise dieser Eigenschaften wird auf die weiterführende Literatur, z.B. Searle (1982) verwiesen.

Definition (Symmetrische Matrix)

Eine Matrix $S \in \mathbb{R}^{n \times n}$ heißt *symmetrisch*, wenn gilt dass $S^T = S$.

Bemerkungen

- Symmetrische Matrizen sind spezielle quadratische Matrizen.
- Symmetrische Matrizen haben viele "gute" Eigenschaften.
- Beispielweise gilt für die Summe zweier symmetrischer Matrizen, dass auch diese wieder symmetrisch ist

$$A = A^T \text{ und } B = B^T \Rightarrow A + B = (A + B)^T \tag{79}$$

  und das die Inverse einer symmetrischen Matrix, sofern sie existiert, auch symmetrisch ist,

$$S^T = S \Rightarrow \left(S^{-1}\right)^T = S^{-1}. \tag{80}$$

- Für Beweise dieser Eigenschaften wird auf die weiterführende Literatur, z.B. Searle (1982), verwiesen.

Definition (Positiv-definite Matrizen und positiv-semidefinite Matrizen)

Eine quadratische Matrix $C \in \mathbb{R}^{n \times n}$ heißt *positiv-definit*, wenn C symmetrisch ist und gilt, dass

$$x^T C x > 0 \text{ für alle } x \in \mathbb{R}^n \text{ mit } x \neq 0_n. \tag{81}$$

Wenn $C \in \mathbb{R}^{n \times n}$ positiv-definit ist, so schreiben wir $C \in \mathbb{R}^{n \times n}$ pd.

Eine quadratische Matrix $C \in \mathbb{R}^{n \times n}$ heißt *positiv-semidefinit*, wenn C symmetrisch ist und gilt, dass

$$x^T C x \geq 0 \text{ für alle } x \in \mathbb{R}^n \text{ mit } x \neq 0_n. \tag{82}$$

Wenn $C \in \mathbb{R}^{n \times n}$ positiv-definit ist, so schreiben wir $C \in \mathbb{R}^{n \times n}$ psd.

Bemerkungen

- Positive-definite und positiv-semidefinite Matrizen sind spezielle symmetrische Matrizen.
- Kovarianzmatrixparameter von multivariaten Normalverteilungen sind positiv definite Matrizen grundlegend.
- Kovarianzmatrizen von Zufallsvektoren sind positiv-semidefinite Matrizen.
- Positiv-definite Matrizen haben viele “gute” Eigenschaften.
- Beispielsweise existiert die Inverse C^{-1} einer positiv-definiten Matrix und ist selbst positiv-definit, für einen Beweis dieser Eigenschaft verweisen wir auf die weiterführende Literatur, z.B. Searle (1982).

Definition (Orthogonale Matrix)

Eine Matrix $Q \in \mathbb{R}^{n \times n}$ heißt *orthogonal*, wenn

$$Q^T Q = I_n. \tag{83}$$

Bemerkung

- Die Spalten einer orthogonalen Matrix sind also paarweise orthogonal, es gilt für

$$Q = \begin{pmatrix} q_1 & \cdots & q_n \end{pmatrix} \text{ mit } q_i \in \mathbb{R}^n \text{ für } 1 \leq i \leq n, \tag{84}$$

dass

$$q_i^T q_j = 0 \text{ für } i \neq j \text{ und } q_i^T q_j = 1 \text{ für } i = j \text{ mit } 1 \leq i, j \leq n. \tag{85}$$

Theorem (Eigenschaften orthogonaler Matrizen)

$Q \in \mathbb{R}^{n \times n}$ sei eine orthogonale Matrix. Dann gelten:

(1) (Inverse) Die Inverse von Q ist Q^T, es gilt

$$Q^{-1} = Q^T. \tag{86}$$

(2) (Transposition) Die Zeilen von Q sind orthonormal, es gilt

$$QQ^T = I_n. \tag{87}$$

(3) (Determinante) Es gilt

$$|Q| = 1 \text{ oder } |Q| = -1. \tag{88}$$

Beweis

(1) Unter der Annahme, dass Q^{-1} existiert, gilt

$$Q^T Q = I_n \Leftrightarrow Q^T Q Q^{-1} = I_n Q^{-1} \Leftrightarrow Q^{-1} = Q^T. \tag{89}$$

(2) Es gilt

$$Q^T Q = I_n \Leftrightarrow QQ^T Q = QI_n \Leftrightarrow QQ^T QQ^T = QQ^T \Leftrightarrow QQ^T = I_n. \tag{90}$$

(3) Mit dem Multiplikationssatz und Transpositionseigenschaft von Determinanten gilt

$$|Q|^2 = |Q||Q| = |Q||Q^T| = |QQ^T| = |I_n| = 1. \tag{91}$$

Aus $|Q|^2 = 1$ folgt dann aber, dass $|Q| = 1$ oder $|Q| = -1$ sein muss.

□

Definition

Operationen

Determinanten

Spezielle Matrizen

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition einer Matrix wieder.
2. Nennen Sie sechs Matrixoperationen.
3. Geben Sie die Definitionen der Matrixaddition und -subtraktion wieder.
4. Geben Sie die Definition der Skalarmultiplikation für Matrizen wieder.
5. Geben Sie die Definition der Matrixtransposition wieder.
6. Es seien
$$A := \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, B := \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \text{ und } c := 2 \tag{92}$$
Berechnen Sie
$$D := c\left(A - B^T\right) \text{ und } E := (cA)^T + B. \tag{93}$$
7. Geben Sie die Definition der Matrixmultiplikation wieder.
8. Es seien $A \in \mathbb{R}^{3\times 2}, B \in \mathbb{R}^{2\times 4}$ und $C \in \mathbb{R}^{3\times 4}$. Prüfen Sie, ob folgende Matrixprodukte definiert sind, und wenn ja, geben Sie die Größe der resultierenden Matrix an:
$$ABC, \quad ABC^T, \quad , A^TCB^T \quad , BAC. \tag{94}$$

Selbstkontrollfragen

9. Es seien

$$A := \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & 2 & 0 \end{pmatrix} B := \begin{pmatrix} 1 & 2 & 2 \\ 1 & 3 & 1 \\ 2 & 0 & 0 \end{pmatrix} \text{ und } C := \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}. \tag{95}$$

   Berechnen Sie die Matrixprodukte

$$AB, \qquad B^T A^T, \qquad \left(B^T A^T\right)^T, \qquad AC. \tag{96}$$

10. Geben Sie die Formel für die Determinante von $A := (A_{ij})_{1 \leq i,j \leq 2} \in \mathbb{R}^2$ wieder.

11. Geben Sie die Formel für die Determinante von $A := (A_{ij})_{1 \leq i,j \leq 3} \in \mathbb{R}^3$ wieder.

12. Berechnen Sie die Determinanten von

$$A := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} B := \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{pmatrix} \text{ und } C := \mathsf{diag}(1, 2, 3). \tag{97}$$

Selbstkontrollfragen

14. Geben Sie den Determinantenmultiplikationssatz wieder.

15. Geben Sie das Theorem zur Invertierbarkeit und Determinante von Matrizen wieder.

16. Geben Sie die Definition einer Diagonalmatrix wieder.

17. Geben Sie die Definition einer symmetrischen Matrix wieder.

18. Geben Sie die Definition einer positiv-definiten und einer positiv-semidefiniten Matrix wieder.

19. Geben Sie die Definition einer orthogonalen Matrix wieder.

Referenzen

Searle, Shayle. 1982. *Matrix Algebra Useful for Statistics*. Wiley-Interscience.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(2) Eigenanalyse

Eigenvektoren und Eigenwerte

Orthonormalzerlegung

Singulärwertzerlegung

Selbstkontrollfragen

Eigenvektoren und Eigenwerte

Orthonormalzerlegung

Singulärwertzerlegung

Selbstkontrollfragen

Definition (Eigenvektor, Eigenwert)

$A \in \mathbb{R}^{m \times m}$ sei eine quadratische Matrix. Dann heißt jeder Vektor $v \in \mathbb{R}^m, v \neq 0_m$, für den gilt, dass

$$Av = \lambda v \tag{1}$$

mit $\lambda \in \mathbb{R}$ ein *Eigenvektor* von A. λ heißt zugehöriger *Eigenwert* von A.

Bemerkungen

- Ein Eigenvektor v von A wird durch A mit einem Faktor λ verlängert.
- Jeder Eigenvektor hat einen zugehörigen Eigenwert.
- Die Eigenwerte verschiedener Eigenvektoren können identisch sein.

Theorem (Multiplikativität von Eigenvektoren)

$A \in \mathbb{R}^{m \times m}$ sei eine quadratische Matrix. Wenn $v \in \mathbb{R}^m$ Eigenvektor von A mit Eigenwert $\lambda \in \mathbb{R}$ ist, dann ist für $c \in \mathbb{R}$ auch $cv \in \mathbb{R}^m$ Eigenvektor von A und zwar wiederum mit Eigenwert $\lambda \in \mathbb{R}$.

Beweis

Es gilt

$$Av = \lambda v \Leftrightarrow cAv = c\lambda v \Leftrightarrow A(cv) = \lambda(cv) \tag{2}$$

Also ist cv ein Eigenvektor von A mit Eigenwert λ.

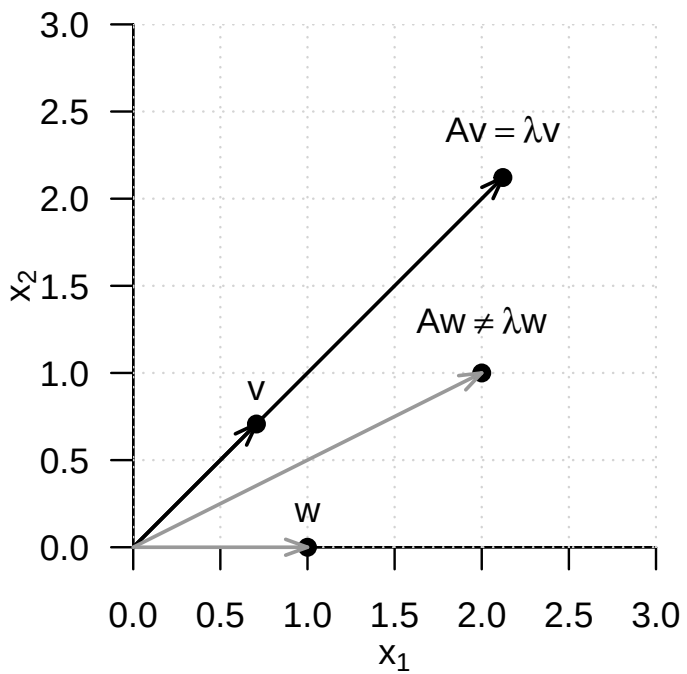
□

Konvention

Wir betrachten im Folgenden nur Eigenvektoren mit $\|v\| = 1$.

Visualisierung eines Eigenvektors

Für $A := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ ist $v := \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ Eigenvektor zum Eigenwert $\lambda = 3, w := \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ist kein Eigenvektor.



Theorem (Bestimmung von Eigenwerten und Eigenvektoren)

$A \in \mathbb{R}^{m \times m}$ sei eine quadratische Matrix. Dann ergeben sich die Eigenwerte von A als die Nullstellen des *charakteristischen Polynoms*

$$\chi_A(\lambda) := |A - \lambda I_m| \tag{3}$$

von A. Weiterhin seien $\lambda_i^*, i = 1, 2, ...$ die auf diese Weise bestimmten Eigenwerte von A. Die entsprechenden Eigenvektoren $v_i, i = 1, 2, ...$ von A können dann durch Lösen der linearen Gleichungssysteme

$$(A - \lambda_i^* I_m) v_i = 0_m \text{ für } i = 1, 2, ... \tag{4}$$

bestimmt werden.

Bemerkungen

- Für kleine Matrizen mit $m \leq 3$ können Eigenwerte und Eigenvektoren manuell bestimmt werden.
- Bei großen Matrizen werden Eigenwerte und Eigenvektor im Allgemeinen numerisch bestimmt.
- R's `eigen()`, Scipy's `linalg.eig()`, Matlab's `eig()`.

Beweis

(1) Bestimmen von Eigenwerten

Wir halten zunächst fest, dass mit der Definition von Eigenvektoren und Eigenwerten gilt, dass

$$Av = \lambda v \Leftrightarrow Av - \lambda v = 0_m \Leftrightarrow (A - \lambda I_m)v = 0_m. \tag{5}$$

Für den Eigenwert λ wird der Eigenvektor v also durch $(A - \lambda I_m)$ auf den Nullvektor 0_m abgebildet. Weil aber per Definition $v \neq 0_m$ gilt, ist die Matrix $(A - \lambda I_m)$ somit nicht invertierbar: sowohl der Nullvektor als auch v werden durch A auf 0_m abgebildet, die lineare Abbildung

$$f : \mathbb{R}^m \to \mathbb{R}^m, x \mapsto (A - \lambda I_m)x \tag{6}$$

ist also nicht bijektiv, und $(A - \lambda I_m)^{-1}$ kann nicht existieren. Die Tatsache, dass $(A - \lambda I_m)$ nicht invertierbar ist, ist aber äquivalent dazu, dass die Determinante von $(A - \lambda I_m)$ Null ist. Also ist

$$\chi_A(\lambda) = |A - \lambda I_m| = 0 \tag{7}$$

notwendige und hinreichende Bedingung dafür, dass λ ein Eigenwert von A ist.

(2) Bestimmen von Eigenvektoren

Es sei λ_i^* ein Eigenwert von A. Dann gilt mit den obigen Überlegungen, dass Auflösen von

$$(A - \lambda_i^* I_m)v_i^* = 0_m \tag{8}$$

nach v_i^* einen Eigenvektor zum Eigenwert λ^* ergibt.

□

Beispiel

Es sei

$$A := \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \tag{9}$$

Wir wollen die Eigenwerte und Eigenvektoren von A bestimmen.

(1) Berechnen von Eigenwerten

Die Eigenwerte von A sind die Nullstellen des charakteristischen Polynoms von A.

Das charakteristische Polynom von A ergibt als

$$\chi_A(\lambda) = \left| \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = \left| \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} \right| = (2-\lambda)^2 - 1. \tag{10}$$

Nullsetzen und Auflösen nach λ ergibt mit der pq-Formel

$$(2-\lambda)^2 - 1 = 0 \Rightarrow \lambda_1 = 3, \lambda_2 = 1. \tag{11}$$

Die Eigenwerte von A sind also $\lambda_1 = 3$ und $\lambda_2 = 1$.

Beispiel (fortgeführt)

(2) Berechnen von Eigenvektoren

Die Eigenvektoren zu den Eigenwerten $\lambda_1 = 3$ und $\lambda_2 = 1$ ergeben sich durch Lösen der linearen Gleichungssysteme

$$(A - \lambda_i I_2)v_i = 0_2 \text{ für } i = 1, 2. \tag{12}$$

Für $\lambda_1 = 3$ ergibt sich

$$(A - 3I_2)v_1 = 0_2 \Leftrightarrow \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} v_{1_1} \\ v_{1_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ ist eine Lösung.} \tag{13}$$

Fpr $\lambda_2 = 1$ ergibt sich

$$(A - 1I_2)v_2 = 0_2 \Leftrightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_{2_1} \\ v_{2_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \text{ ist eine Lösung.} \tag{14}$$

Weiterhin gilt $v_1^T v_2 = 0$ und $||v_1|| = ||v_2|| = 1$.

Theorem (Eigenwerte positiv semidefiniter und positiv definiter Matrizen)

(1) $C \in \mathbb{R}^{m \times m}$ sei eine positiv semidefinite Matrix. Dann sind alle Eigenwerte von C nicht-negativ.

(2) $C \in \mathbb{R}^{m \times m}$ sei eine positiv definite Matrix. Dann sind alle Eigenwerte von C positiv.

Beweis

(1) Mit der Definition von Eigenvektor und Eigenwert einer quadratischen Matrix gilt für jeden Eigenwert λ und zugehörigen Eigenvektor $x \in \mathbb{R}^m, x \neq 0_m$

$$Cx = \lambda x \Leftrightarrow x^T Cx = x^T(\lambda x) = \lambda x^T x. \tag{15}$$

Mit der positiven Semidefinitheit von C und $x^T x \geq 0$ für alle $x \in \mathbb{R}^m$ mit $x \neq 0_m$ gilt dann aber

$$x^T Cx \geq 0 \Leftrightarrow \lambda x^T x \geq 0 \Rightarrow \lambda \geq 0. \tag{16}$$

Also ist jeder Eigenwert von C nichtnegativ.

(2) Der Beweis erfolgt analog zu (1) unter Ersetzung von $\geq$ durch $>$.

□

Bemerkungen

- Die Eigenwertnichtnegativität wird manchmal auch zur Definition der positiven Semidefinitheit genutzt.
- Die Eigenwertpositivität wird manchmal auch zur Definition der positiven Definitheit genutzt.

Eigenvektoren und Eigenwerte

Orthonormalzerlegung

Singulärwertzerlegung

Selbstkontrollfragen

Theorem (Eigenwerte und Eigenvektoren symmetrischer Matrizen)

$S \in \mathbb{R}^{m \times m}$ sei eine symmetrische Matrix. Dann gelten

(1) Die Eigenwerte von S sind reell.

(2) Die Eigenvektoren zu je zwei verschiedenen Eigenwerten von S sind orthogonal.

Bemerkung

- In nachfolgendem Beweis setzen wir die Tatsache dass eine symmetrische m reelle Eigenwerte hat als gegeben voraus und zeigen lediglich, dass die Eigenvektoren zu je zwei verschiedenen Eigenwerten einer symmetrischen Matrix orthogonal sind. Ein vollständiger Beweis des Theorems findet sich in Strang (2009), Kapitel 6.4.
- Da wir als Eigenvektoren nur Eigenvektoren der Länge 1 betrachten, sind die hier angesprochenen orthogonalen Eigenvektoren insbesondere auch orthonormal.

Orthonormalzerlegung

Beweis

Ohne Beschränkung der Allgemeinheit seien λ_i und λ_j mit $1 \leq i, j \leq m$ und $\lambda_i \neq \lambda_j$ zwei verschiedenen Eigenwerte von S mit zugehörigen Eigenvektoren q_i und q_j, respektive. Dann ergibt sich wie unten gezeigt, dass

$$\lambda_i q_i^T q_j = \lambda_j q_i^T q_j. \tag{17}$$

Mit $q_i \neq 0_m, q_j \neq 0_m$ und $\lambda_i \neq \lambda_j$ folgt damit $q_i^T q_j = 0$, weil weil es keine andere Zahl c als die Null gibt, für die bei $a, b \in \mathbb{R}$ und $a \neq b$ gilt, dass

$$ac = bc. \tag{18}$$

Um

$$\lambda_i q_i^T q_j = \lambda_j q_i^T q_j. \tag{19}$$

zu zeigen, halten wir zunächst fest, dass

$$Sq_i = \lambda_i q_i \Leftrightarrow (Sq_i)^T = (\lambda_i q_i)^T \Leftrightarrow q_i^T S^T = q_i^T \lambda_i^T \Leftrightarrow q_i^T S = q_i^T \lambda_i \Leftrightarrow q_i^T S q_j = \lambda_i q_i^T q_j \tag{20}$$

und

$$Sq_j = \lambda_j q_j \Leftrightarrow q_i^T S q_j = \lambda_j q_i^T q_j \tag{21}$$

gelten. Sowohl $\lambda_i q_i^T q_j$ als auch $\lambda_j q_i^T q_j$ sind also mit $q_i^T S q_j$ und damit auch miteinander identisch.

□

Theorem (Orthonormale Zerlegung einer symmetrischen Matrix)

$S \in \mathbb{R}^{m \times m}$ sei eine symmetrische Matrix mit m verschiedenen Eigenwerten. Dann kannn S geschrieben werden als

$$S = Q \Lambda Q^T, \tag{22}$$

wobei $Q \in \mathbb{R}^{m \times m}$ eine orthogonale Matrix ist und $\Lambda \in \mathbb{R}^{m \times m}$ eine Diagonalmatrix ist.

Beweis

Es seien $\lambda_1 > \lambda_2 > \cdots > \lambda_m$ die der Größe nach geordneten Eigenwerte von S und $q_1, q_2, \ldots, q_m$ seien die jeweils zugehörigen orthonormalen Eigenvektoren. Mit

$$Q := \begin{pmatrix} q_1 & q_2 & \cdots & q_m \end{pmatrix} \in \mathbb{R}^{m \times m} \text{ und } \Lambda := \text{diag}\begin{pmatrix} \lambda_1, \lambda_2, \ldots, \lambda_m \end{pmatrix} \in \mathbb{R}^{m \times m}, \tag{23}$$

folgt dann mit den Definitionen von Eigenwerten und Eigenvektoren zunächst, dass

$$S q_i = \lambda_i q_i \text{ für } i = 1, \ldots, m \Leftrightarrow SQ = Q\Lambda. \tag{24}$$

Rechtseitige Multiplikation mit Q^T ergibt dann mit $QQ^T = I_m$, dass

$$SQQ^T = Q\Lambda Q^T \Leftrightarrow SI_m = Q\Lambda Q^T \Leftrightarrow S = Q\Lambda Q^T. \tag{25}$$

□

Bemerkungen

- $S = Q\Lambda Q^T$ heißt auch *Diagonalisierung* von S.
- Man wählt man als Diagonalelemente von Λ die der Größe nach geordneten Eigenwerte von S.
- Man wählt man als Spalten von Q die zugehörigen orthonormalen Eigenvektoren von S.

Beispiel (fortgeführt)

Für die symmetrische Matrix

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \tag{26}$$

mit den oben bestimmten Eigenwerten $\lambda_1 = 3, \lambda_2 = 1$ und zugehörigen orthonormalen Eigenvektoren

$$v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{27}$$

seien

$$Q := \begin{pmatrix} v_1 & v_2 \end{pmatrix} \text{ und } \Lambda = \mathsf{diag}(\lambda_1, \lambda_2). \tag{28}$$

Dann ergibt sich

$$\begin{aligned} Q\Lambda Q^T &= \begin{pmatrix} v_1 & v_2 \end{pmatrix} \mathsf{diag}(\lambda_1, \lambda_2) \begin{pmatrix} v_1 & v_2 \end{pmatrix}^T \\ &= \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \right) \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \right) \\ &= \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 3 & -1 \\ 3 & 1 \end{pmatrix} \right) \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \right) \\ &= \frac{1}{2} \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \\ &= A. \end{aligned}$$

Theorem (Spur und Determinante einer symmetrischen Matrix)

$S \in \mathbb{R}^{m \times m}$ sei eine symmetrische Matrix mit verschiedenden Eigenwerten $\lambda_1, ..., \lambda_m$. Dann gelten

$$|S| = \prod_{i=1}^{m} \lambda_i \text{ und } \mathsf{tr}(S) = \sum_{i=1}^{m} \lambda_i \tag{29}$$

Beweis

Mit dem Theorem zur Zerlegung einer symmetrischen Matrix mit verschiedenen Eigenwerten gilt, dass

$$|S| = |Q \Lambda Q^T| \tag{30}$$

wobei $Q \in \mathbb{R}^{m \times m}$ eine orthogonale Matrix und $\Lambda \in \mathbb{R}^{m \times m}$ die Diagonalmatrix der m verschiedenen Eigenwerte von S ist. Mit dem Determinantenmultiplikationssatz, der Determinanteneigenschaft von orthogonalen Matrizen und der Tatsache, dass die Determinante einer Diagonalmatrix dem Produkt ihrer Diagonalelemente entspricht, gilt dann weiterhin

$$|S| = |Q \Lambda Q^T| = |Q||\Lambda||Q^T| = |\Lambda| = \prod_{i=1}^{m} \lambda_i . \tag{31}$$

Wiederrum mit dem Theorem zur Zerlegung einer symmetrischen Matrix mit verschiedenen Eigenwerten gilt, dass

$$\mathsf{tr}(S) = \mathsf{tr}\left(Q \Lambda Q^T\right). \tag{32}$$

Mit der zyklischen Permutationsinvarianz der Spur, der Inversionseigenschaft orthogonaler Matrizen und der Definition der Spur gilt dann weiterhin

$$\mathsf{tr}(S) = \mathsf{tr}\left(Q \Lambda Q^T\right) = \mathsf{tr}\left(Q^T Q \Lambda\right) = \mathsf{tr}(\Lambda) = \sum_{i=1}^{m} \lambda_i . \tag{33}$$

□

Eigenvektoren und Eigenwerte

Orthonormalzerlegung

Singulärwertzerlegung

Selbstkontrollfragen

Definition (Singulärwertzerlegung)

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

$$Y = USV^T, \tag{34}$$

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 Y. Die Diagonalelemente von S heißen die *Singulärwerte* von Y.

Bemerkungen

- Für eine ausführliche Diskussion der Singulärwertzerlegung siehe z.B. Strang (2009), Kapitel 7.
- Singulärwertzerlegungen können in R mit `svd()` berechnet werden.

Theorem (Singulärwertzerlegung und Eigenanalyse)

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

$$Y = USV^T \tag{35}$$

sei ihre Singulärwertzerlegung. Dann gilt:

- Die Spalten von U sind die Eigenvektoren von YY^T,
- die Spalten von V sind die Eigenvektoren von Y^TY und
- die entsprechenden Singulärwerte sind die Quadratwurzeln der zugehörigen Eigenwerte.

Bemerkung

- Singulärwertzerlegung und Eigenanalyse sind eng verwandt.

Beweis

Wir halten zunächst fest, dass mit

$$\left(YY^T\right)^T = YY^T \text{ und } \left(Y^TY\right)^T = Y^TY \tag{36}$$

YY^T und Y^TY symmetrische Matrizen sind und somit Orthornomalzerlegungen haben. Wir halten weiterhin fest, dass mit $V^TV = I_n$, $U^TU = I_m$ gilt, dass

$$YY^T = USV^T\left(USV^T\right)^T = USV^TVS^TU^T = USS^TU^T =: U\Lambda_U U^T, \tag{37}$$

wobei wir $\Lambda_U := SS^T$ definiert haben und

$$Y^TY = \left(USV^T\right)^T USV^T = VS^TU^TUSV^T =: V\Lambda_V V^T, \tag{38}$$

wobei wir $\Lambda_V := S^TS$ definiert haben. Weil das Produkt von Diagonalmatrizen wieder eine Diagonalmatrix ist, sind Λ_U und Λ_V Diagonalmatrizen und per Definition sind U und V orthogonale Matrizen. Wir haben also YY^T und Y^TY also in Form der Orthonormalzerlegungen

$$YY^T = U\Lambda_U U^T \text{ und } Y^TY = V\Lambda_V V^T \tag{39}$$

geschrieben, wobei für die Diagonalelemente von Λ_U und Λ_V gilt, dass sie die quadrierten Werte der Diagonalemente von S sind. Damit folgen die Aussagen des Theorems direkt. □

Eigenvektoren und Eigenwerte

Orthonormalzerlegung

Singulärwertzerlegung

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition eines Eigenvektors und eines Eigenwertes einer quadratischen Matrix wieder.
2. Geben Sie das Theorem zur Bestimmung von Eigenwerten und Eigenvektoren wieder.
3. Geben Sie das Theorem zu den Eigenwerten und Eigenvektoren symmetrischer Matrizen wieder.
4. Geben Sie das Theorem zur orthonormalen Zerlegung einer symmetrischen Matrix wieder.
5. Geben Sie die Definition einer Singulärwertzerlegung wieder.
6. Geben Sie das Theorem zum Zusammenhang von Singulärwertzerlegung und Eigenanalyse wieder.

Referenzen

Strang, Gilbert. 2009. *Introduction to Linear Algebra*. Cambridge University Press.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(3) Zufallsvektoren

Definition und multivariate Verteilungen

Marginale und bedingte Verteilungen

Normalverteilungen

Selbstkontrollfragen

Definition und multivariate Verteilungen

Marginale und bedingte Verteilungen

Normalverteilungen

Selbstkontrollfragen

Definition (Zufallsvektor)

$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum und $(\mathcal{X}, \mathcal{S})$ sei ein m-dimensionaler Messraum. Ein m-dimensionaler *Zufallsvektor* ist definiert als eine Abbildung

$$\xi : \Omega \to \mathcal{X}, \omega \mapsto \xi(\omega) := \begin{pmatrix} \xi_1(\omega) \\ \vdots \\ \xi_m(\omega) \end{pmatrix} \tag{1}$$

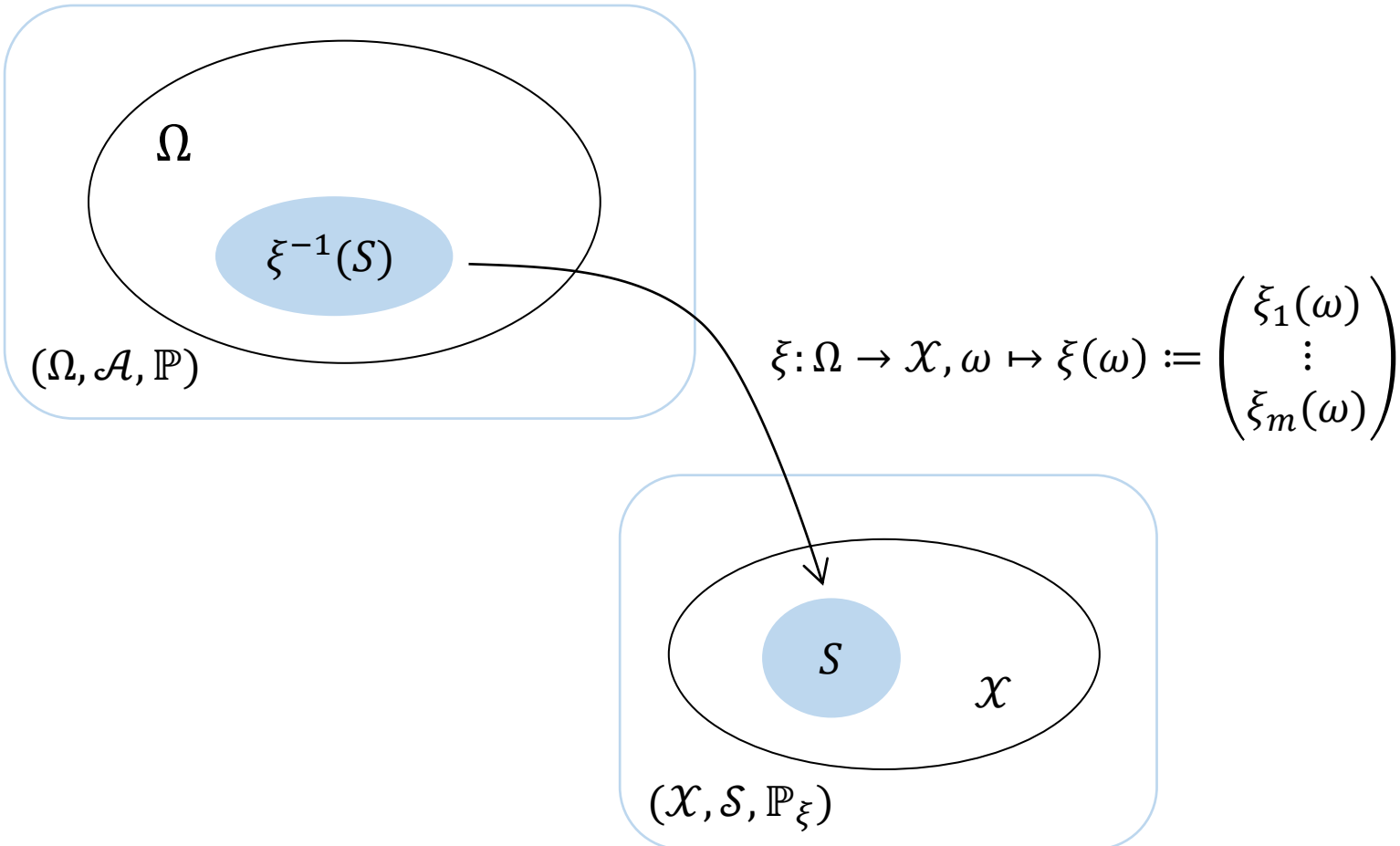
mit der *Messbarkeitseigenschaft*

$$\{\omega \in \Omega | \xi(\omega) \in S\} \in \mathcal{A} \text{ für alle } S \in \mathcal{S}. \tag{2}$$

Bemerkungen

- ξ ist hier eine univariate, vektorwertige Abbildung.
- Das Standardbeispiel für $(\mathcal{X}, \mathcal{S})$ ist $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$.
- Wir verzichten auf eine explizite Einführung m-dimensionaler σ-Algebren wie $\mathcal{B}(\mathbb{R}^m)$.
- Ohne Beweis halten wir fest, dass ξ messbar ist, wenn die Funktionen $\xi_1, \ldots, \xi_m$ messbar sind.
- Die Komponentenfunktionen eines Zufallsvektors sind Zufallsvariablen.
- Ein m-dimensionaler Zufallsvektor ist die Konkatenation von m Zufallsvariablen.
- Für $m := 1$ ist ein Zufallsvektor eine Zufallsvariable.

Definition und multivariate Verteilungen



$$\mathbb{P}\left(\xi^{-1}(S)\right) = \mathbb{P}(\{\omega \in \Omega | \xi(\omega) \in S\}) =: \mathbb{P}_\xi(S)$$

Definition (Multivariate Verteilung)

$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein m-dimensionaler Messraum und

$$\xi : \Omega \to \mathcal{X}, \omega \mapsto \xi(\omega) \tag{3}$$

sei ein Zufallsvektor. Dann heißt das Wahrscheinlichkeitsmaß $\mathbb{P}_\xi$, definiert durch

$$\mathbb{P}_\xi : \mathcal{S} \to [0,1], S \mapsto \mathbb{P}_\xi(S) := \mathbb{P}(\xi^{-1}(S)) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) \in S\}\right) \tag{4}$$

die *multivariate Verteilung des Zufallsvektor* ξ.

Bemerkungen

- Der Einfachheit halber spricht man oft auch nur von "der Verteilung des Zufallsvektors ξ".
- Die Notationskonventionen für Zufallsvariablen gelten für Zufallsvektoren analog, z.B.

$$\begin{aligned}
\mathbb{P}_\xi(\xi \in S) &:= \mathbb{P}\left(\{\xi \in S\}\right) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) \in S\}\right) \\
\mathbb{P}_\xi(\xi = x) &:= \mathbb{P}\left(\{\xi = x\}\right) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) = x\}\right) \\
\mathbb{P}_\xi(\xi \leq x) &:= \mathbb{P}\left(\{\xi \leq x\}\right) = \mathbb{P}\left(\{\omega \in \Omega | \xi(\omega) \leq x\}\right) \\
\mathbb{P}_\xi(x_1 \leq \xi \leq x_2) &:= \mathbb{P}\left(\{x_1 \leq \xi \leq x_2\}\right) = \mathbb{P}\left(\{\omega \in \Omega | x_1 \leq \xi(\omega) \leq x_2\}\right)
\end{aligned} \tag{5}$$

- Relationsoperatoren wie $\leq$ werden hier *komponentenweise* verstanden.
- Zum Beispiel heißt $x \leq y$ für $x, y \in \mathbb{R}^m$, dass $x_i \leq y_i$ für alle $i = 1, \ldots, m$.

Definition (Diskreter Zufallsvektor, Multivariate WMF)

ξ sei ein Zufallsvektor mit Ergebnisraum $\mathcal{X}$. ξ heißt *diskreter Zufallsvektor* wenn der Ergebnisraum $\mathcal{X}$ endlich oder abzählbar ist und eine Funktion

$$p_\xi : \mathcal{X} \to [0,1], x \mapsto p_\xi(x) \tag{6}$$

existiert, für die gilt

(1) $\sum_{x \in \mathcal{X}} p_\xi(x) = 1$ und

(2) $\mathbb{P}_\xi(\xi = x) = p_\xi(x)$ für alle $x \in \mathcal{X}$.

Ein entsprechende Funktion p heißt *multivariate Wahrscheinlichkeitsmassefunktion (WMF)* von ξ.

Bemerkungen

- Der Begriff der multivariaten WMF ist analog zum Begriff der WMF.
- Man spricht oft einfach von der WMF eines Zufallsvektors.
- Wie univariate WMFen sind multivariate WMFen nicht-negativ und normiert.

Beispiel (Multivariate Wahrscheinlichkeitsmassefunktion)

Wir betrachten einen zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ der Werte in $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ annimmt, wobei $\mathcal{X}_1 := \{1, 2, 3\}$ und $\mathcal{X}_2 = \{1, 2, 3, 4\}$ seien.

Dann entspricht der Ergebnisraum von ξ der in untenstehender Tabelle spezifizierten Menge an Tupeln (x_1, x_2)

(x_1, x_2)	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 1$	$(1, 1)$	$(1, 2)$	$(1, 3)$	$(1, 4)$
$x_1 = 2$	$(2, 1)$	$(2, 2)$	$(2, 3)$	$(2, 4)$
$x_1 = 3$	$(3, 1)$	$(3, 2)$	$(3, 3)$	$(3, 4)$

Beispiel (Multivariate Wahrscheinlichkeitsmassefunktion)

Wir betrachten einen zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ der Werte in $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ annimmt, wobei $\mathcal{X}_1 := \{1, 2, 3\}$ und $\mathcal{X}_2 = \{1, 2, 3, 4\}$ seien.

Eine exemplarische bivariate WMF der Form

$$p_\xi : \{1, 2, 3\} \times \{1, 2, 3, 4\} \to [0, 1], (x_1, x_2) \mapsto p_\xi(x_1, x_2) \tag{7}$$

ist dann durch nachfolgende Tabelle definiert:

$p_\xi(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 1$	0.1	0.0	0.2	0.1
$x_1 = 2$	0.1	0.2	0.0	0.0
$x_1 = 3$	0.0	0.1	0.1	0.1

Man beachte, dass $\sum_{x_1=1}^{3} \sum_{x_2=1}^{4} p_\xi(x_1, x_2) = 1$.

Definition (Kontinuierlicher Zufallsvektor, Multivariate WDF)

Ein Zufallsvektor ξ heißt *kontinuierlich*, wenn $\mathbb{R}^m$ der Ergebnisraum von ξ ist und eine Funktion

$$p_\xi : \mathbb{R}^m \to \mathbb{R}_{\geq 0}, x \mapsto p_\xi(x), \tag{8}$$

existiert, für die gilt

(1) $\int_{\mathbb{R}^m} p_\xi(x)\, dx = 1$ und

(2) $\mathbb{P}_\xi(x_1 \leq \xi \leq x_2) = \int_{x_1{}_1}^{x_2{}_1} \cdots \int_{x_1{}_m}^{x_2{}_m} p_\xi(s_1, ..., s_m)\, ds_1 \cdots ds_m$.

Eine entsprechende Funktion p heißt *multivariate Wahrscheinlichkeitsdichtefunktion (WDF)* von ξ.

Bemerkungen

- Der Begriff der multivariaten WDF ist analog zum Begriff der WDF.
- Man spricht häufig auch einfach von der WDF eines Zufallsvektors
- Wie univariate WDFen sind multivariate WDFen nicht-negativ und normiert.
- Wie für kontinuierliche Zufallsvariablen gilt für kontinuierliche Zufallsvektoren

$$\mathbb{P}_\xi(\xi = x) = \mathbb{P}_\xi(x \leq \xi \leq x) = \int_{x_1}^{x_1} \cdots \int_{x_m}^{x_m} p_\xi(s_1, ..., s_m)\, ds_1 \cdots ds_m = 0 \tag{9}$$

- Mit den multivariaten Normalverteilungen diskutieren wir unten ein ausführliches Beispiel. –>

Definition (Erwartungswert eines Zufallsvektors)

ξ sei ein m-dimensionaler Zufallvektor. Der *Erwartungwert von* ξ ist definiert als der m-dimensionale Vektor

$$\mathbb{E}(\xi) := \begin{pmatrix} \mathbb{E}(\xi_1) \\ \vdots \\ \mathbb{E}(\xi_m) \end{pmatrix}. \tag{10}$$

Bemerkung

- Der Erwartungswert eines Zufallsvektors ist der Vektor der Erwartungswerte seiner Komponenten.

Definition (Erwartungswert einer Zufallsmatrix)

Ξ sei eine durch die spaltenweise Konkatenation von $\xi^1, \ldots, \xi^n$ m-dimensionalen Zufallsvektoren gebildete Zufallsmatrix. Dann ist der *Erwartungwert von* Ξ ist definiert als die $m \times n$ Matrix

$$\mathbb{E}(\Xi) := \begin{pmatrix} \mathbb{E}\left(\xi_1^1\right) & \cdots & \mathbb{E}\left(\xi_1^n\right) \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left(\xi_m^1\right) & \cdots & \mathbb{E}\left(\xi_m^n\right) \end{pmatrix}. \tag{11}$$

Bemerkung

- Der Erwartungswert einer Zufallsmatrix ist von Ξ ist die Matrix der spaltenweise konkatenierten Erwartungswerte $\mathbb{E}\left(\xi^1\right), \ldots, \mathbb{E}\left(\xi^n\right)$.

Theorem (Eigenschaften von Erwartungswerten)

(1) (Linear-affine Transformation einer Zufallsmatrix) Ξ sei ein $m \times n$-dimensionale Zufallsmatrix und es seien $A \in \mathbb{R}^{l \times m}$, $B \in \mathbb{R}^{n \times p}$ und $C \in \mathbb{R}^{l \times p}$. Dann gilt

$$\mathbb{E}(A\xi B + C) = A\mathbb{E}(\xi)B + C. \tag{12}$$

(2) (Linear-affine Transformation eines Zufallsvektors) ξ sei ein m-dimensionaler Zufallvektor und es seien $A \in \mathbb{R}^{n \times m}$ und $b \in \mathbb{R}^n$. Dann gilt

$$\mathbb{E}(A\xi + b) = A\mathbb{E}(\xi) + b. \tag{13}$$

(3) (Lineare Kombination zweier Zufallsvektoren) ξ und υ seien m-dimensionale Zufallvektoren und es seien $A, B \in \mathbb{R}^{n \times m}$. Dann gilt

$$\mathbb{E}\left(A\xi + B\upsilon\right) = A\mathbb{E}(\xi) + B\mathbb{E}(\upsilon). \tag{14}$$

Bemerkungen

- Die Aussagen sind im Wesentlichen analog zu den Eigenschaften des Erwartungswerts bei Zufallsvariablen.
- Eigenschaft (2) ist ein Spezialfall von Eigenschaft (1).

Beweis (fortgeführt)

Wir halten zunächst fest, dass für $i = 1, \ldots, l$ und $j = 1, \ldots p$ mit den Regeln von Matrixmultiplikation und Matrixaddition und den Eigenschaften des Erwartungswertes von Zufallsvariablen gilt, dass

$$\mathbb{E}(A\xi B + C)_{ij} = \mathbb{E}\left(\sum_{r=1}^{m}\sum_{s=1}^{n} a_{ir}\xi_r^s b_{sj} + c_{ij}\right) = \sum_{r=1}^{m}\sum_{s=1}^{n} a_{ir}\mathbb{E}(\xi_r^s)b_{sj} + c_{ij} \tag{15}$$

. Dann aber gilt mit der Definition des Erwartungswerts einer Zufallsmatrix, dass

$$\begin{aligned}
\mathbb{E}(A\xi B + C) &= (\mathbb{E}(A\xi B + C)_{ij})_{1\leq i\leq l, 1\leq j\leq p} \\
&= \left(\sum_{r=1}^{m}\sum_{s=1}^{n} a_{ir}\mathbb{E}(\xi_r^s)b_{sj} + c_{ij}\right)_{1\leq i\leq l, 1\leq j\leq p} \\
&= \left(\sum_{r=1}^{m}\sum_{s=1}^{n} a_{ir}\mathbb{E}(\xi_r^s)b_{sj}\right)_{1\leq i\leq l, 1\leq j\leq p} + \left(c_{ij}\right)_{1\leq i\leq l, 1\leq j\leq p} \\
&= A\mathbb{E}(\xi)B + C.
\end{aligned} \tag{16}$$

Beweis (fortgeführt)

(2) Für $i = 1, \ldots, n$ gilt mit der Definition des Erwartungswert eines Zufallsvektors und den Eigenschaften des Erwartungswert einer Zufallsvariable, dass

$$\mathbb{E}(A\xi + b)_i = \mathbb{E}\left(\sum_{j=1}^{m} a_{ij}\xi_j + b_i\right) = \sum_{j=1}^{m} a_{ij}\mathbb{E}(\xi_j) + b_i. \tag{17}$$

Also gilt

$$\mathbb{E}(A\xi + b) = \begin{pmatrix} \sum_{j=1}^{m} a_{1j}\mathbb{E}(\xi_j) + b_1 \\ \vdots \\ \sum_{j=1}^{m} a_{nj}\mathbb{E}(\xi_j) + b_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{m} a_{1j}\mathbb{E}(\xi_j) \\ \vdots \\ \sum_{j=1}^{m} a_{nj}\mathbb{E}(\xi_j) \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = A\mathbb{E}(\xi) + b. \tag{18}$$

(3) Für $i = 1, \ldots, n$ gilt mit der Definition des Erwartungswert eines Zufallsvektors und den Eigenschaften des Erwartungswert einer Zufallsvariable, dass

$$\mathbb{E}\left(A\xi + B\upsilon\right)_i = \mathbb{E}\left(\sum_{j=1}^{m} a_{ij}\xi_j + \sum_{j=1}^{m} b_{ij}\upsilon_j\right) = \sum_{j=1}^{m} a_{ij}\mathbb{E}\left(\xi_j\right) + \sum_{j=1}^{m} b_{ij}\mathbb{E}\left(\upsilon_j\right) \tag{19}$$

Also gilt

$$\mathbb{E}\left(A\xi + B\upsilon\right) = \begin{pmatrix} \sum_{j=1}^{m} a_{1j}\mathbb{E}\left(\xi_j\right) + \sum_{j=1}^{m} b_{1j}\mathbb{E}\left(\upsilon_j\right) \\ \vdots \\ \sum_{j=1}^{m} a_{nj}\mathbb{E}\left(\xi_j\right) + \sum_{j=1}^{m} b_{nj}\mathbb{E}\left(\upsilon_j\right) \end{pmatrix} = A\mathbb{E}(\xi) + B\mathbb{E}(\upsilon). \tag{20}$$

Definition (Kovarianzmatrix eines Zufallsvektors)

ξ sei ein m-dimensionaler Zufallvektor. Dann ist die *Kovarianzmatrix von* ξ definiert als die $m \times m$ Matrix

$$\mathbb{C}(\xi) := \mathbb{E}\left((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right). \tag{21}$$

Bemerkungen

- Die Kovarianzmatrix ist formal analog zur Kovarianz zweier Zufallsvariablen definiert.
- Der äußere Erwartungswert ist der Erwartungswert einer Matrix, die inneren Erwartungswerte von Vektoren.

Theorem (Eigenschaften der Kovarianzmatrix)

ξ sei ein m-dimensionaler Zufallsvektor und $\mathbb{C}(\xi)$ sei seine Kovarianzmatrix. Dann gelten

(1) (Elemente) Die Elemente von $\mathbb{C}(\xi)$ sind die Kovarianzen der Komponenten von ξ,

$$\mathbb{C}(\xi) = \left(\mathbb{C}(\xi_i, \xi_j)\right)_{1 \leq i,j \leq m}. \tag{22}$$

(2) (Kovarianzmatrixverschiebungssatz) Es gilt

$$\mathbb{C}(\xi) = \mathbb{E}\left(\xi\xi^T\right) - \mathbb{E}(\xi)\mathbb{E}(\xi)^T. \tag{23}$$

(3) (Linear-affine Transformation) Für $A \in \mathbb{R}^{n \times m}$ und $b \in \mathbb{R}^n$ gilt

$$\mathbb{C}(A\xi + b) = A\mathbb{C}(\xi)A^T. \tag{24}$$

(4) (Matrixeigenschaften) $\mathbb{C}(\xi)$ ist symmetrisch und positiv-semidefinit.

Bemerkungen

- Die Diagonalelemente von $\mathbb{C}(\xi)$ sind die Varianzen der Komponenten von ξ, da

$$\mathbb{V}(\xi_i) = \mathbb{C}(\xi_i, \xi_i) \text{ für } i = 1, \ldots m. \tag{25}$$

- Eigenschaften (2) und (3) sind im Wesentlichen analog zu den Eigenschaften der Varianz.

Beweis

(1) Es gilt

$$
\begin{aligned}
\mathbb{C}(\xi) &:= \mathbb{E}\left((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right) \\
&= \mathbb{E}\left(\left(\begin{pmatrix}\xi_1\\ \vdots \\ \xi_n\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\xi_1)\\ \vdots \\ \mathbb{E}(\xi_n)\end{pmatrix}\right)\left(\begin{pmatrix}\xi_1\\ \vdots \\ \xi_n\end{pmatrix} - \begin{pmatrix}\mathbb{E}(\xi_1)\\ \vdots \\ \mathbb{E}(\xi_n)\end{pmatrix}\right)^T\right) \\
&= \mathbb{E}\left(\begin{pmatrix}\xi_1 - \mathbb{E}(\xi_1)\\ \vdots \\ \xi_n - \mathbb{E}(\xi_n)\end{pmatrix}\begin{pmatrix}\xi_1 - \mathbb{E}(\xi_1)\\ \vdots \\ \xi_n - \mathbb{E}(\xi_n)\end{pmatrix}^T\right) \\
&= \mathbb{E}\left(\begin{pmatrix}\xi_1 - \mathbb{E}(\xi_1)\\ \vdots \\ \xi_n - \mathbb{E}(\xi_n)\end{pmatrix}\begin{pmatrix}\xi_1 - \mathbb{E}(\xi_1) & \dots & \xi_n - \mathbb{E}(\xi_n)\end{pmatrix}\right) \\
&= \mathbb{E}\begin{pmatrix}(\xi_1 - \mathbb{E}(\xi_1))(\xi_1 - \mathbb{E}(\xi_1)) & \dots & (\xi_1 - \mathbb{E}(\xi_1))(\xi_n - \mathbb{E}(\xi_n)\\ \vdots & \ddots & \vdots \\ (\xi_n - \mathbb{E}(\xi_n))(\xi_1 - \mathbb{E}(\xi_1)) & \dots & (\xi_n - \mathbb{E}(\xi_n))(\xi_n - \mathbb{E}(\xi_n))\end{pmatrix} \\
&= \left(\mathbb{E}\left((\xi_i - \mathbb{E}(\xi_i))(\xi_j - \mathbb{E}(\xi_j))\right)\right)_{1\le i,j\le n} \\
&= \left(\mathbb{C}(\xi_i, \xi_j)\right)_{1\le i,j\le n}.
\end{aligned}
\tag{26}
$$

Beweis (fortgeführt)

(2) Mit den Eigenschaften von Erwartungswerten gilt

$$\begin{aligned}\mathbb{C}(\xi) &= \mathbb{E}\left((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right) \\ &= \mathbb{E}\left(\xi\xi^T - \xi\mathbb{E}(\xi)^T - \mathbb{E}(\xi)\xi^T + \mathbb{E}(\xi)\mathbb{E}(\xi)^T\right) \\ &= \mathbb{E}\left(\xi\xi^T\right) - \mathbb{E}(\xi)\mathbb{E}(\xi)^T - \mathbb{E}(\xi)\mathbb{E}(\xi)^T + \mathbb{E}(\xi)\mathbb{E}(\xi)^T \\ &= \mathbb{E}\left(\xi\xi^T\right) - \mathbb{E}(\xi)\mathbb{E}(\xi)^T.\end{aligned} \tag{27}$$

(3) Mit den Eigenschaften von Erwartungswerten gilt

$$\begin{aligned}\mathbb{C}(A\xi + b) &= \mathbb{E}\left((A\xi + b - \mathbb{E}(A\xi + b))(A\xi + b - \mathbb{E}(A\xi + b))^T\right) \\ &= \mathbb{E}\left((A\xi + b - A\mathbb{E}(\xi) - b)(A\xi + b - A\mathbb{E}(\xi) - b)^T\right) \\ &= \mathbb{E}\left((A(\xi - \mathbb{E}(\xi)))(A(\xi - \mathbb{E}(\xi)))^T\right) \\ &= \mathbb{E}\left(A(\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T A^T\right) \\ &= A\mathbb{E}\left((\xi - \mathbb{E}(\xi))(\xi - \mathbb{E}(\xi))^T\right) A^T \\ &= A\mathbb{C}(\xi)A^T.\end{aligned} \tag{28}$$

Beweis (fortgeführt)

(4) Die Symmetrie von $\mathbb{C}(\xi)$ folgt aus der Symmetrie der Kovarianz einer Zufallsvariable mit

$$\mathbb{C}(\xi_i, \xi_j) = \mathbb{C}(\xi_j, \xi_i) \text{ für alle } i = 1, ..., m, j = 1, ..., m. \tag{29}$$

Um die positive Semidefinitheit von $\mathbb{C}(\xi)$ nachzuweisen, ist zu zeigen, dass $a^T \mathbb{C}(\xi) a \geq 0$ für alle $a \in \mathbb{R}^m$ mit $a \neq 0_m$. Sei also $a \in \mathbb{R}^m$ mit $a \neq 0$. Dann gilt mit Aussage (3) für $A := a^T \in \mathbb{R}^{1 \times m}$, dass

$$a^T \mathbb{C}(\xi) a = \mathbb{C}(a^T \xi). \tag{30}$$

Weiterhin gilt mit der Definition der Kovarianzmatrix aber, dass

$$\mathbb{C}(a^T \xi) = \mathbb{E}\left(\left(a^T \xi - \mathbb{E}(a^T \xi) \right)^2 \right) = \mathbb{V}\left(a^T \xi \right). \tag{31}$$

Da mit den Eigenschaften der Varianz die Varianz der Zufallsvariable $a^T \xi$ aber immer nichtnegativ ist, folgt

$$a^T \mathbb{C}(\xi) a = \mathbb{V}(a^T \xi) \geq 0 \tag{32}$$

und damit die positive Semidefinitheit von $\mathbb{C}(\xi)$.

Definition (Korrelationsmatrix)

ξ sei ein m-dimensionaler Zufallsvektor. Dann ist die *Korrelationsmatrix* von ξ definiert als die $m \times m$ Matrix

$$\mathbb{R}(\xi) := \left(\rho_{ij}\right)_{1 \leq i,j \leq m} = \left(\frac{\mathbb{C}(\xi_i, \xi_j)}{\sqrt{\mathbb{V}(\xi_i)}\sqrt{\mathbb{V}(\xi_j)} } \right)_{1 \leq i,j \leq m} \tag{33}$$

Bemerkungen

- Mit $\mathbb{V}(\xi_i) = \mathbb{C}(\xi_i, \xi_i), i = 1, \ldots, m$ ist die Korrelationsmatrix in der Kovarianzmatrix implizit.
- Es gelten $\rho_{ij} \in [-1, 1]$ für $1 \leq i, j \leq m$ und $\rho_{ii} = 1$ für $1 \leq i \leq m$.

Definition und multivariate Verteilungen

Marginale und bedingte Verteilungen

Normalverteilungen

Selbstkontrollfragen

Definition (Univariate Marginalverteilung)

$(\Omega, \mathcal{A}, \mathbb{P})$ sei ein Wahrscheinlichkeitsraum, $(\mathcal{X}, \mathcal{S})$ sei ein m-dimensionaler Messraum, $\xi : \Omega \to \mathcal{X}$ sei ein Zufallsvektor, $\mathbb{P}_\xi$ sei die Verteilung von ξ, $\mathcal{X}_i \subset \mathcal{X}$ sei der Ergebnisraum der $i$ten Komponente ξ_i von ξ, und $\mathcal{S}_i$ sei eine σ-Algebra auf ξ_i. Dann heißt die durch

$$\mathbb{P}_{\xi_i} : \mathcal{S}_i \to [0,1], S \mapsto \mathbb{P}_\xi \left(\mathcal{X}_1 \times \cdots \times \mathcal{X}_{i-1} \times S \times \mathcal{X}_{i+1} \times \cdots \times \mathcal{X}_m \right) \text{ für } S \in \mathcal{S}_i \tag{34}$$

definierte Verteilung die *$i$te univariate Marginalverteilung von ξ*.

Bemerkungen

- Univariate Marginalverteilungen sind die Verteilungen der Komponenten eines Zufallsvektors.
- Univariate Marginalverteilungen sind Verteilungen von Zufallsvariablen.
- Die Festlegung der multivariaten Verteilung von ξ legt auch die Verteilungen der ξ_i fest.

Theorem (Marginale Wahrscheinlichkeitsmasse- und dichtefunktionen)

(1) $\xi = (\xi_1, ..., \xi_m)^T$ sei ein m-dimensionaler diskreter Zufallsvektor mit Wahrscheinlichkeitsmassefunktion p_ξ und Komponentenergebnisräumen $\mathcal{X}_1, ..., \mathcal{X}_m$. Dann ergibt sich die Wahrscheinlichkeitsmassefunktion der $i$ten Komponente ξ_i von ξ als

$$p_{\xi_i} : \mathcal{X}_i \to [0,1], x_i \mapsto p_{\xi_i}(x_i) := \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_m} p_\xi(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_m). \tag{35}$$

(2) $\xi = (\xi_1, ..., \xi_m)^T$ sei ein m-dimensionaler kontinuierlicher Zufallsvektor mit Wahrscheinlichkeitsdichtefunktion p_ξ und Komponentenergebnisraum $\mathbb{R}$. Dann ergibt sich die Wahrscheinlichkeitsdichtefunktion der $i$ten Komponente ξ_i von ξ als

$$p_{\xi_i} : \mathbb{R} \to \mathbb{R}_{\geq 0}, x_i \mapsto p_{\xi_i}(x_i) := \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_m} p_\xi(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_m)\, dx_1 \ldots dx_{i-1}\, dx_{i+1} \ldots dx_m. \tag{36}$$

Bemerkungen

- Wir verzichten auf einen Beweis.
- Die WMFen der Marginalverteilungen diskreter Zufallsvektoren ergeben sich durch Summation.
- Die WDFen der Marginalverteilungen kontinuierlicher Zufallsvektoren ergeben sich durch Integration.

Beispiel (Marginale Wahrscheinlichkeitsmassefunktionen)

Wir betrachten erneut den zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ der Werte in $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ annimmt, wobei $\mathcal{X}_1 := \{1, 2, 3\}$ und $\mathcal{X}_2 = \{1, 2, 3, 4\}$ seien.

Basierend auf der oben definierten WMF ergeben sich folgende marginale WMFen p_{ξ_1} und p_{ξ_2}:

$p_\xi(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$	$p_{\xi_1}(x_1)$
$x_1 = 1$	0.1	0.0	0.2	0.1	0.4
$x_1 = 2$	0.1	0.2	0.0	0.0	0.3
$x_1 = 3$	0.0	0.1	0.1	0.1	0.3
$p_{\xi_2}(x_2)$	0.2	0.3	0.3	0.2	

Man beachte, dass $\sum_{x_1=1}^{3} p_{\xi_1}(x_1) = 1$ und $\sum_{x_2=1}^{3} p_{\xi_2}(x_2) = 1$ gilt.

Beispiel (Marginale Wahrscheinlichkeitsmassefunktionen)

Ein Realisierungsbeispiel mithilfe relativer Häufigkeiten mag den Begriff der marginalen WMF intuitiv verdeutlichen. Nehmen wir an, wir hätten $n = 100$ (unabhängige) Realisierungen von ξ vorliegen.

Um die Wahrscheinlichkeiten $p_\xi(x_1, x_2)$ zu schätzen, würden wir die Anzahl der Realisierungen von (x_1, x_2) zählen und durch n teilen. Hätten wir beispielsweise 12 Realisierungen von $(3, 2)$ vorliegen, so würden wir $p_\xi(3, 2) \approx 12/100 = 0.12$ schätzen.

Die Frage nach der marginalen Wahrscheinlichkeit von $x_2 = 2$ entspräche dann der Frage, wie oft unter den Realisierungen zu finden sind, bei denen $x_2 = 2$ ist, irrespektive des Wertes von x_1. Dies wäre gerade die Anzahl der Realisierungen der Form $(1, 2), (2, 2)$ und $(3, 2)$. Gäbe es von diesen beispielsweise $0, 22$ und 12 respektive, so würde man die Wahrscheinlichkeit $p_{\xi_2}(2)$ natürlicherweise durch

$$\frac{0 + 22 + 12}{100} = \frac{0}{100} + \frac{22}{100} + \frac{12}{100} = 0.00 + 0.22 + 0.12 = 0.34 \tag{37}$$

schätzen. Anstelle der Wahrscheinlichkeiten $p_\xi(1, 2)$, $p_\xi(2, 2)$, $p_\xi(3, 2)$ addiert man hier also die entsprechenden relativen Häufigkeiten.

Vorbemerkungen zu bedingten Verteilungen

Wir erinnern uns, dass für einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ und zwei Ereignisse $A, B \in \mathcal{A}$ mit $\mathbb{P}(B) > 0$ die bedingte Wahrscheinlichkeit von A gegeben B definiert ist als

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{38}$$

Analog wird für zwei Zufallsvariablen ξ_1, ξ_2 mit Ereignisräumen $\mathcal{X}_1, \mathcal{X}_2$ und (messbaren) Mengen $S_1 \in \mathcal{X}_1, S_2 \in \mathcal{X}_2$ die bedingte Verteilung von ξ_1 gegeben ξ_2 mithilfe der Ereignisse

$$A := \{\xi_1 \in S_1\} \text{ und } B := \{\xi_2 \in S_2\} \tag{39}$$

definiert.

So ergibt sich zum Beispiel die bedingte Wahrscheinlichkeit, dass $\xi_1 \in S_1$ gegeben dass $\xi_2 \in S_2$ unter der Annahme, dass $\mathbb{P}(\{\xi_2 \in S_2\}) > 0$, zu

$$\mathbb{P}(\{\xi_1 \in S_1\}|\{\xi_2 \in S_2\}) = \frac{\mathbb{P}(\{\xi_1 \in S_1\} \cap \{\xi_2 \in S_2\})}{\mathbb{P}(\{\xi_2 \in S_2\})}. \tag{40}$$

Wir betrachten zunächst durch WMFen/WDFen zweidimensionaler Zufallsvektoren definierte bedingte Verteilungen.

Definition (Bedingte WMF, diskrete bedingte Verteilung)

$\xi := (\xi_1, \xi_2)$ sei ein diskreter Zufallsvektor mit Ergebnisraum $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$, WMF $p_\xi = p_{\xi_1,\xi_2}$ und marginalen WMFen p_{ξ_1} und p_{ξ_2}. Die bedingte WMF von ξ_1 gegeben $\xi_2 = x_2$ ist dann für $p_{\xi_2}(x_2) > 0$ definiert als

$$p_{\xi_1|\xi_2=x_2} : \mathcal{X}_1 \to [0,1], x_1 \mapsto p_{\xi_1|\xi_2=x_2}(x_1|x_2) := \frac{p_{\xi_1,\xi_2}(x_1,x_2)}{p_{\xi_2}(x_2)} \tag{41}$$

Analog ist für $p_{\xi_1}(x_1) > 0$ die bedingte WMF von ξ_2 gegeben $\xi_1 = x_1$ definiert als

$$p_{\xi_2|\xi_1=x_1} : \mathcal{X}_2 \to [0,1], x_2 \mapsto p_{\xi_2|\xi_1=x_1}(x_2|x_1) := \frac{p_{\xi_1,\xi_2}(x_1,x_2)}{p_{\xi_1}(x_1)} \tag{42}$$

Die bedingten Verteilungen mit WMFen $p_{\xi_1|\xi_2=x_2}$ und $p_{\xi_2|\xi_1=x_1}$ heißen dann die *diskreten bedingten Verteilungen von ξ_1 gegeben $\xi_2 = x_2$ und ξ_2 gegeben $\xi_1 = x_1$*, respektive.

Bemerkungen

- In Analogie zur Definition der bedingten Wahrscheinlichkeit von Ereignissen gilt also

$$p_{\xi_1|\xi_2}(x_1|x_2) = \frac{p_{\xi_1,\xi_2}(x_1,x_2)}{p_{\xi_2}(x_2)} = \frac{\mathbb{P}(\{\xi_1 = x_1\} \cap \{\xi_2 = x_2\})}{\mathbb{P}(\{\xi_2 = x_2\})}. \tag{43}$$

- Bedingte Verteilungen sind (lediglich) normalisierte gemeinsame Verteilungen.

Beispiel (Bedingte Wahrscheinlichkeitsmassefunktionen)

Wir betrachten erneut den zweidimensionalen Zufallsvektor $\xi := (\xi_1, \xi_2)$ der Werte in $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$ annimmt, wobei $\mathcal{X}_1 := \{1, 2, 3\}$ und $\mathcal{X}_2 = \{1, 2, 3, 4\}$ seien.

Basierend auf der oben definierten WMF und den entsprechenden oben evaluierten marginalen WMFen ergeben sich folgende bedingte WMFen für $p_{\xi_2|\xi_1=x_1}$

$p_{\xi_2\|\xi_1}(x_2\|x_1)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$p_{\xi_2\|\xi_1=1}(x_2\|x_1 = 1)$	$\frac{0.1}{0.4} = 0.25$	$\frac{0.0}{0.4} = 0.00$	$\frac{0.2}{0.4} = 0.50$	$\frac{0.1}{0.4} = 0.25$
$p_{\xi_2\|\xi_1=2}(x_2\|x_1 = 2)$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.2}{0.3} = 0.6\bar{6}$	$\frac{0.0}{0.3} = 0.00$	$\frac{0.0}{0.3} = 0.00$
$p_{\xi_2\|\xi_1=3}(x_2\|x_1 = 3)$	$\frac{0.0}{0.3} = 0.00$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.3} = 0.3\bar{3}$	$\frac{0.1}{0.3} = 0.3\bar{3}$

Bemerkungen

- Man beachte, dass $\sum_{x_2=1}^{4} p_{\xi_2|\xi_1=x_1}(x_2|x_1) = 1$ für alle $x_1 \in \mathcal{X}_1$.
- Man beachte die qualitative Ähnlichkeit der WMFen $p_{\xi_1,\xi_2}(x_1, x_2)$ und $p_{\xi_2|\xi_1}(x_2|x_1)$.
- Bedingte Verteilungen sind (lediglich) normalisierte gemeinsame Verteilungen.

Definition (Bedingte WDF, kontinuierliche bedingte Verteilungen)

$\xi := (\xi_1, \xi_2)$ sei ein kontinuierlicher Zufallsvektor mit Ergebnisraum $\mathbb{R}^2$, WDF $p_\xi = p_{\xi_1,\xi_2}$ und marginalen WDFen p_{ξ_1} und p_{ξ_2}. Die bedingte WDF von ξ_1 gegeben $\xi_2 = x_2$ ist dann für $p_{\xi_2}(x_2) > 0$ definiert als

$$p_{\xi_1|\xi_2=x_2} : \mathbb{R} \to \mathbb{R}_{\geq 0}, x_1 \mapsto p_{\xi_1|\xi_2=x_2}(x_1|x_2) := \frac{p_{\xi_1,\xi_2}(x_1, x_2)}{p_{\xi_2}(x_2)} \tag{44}$$

Analog ist für $p_{\xi_1}(x_1) > 0$ die bedingte WMF von ξ_2 gegeben $\xi_1 = x_1$ definiert als

$$p_{\xi_2|\xi_1=x_1} : \mathbb{R} \to \mathbb{R}_{\geq 0}, x_2 \mapsto p_{\xi_2|\xi_1=x_1}(x_2|x_1) := \frac{p_{\xi_1,\xi_2}(x_1, x_2)}{p_{\xi_1}(x_1)} \tag{45}$$

Die Verteilungen mit WDFen $p_{\xi_1|\xi_2=x_2}$ und $p_{\xi_2|\xi_1=x_1}$ heißen dann die *kontinuierlichen bedingten Verteilungen von* ξ_1 *gegeben* $\xi_2 = x_2$ *und* ξ_2 *gegeben* $\xi_1 = x_1$, respektive.

Bemerkung

- Im kontinuierlichen Fall gilt zwar $\mathbb{P}(\xi = x) = 0$, aber nicht notwendig auch $p_\xi(x) = 0$.

Definition und multivariate Verteilungen

Marginale und bedingte Verteilungen

Normalverteilungen

Selbstkontrollfragen

Definition (Normalverteilte Zufallsvariable)

ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}$ und WDF

$$p : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{46}$$

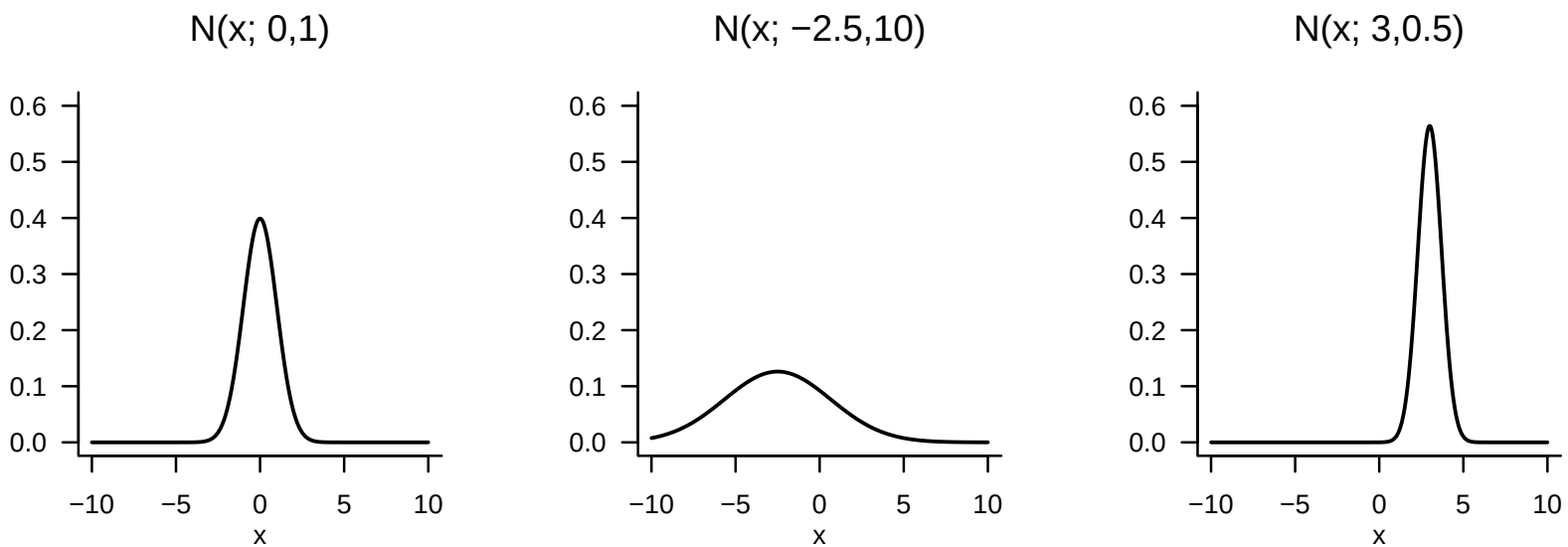
Dann sagen wir, dass ξ einer *Normalverteilung (oder Gauß-Verteilung) mit Erwartungswertparameter* $\mu \in \mathbb{R}$ *und Varianzparameter* $\sigma^2 > 0$ unterliegt und nennen ξ eine *normalverteilte Zufallsvariable*. Wir kürzen dies mit $\xi \sim N(\mu, \sigma^2)$ ab. Die WDF einer normalverteilten Zufallsvariable bezeichnen wir mit

$$N\left(x; \mu, \sigma^2\right) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{47}$$

Bemerkungen

- Es gelten $\mathbb{E}(\xi) = \mu$ und $\mathbb{V}(\xi) = \sigma^2$.
- Der Parameter μ entspricht dem Wert höchster Wahrscheinlichkeitsdichte.
- Der Parameter σ^2 spezifiziert die Breite der WDF.
- $\xi \sim N(0, 1)$ heißt auch *standardnormalverteilt*.

Visualisierung univariater Normalverteilungsdichtefunktionen



Theorem (Konstruktion bivariater Normalverteilungen)

$\zeta_1 \sim N(0,1)$ und $\zeta_2 \sim N(0,1)$ seien zwei unabhängige standardnormalverteilte Zufallsvariablen. Weiterhin seien $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$ und $\rho \in]-1, 1[$. Schließlich seien

$$\begin{aligned} \xi_1 &:= \sigma_1 \zeta_1 + \mu_1 \\ \xi_2 &:= \sigma_2 \left(\rho \zeta_1 + (1-\rho^2)^{1/2} \zeta_2 \right) + \mu_2 . \end{aligned} \tag{48}$$

Dann hat die WDF des Zufallsvektors $\xi := (\xi_1, \xi_2)^T$, also der gemeinsamen Verteilung von ξ_1 und ξ_2, die Form

$$p : \mathbb{R}^2 \to \mathbb{R}_{>0},\ x \mapsto p(x) := (2\pi)^{-1} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right), \tag{49}$$

wobei

$$n := 2, \mu := \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ und } \Sigma := \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix} \tag{50}$$

Bemerkungen

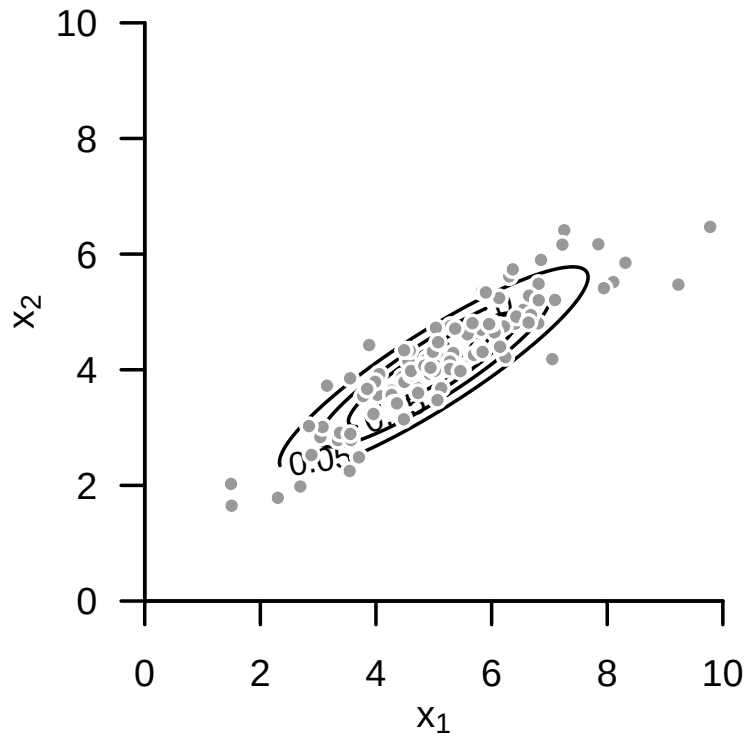
- Man nennt die gemeinsame Verteilung von ξ_1 und ξ_2 *bivariate Normalverteilung*.

Konstruktion bivariater Normalverteilungen

$$\mu_1 := 5.0, \mu_2 := 4.0, \sigma_1 := 1.5, \sigma_2 := 1.0, \rho := 0.9$$

- Realisierungen von $\xi = (\xi_1, \xi_2)^T$

– Isokonturen (Linien gleicher Wahrscheinlichkeitsdichte) von p



Definition (Multivariate Normalverteilung)

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

$$p_\xi : \mathbb{R}^m \to \mathbb{R}_{>0},\ x \mapsto p_\xi(x) := (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \tag{51}$$

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

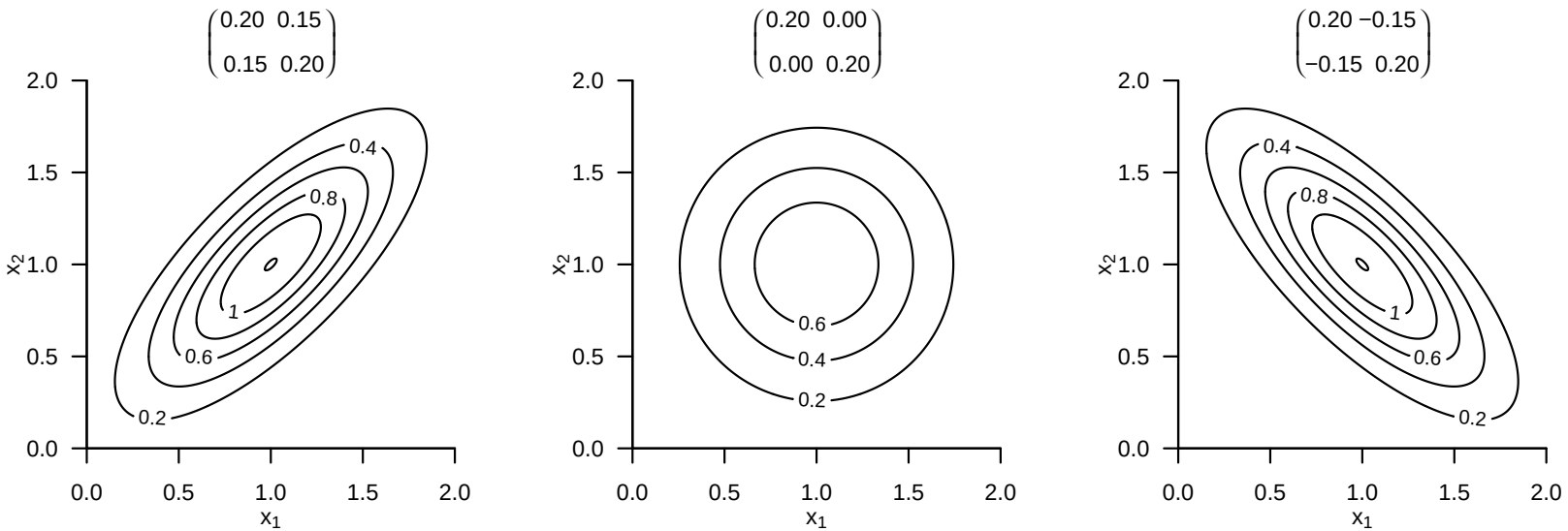
$$N(x; \mu, \Sigma) := (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right). \tag{52}$$

Bemerkungen

- Der Parameter $\mu \in \mathbb{R}^m$ entspricht dem Wert höchster Wahrscheinlichkeitsdichte
- Die Diagonalelemente von Σ spezifizieren die Breite der WDF bezüglich $\xi_1, ..., \xi_m$.
- Das i,jte Element von Σ spezifiziert die Kovarianz on ξ_i und ξ_j.
- Der Term $(2\pi)^{-m/2}|\Sigma|^{-1/2}$ ist die Normalisierungskonstante für den Exponentialfunktionsterm.

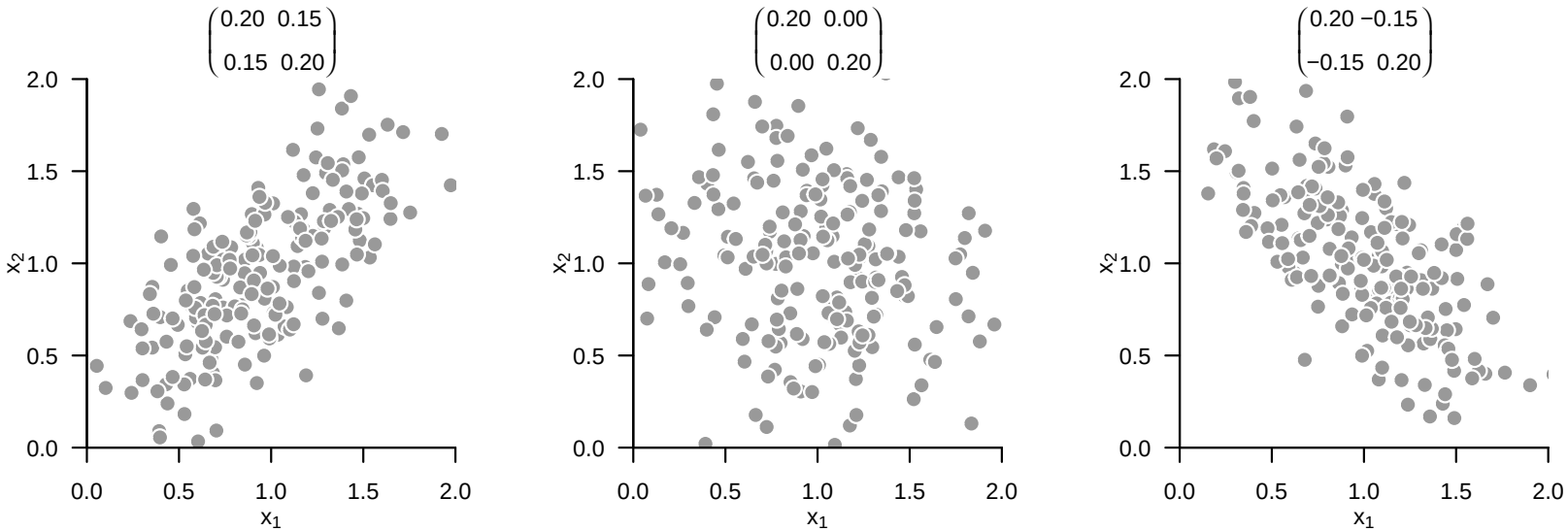
Visualisierung bivariater Normalverteilungsdichtefunktionen

$\mu = (1, 1)^T, \Sigma \in \mathbb{R}^{2\times 2}$ wie im Abbildungstitel vermerkt.



Realisierung bivariater normalverteilter Zufallsvektoren

$\mu = (1, 1)^T, \Sigma \in \mathbb{R}^{2\times 2}$ wie im Abbildungstitel vermerkt.



Theorem (Erwartungswert und Kovarianzmatrix)

$\xi \sim N(\mu, \Sigma)$ sei ein multivariate normalverteilter Zufallsvektor mit Erwartungswertparameter $\mu \in \mathbb{R}^m$ und Kovarianzmatrixparameter $\Sigma \in \mathbb{R}^{m \times m}$pd . Dann gelten für den Erwartungswert und die Kovarianzmatrix von ξ, dass

$$\mathbb{E}(\xi) = \mu \text{ und } \mathbb{C}(\xi) = \Sigma, \tag{53}$$

respektive.

Bemerkungen

- Wir verzichten auf einen Beweis.
- Das Theorem ist die direkte Generalisierung der Eigenschaften univariater normalverteilter Zufallsvariablen

Einige Eigenschaften von multivariaten Normalverteilungen

Wie bei univariaten Normalverteilungen gilt bei multivariaten Normalverteilungen, dass linear-affine Transformationen wiederrum auf Normalverteilugen führen, deren Parameter anhand der Ausgangsparameter und der Transformationsparameter errechnet werden können.

Multivariate Normalverteilungen haben weiterhin die Eigenschaft, dass auch alle anderen assoziierten Verteilung Normalverteilungen sind und deren Erwartungswert- und Kovarianzmatrixparameter aus den Parametern der jeweils komplementären Verteilung errechnet werden können.

Insbesondere gelten:

(1) Die uni- und multivariaten Marginalverteilungen multivariater Normalverteilungen sind Normalverteilungen.
(2) Wie alle multivariaten Verteilungen lassen sich multivariate Normalverteilungen multiplikativ in eine marginale und eine bedingte Verteilung zerlegen. Insbesondere sind bei multivariaten Normalverteilungen diese Verteilungen auch Normalverteilungen, deren Parameter aus den Parametern der gemeinsame Verteilung errechnet werden können und umgekehrt.

Wir fassen diese Einsichten in den folgenden Theoremen zusammen.

Theorem (Linear-affine Transformation)

$\xi \sim N(\mu, \Sigma)$ sei ein normalverteilter m-dimensionaler Zufallsvektor und es sei

$$v := A\xi + b \text{ mit } A \in \mathbb{R}^{n \times m} \text{ und } b \in \mathbb{R}^n. \tag{54}$$

Dann gilt

$$v \sim N\left(A\mu + b, A\Sigma A^T\right). \tag{55}$$

Bemerkung

- Die linear-affine Transformation eines multivariate normalverteilten Zuallsvektors ergibt wieder ein normalverteilten Zufallsvektor. Die Parameter des resultierenden normalverteilten Zufallsvektors ergeben sich dabei aus den Parametern des ursprünglichen Zufallsvektors und der Transformationsparamter.

Theorem (Marginale Normalverteilungen)

Es sei $m := k + l$ und $\xi = (\xi_1, ..., \xi_m)^T$ sei ein m-dimensionaler normalverteilter Zufallsvektor mit Erwartungswertparameter

$$\mu = \begin{pmatrix} \mu_\upsilon \\ \mu_\zeta \end{pmatrix} \in \mathbb{R}^m, \tag{56}$$

mit $\mu_\upsilon \in \mathbb{R}^k$ and $\mu_\zeta \in \mathbb{R}^l$ und Kovarianzmatrixparameter

$$\Sigma = \begin{pmatrix} \Sigma_{\upsilon\upsilon} & \Sigma_{\upsilon\zeta} \\ \Sigma_{\zeta\upsilon} & \Sigma_{\zeta\zeta} \end{pmatrix} \in \mathbb{R}^{m \times m}, \tag{57}$$

mit $\Sigma_{\upsilon\upsilon} \in \mathbb{R}^{k \times k}$, $\Sigma_{\upsilon\zeta} \in \mathbb{R}^{k \times l}$, $\Sigma_{\zeta\upsilon} \in \mathbb{R}^{l \times k}$, und $\Sigma_{\zeta\zeta} \in \mathbb{R}^{l \times l}$. Dann sind $\upsilon := (\xi_1, ..., \xi_k)^T$ und $\zeta := (\xi_{k+1}, ..., \xi_m)^T$ k- und l-dimensionale normalverteilte Zufallsvektoren, respektive, und es gilt

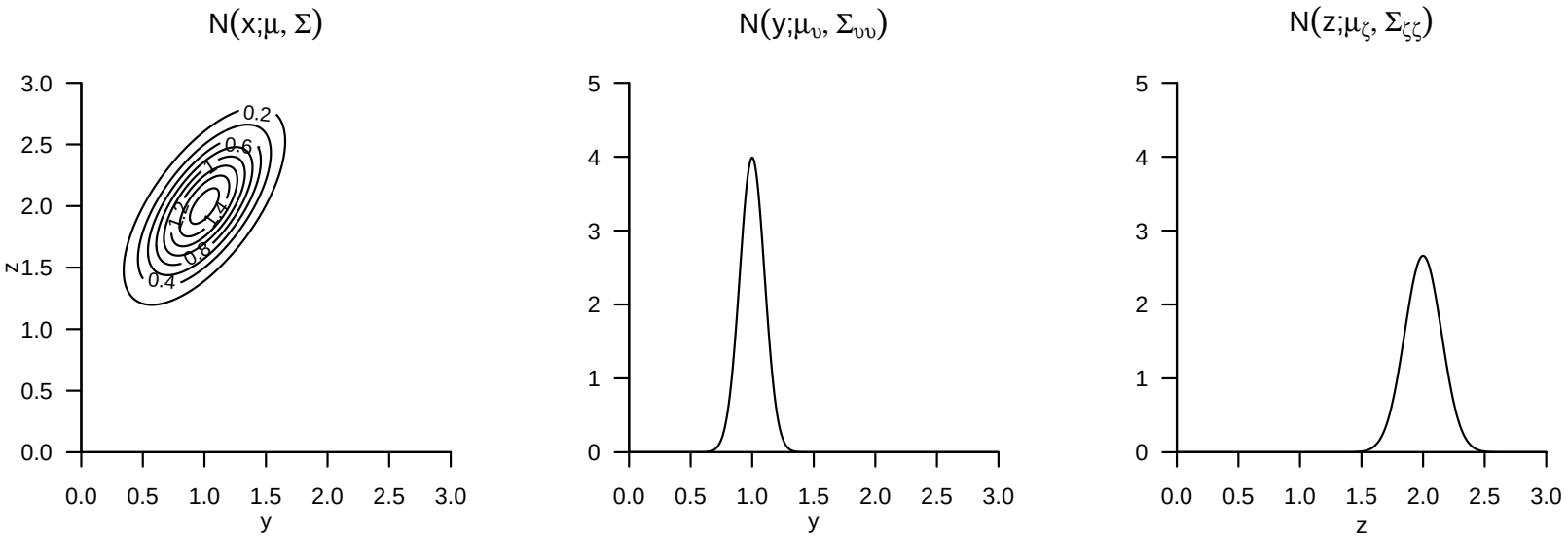
$$\upsilon \sim N(\mu_\upsilon, \Sigma_{\upsilon\upsilon}) \text{ and } \zeta \sim N(\mu_\zeta, \Sigma_{\zeta\zeta}), \tag{58}$$

Bemerkungen

- Die Marginalverteilungen einer multivariaten Normalverteilung sind auch Normalverteilungen.
- Die Parameter der Marginalverteilungen ergeben sich aus den Parametern der gemeinsamen Verteilung.

Marginale Normalverteilungen

$$m := 2, k = 1, l = 1, \mu := \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \Sigma := \begin{pmatrix} 0.10 & 0.08 \\ 0.08 & 0.15 \end{pmatrix}$$



Theorem (Gemeinsame Normalverteilungen)

ξ sei ein m-dimensionaler normalverteilter Zufallsvektor mit WDF

$$p_\xi : \mathbb{R}^m \to \mathbb{R}_{>0},\ x \mapsto p_\xi(x) := N(x; \mu_\xi, \Sigma_{\xi\xi}) \text{ mit } \mu_\xi \in \mathbb{R}^m, \Sigma_{\xi\xi} \in \mathbb{R}^{m\times m}, \tag{59}$$

$A \in \mathbb{R}^{n\times m}$ sei eine Matrix, $b \in \mathbb{R}^n$ sei ein Vektor und υ sei ein n-dimensionaler bedingt normalverteilter Zufallsvektor mit bedingter WDF

$$p_{\upsilon|\xi}(\cdot|x) : \mathbb{R}^n \to \mathbb{R}_{>0},\ y \mapsto p_{\upsilon|\xi}(y|x) := N(y; A\xi + b, \Sigma_{\upsilon\upsilon}) \text{ mit } \Sigma_{\upsilon\upsilon} \in \mathbb{R}^{n\times n}. \tag{60}$$

Dann ist der $m+n$-dimensionale Zufallsvektor $(\xi, \upsilon)^T$ normalverteilt mit (gemeinsamer) WDF

$$p_{\xi,\upsilon} : \mathbb{R}^{m+n} \to \mathbb{R}_{>0},\ \begin{pmatrix} x \\ y \end{pmatrix} \mapsto p_{\xi,\upsilon}\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \mu_{\xi,\upsilon}, \Sigma_{\xi,\upsilon}\right), \tag{61}$$

mit $\mu_{\xi,\upsilon} \in \mathbb{R}^{m+n}$ and $\Sigma_{\xi,\upsilon} \in \mathbb{R}^{m+n\times m+n}$ und insbesondere

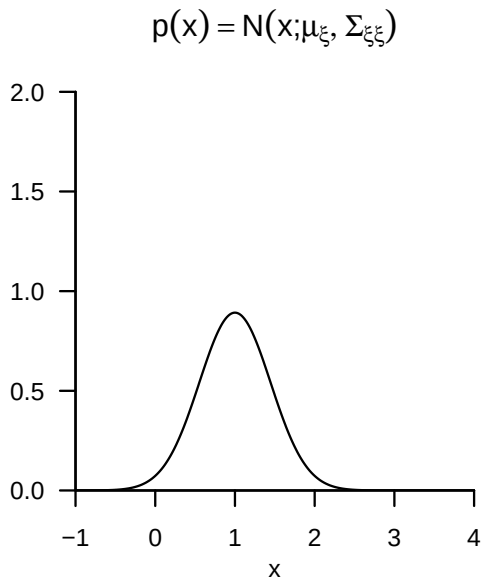
$$\mu_{\xi,\upsilon} = \begin{pmatrix} \mu_\xi \\ A\mu_\xi + b \end{pmatrix} \text{ und } \Sigma_{\xi,\upsilon} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\xi}A^T \\ A\Sigma_{\xi\xi} & \Sigma_{\upsilon\upsilon} + A\Sigma_{\xi\xi}A^T \end{pmatrix}. \tag{62}$$

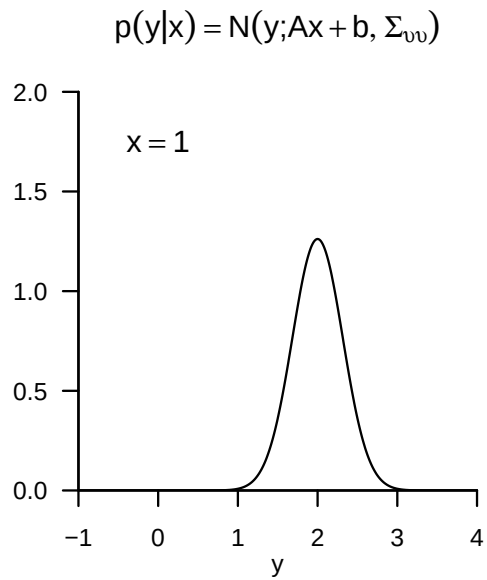
Bemerkungen

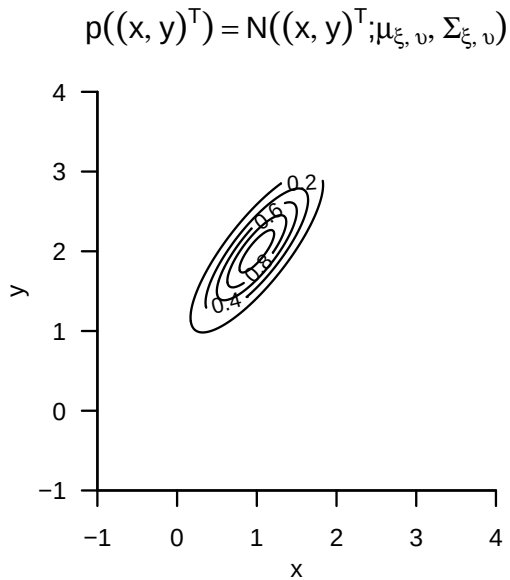
- Eine marginale und eine bedingte multivariate Normalverteilung induzieren eine gemeinsame Normalverteilung.
- Die Parameter der gemeinsamen Verteilungen ergeben sich als linear-affine Transformation der Parameter der induzierenden Verteilungen.

Gemeinsame Normalverteilungen

$$m := 1, n := 1, \mu_\xi := 1, \Sigma_{\xi\xi} := 0.2, A := 1, b := 1, \Sigma_{\upsilon\upsilon} := 0.1$$







Theorem (Bedingte Normalverteilungen)

(ξ, υ) sei ein $m+n$-dimensionaler normalverteilter Zufallsvektor mit WDF

$$p_{\xi,\upsilon} : \mathbb{R}^{m+n} \to \mathbb{R}_{>0}, \begin{pmatrix} x \\ y \end{pmatrix} \mapsto p_{\xi,\upsilon}\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) := N\left(\begin{pmatrix} x \\ y \end{pmatrix}; \mu_{\xi,\upsilon}, \Sigma_{\xi,\upsilon}\right), \tag{63}$$

mit

$$\mu_{\xi,\upsilon} = \begin{pmatrix} \mu_\xi \\ \mu_\upsilon \end{pmatrix}, \Sigma_{\xi,\upsilon} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\upsilon} \\ \Sigma_{\upsilon\xi} & \Sigma_{\upsilon\upsilon} \end{pmatrix}, \tag{64}$$

mit $x, \mu_\xi \in \mathbb{R}^m, y, \mu_\upsilon \in \mathbb{R}^n$ and $\Sigma_{\xi\xi} \in \mathbb{R}^{m\times m}, \Sigma_{\xi\upsilon} \in \mathbb{R}^{m\times n}, \Sigma_{\upsilon\upsilon} \in \mathbb{R}^{n\times n}$. Dann ist die bedingte Verteilung von ξ gegeben υ eine m-dimensionale Normalverteilung mit bedingter WDF

$$p_{\xi|\upsilon}(\cdot|y) : \mathbb{R}^m \to \mathbb{R}_{>0}, x \mapsto p_{\xi|\upsilon}(x|y) := N(x; \mu_{\xi|\upsilon}, \Sigma_{\xi|\upsilon}) \tag{65}$$

mit

$$\mu_{\xi|\upsilon} = \mu_\xi + \Sigma_{\xi\upsilon}\Sigma_{\upsilon\upsilon}^{-1}(y - \mu_\upsilon) \in \mathbb{R}^m \tag{66}$$

und

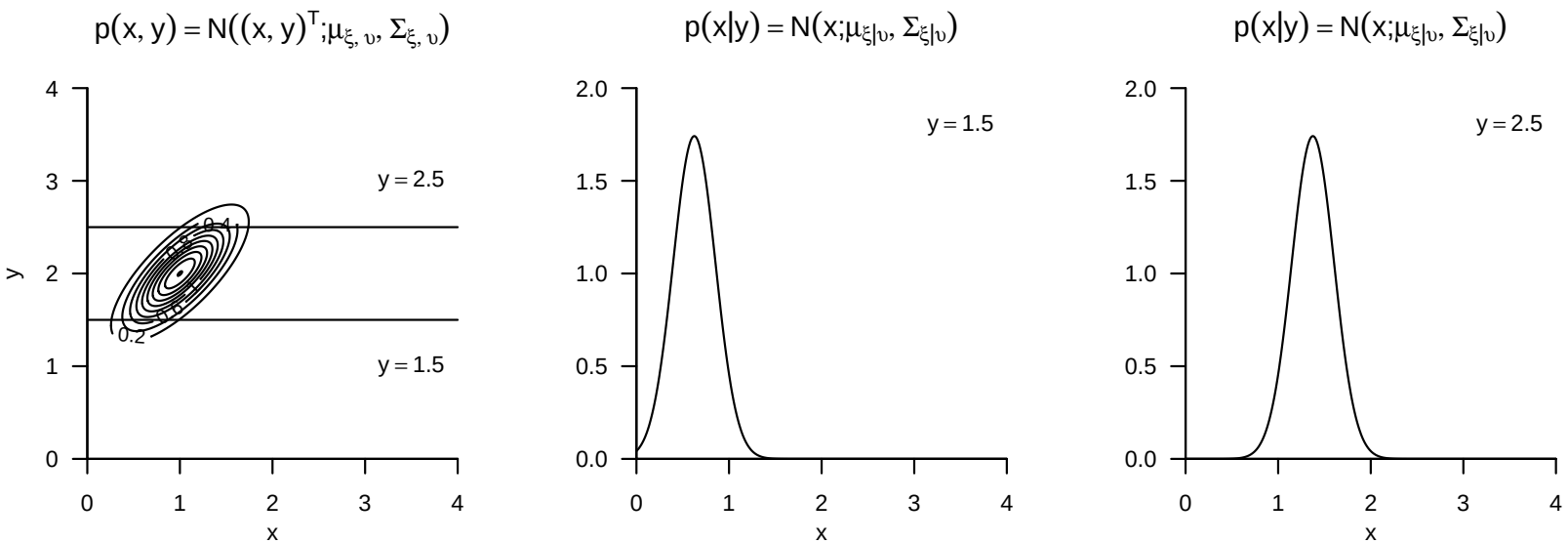
$$\Sigma_{\xi|\upsilon} = \Sigma_{\xi\xi} - \Sigma_{\xi\upsilon}\Sigma_{\upsilon\upsilon}^{-1}\Sigma_{\upsilon\xi} \in \mathbb{R}^{m\times m}. \tag{67}$$

Bemerkungen

- Die Parameter einer bedingten (multivariaten) Normalverteilung ergeben sich aus den Parametern einer gemeinsamen multivariaten Normalverteilung. Im Zusammenspiel mit den vorherigen Theoremen können die Parameter bedingter und marginale Normalverteilungen aus den Parametern der komplementären bedingten und marginalen Normalverteilungen bestimmt werden.

Bedingte Normalverteilungen

$$m := 2, n := 1, \mu := \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \Sigma := \begin{pmatrix} 0.12 & 0.09 \\ 0.09 & 0.12 \end{pmatrix}$$



Definition und multivariate Verteilungen

Marginale und bedingte Verteilungen

Normalverteilungen

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie Definition eines Zufallsvektors wieder.
2. Geben Sie Definition der multivariaten Verteilung eines Zufallsvektors wieder.
3. Geben Sie Definition einer multivariaten WMF wieder.
4. Geben Sie Definition einer multivariaten WDF wieder.
5. Geben Sie die Definition des Erwartungswerts eines Zufallsvektors wieder.
6. Geben Sie die Definition der Kovarianzmatrix eines Zufallsvektors wieder.
7. Was repräsentieren die Diagonalelemente der Kovarianzmatrix eines Zufallsvektors?
8. Was repräsentieren die Nichtdiagonalelemente der Kovarianzmatrix eines Zufallsvektors?
9. Geben Sie die Definition der Korrelationsmatrix eines Zufallsvektors wieder.
10. Geben Sie die Definition der univariaten Marginalverteilung eines Zufallsvektors wieder.
11. Wie berechnet man die WMF der $i$ten Komponente eines diskreten Zufallsvektors?
12. Wie berechnet man die WDF der $i$ten Komponente eines kontinuierlichen Zufallsvektors?
13. Geben Sie Definition der bedingten WMF und der diskreten bedingten Verteilung wieder.
14. Geben Sie Definition der bedingten WDF und der kontinuierlichen bedingten Verteilung wieder.
15. Geben Sie die Definition der WDF eines multivariaten normalverteilten Zufallsvektors wieder.
16. Erläutern Sie die Komponenten der WDF eines multivariaten normalverteilten Zufallsvektors.
17. Welche Werte haben der Erwartungswert und die Kovarianzmatrix eines normalverteilten Zufallsvektors?

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(4) Deskription und Inferenz

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Datenanalyseszenarien

UV	AV	Datenanalysemethoden
Univariat	Univariat	Korrelation, Einfache Regression,T-Tests
Multivariat	Univariat	Multiple Korrelation, Multiple Regression, Allgemeines Lineares Modell
Univariat	Multivariat	Einstichproben-T^2-Tests, Einfaktorielle multivariate Varianzanalyse
Multivariat	Multivariat	Kanonische Korrelation, Multivariates Allgemeines Lineares Modell

Korrelation, Einfache Regression, T-Tests

UV	AV
x_1	y_1
x_{11}	y_{11}
x_{12}	y_{12}
x_{13}	y_{13}
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_{1n}	y_{1n}

Multiple Korrelation, Multiple Regression, Allgemeines Lineares Modell

UV			AV
x_1	$\cdots$	x_m	y_1
x_{11}	$\cdots$	x_{m1}	y_{11}
x_{12}	$\cdots$	x_{m2}	y_{12}
x_{13}	$\cdots$	x_{m3}	y_{13}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	x_{mn}	y_{1n}

Einstichproben-T^2-Tests, Einfaktorielle multivariate Varianzanalyse

UV	AV		
x_1	y_1	$\cdots$	y_m
x_{11}	y_{12}	$\cdots$	y_{m1}
x_{12}	y_{13}	$\cdots$	y_{m2}
x_{13}	y_{14}	$\cdots$	y_{m3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	y_{1n}	$\cdots$	y_{mn}

Kanonische Korrelationsanalyse

Multivariates Allgemeines Lineares Modell

UV			AV		
x_1	$\cdots$	x_{m_x}	y_1	$\cdots$	y_{m_y}
x_{11}	$\cdots$	$x_{m_x 1}$	y_{11}	$\cdots$	$y_{m_y 1}$
x_{12}	$\cdots$	$x_{m_x 2}$	y_{12}	$\cdots$	$y_{m_y 2}$
x_{13}	$\cdots$	$x_{m_x 3}$	y_{13}	$\cdots$	$y_{m_y 3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	$x_{m_x n}$	y_{1n}	$\cdots$	$y_{m_y n}$

Multivariate Generalisierungen bekannter Frequentistischer Verfahren WiSe 23/24

Einstichproben-T^2-Tests als Generalisierung von Einstichproben-T-Tests

- Inferenz für ein bis zwei Gruppen multivariater Daten

Einfaktorielle multivariate Varianzanalyse als Generalisierung der einfaktoriellen Varianzanalyse

- Inferenz für drei oder mehr Gruppen multivariater Daten

Kanonische Korrelationsanalyse als Generalisierung der Korrelation

- Zusammenhangsmaß für multivariate unabhängige und abhängige Variablen

Zur Revision univariater Frequentistischer Verfahren

- Wahrscheinlichkeitstheorie und Frequentistische Inferenz 2022/23
- Allgemeines Lineares Modell 2023

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Definition (Multivariate Deskriptivstatistiken)

$\upsilon_1, ..., \upsilon_n$ seien m-dimensionale Zufallsvektoren.

- Das *Stichprobenmittel* der $\upsilon_1, ..., \upsilon_n$ ist definiert als der m-dimensionale Vektor

$$\bar{\upsilon} := \frac{1}{n} \sum_{i=1}^{n} \upsilon_i. \tag{1}$$

- Die *Stichprobenkovarianzmatrix* der $\upsilon_1, ..., \upsilon_n$ ist definiert als die $m \times m$-dimensionale Matrix

$$C := \frac{1}{n-1} \sum_{i=1}^{n} (\upsilon_i - \bar{\upsilon})(\upsilon_i - \bar{\upsilon})^T. \tag{2}$$

- Die *Stichprobenkorrelationsmatrix* der $\upsilon_1, ..., \upsilon_n$ definiert als die $m \times m$-dimensionale Matrix

$$R := \left(\frac{(C)_{ij}}{\sqrt{(C)_{ii}}\sqrt{(C)_{jj}}} \right)_{1 \leq i,j \leq m}. \tag{3}$$

Bemerkungen

- Bei unabhängig und identisch verteilten $\upsilon_1, ..., \upsilon_n$ ist $\bar{\upsilon}$ ein unverzerrter Schätzer von $\mathbb{E}(\upsilon_i), i = 1, ..., n$.
- Bei unabhängig und identisch verteilten $\upsilon_1, ..., \upsilon_n$ ist C ein unverzerrter Schätzer von $\mathbb{C}(\upsilon_i), i = 1, ..., n$.

Theorem (Datenmatrix und multivariate Deskriptivstatistiken)

$$\Upsilon := \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix} \tag{4}$$

sei eine $m \times n$ *Datenmatrix*, die durch die spaltenweise Konkatenation von m-dimensionaler Zufallvektoren $\upsilon_1, \ldots, \upsilon_n$ gegeben sei. Dann ergeben sich

- für das Stichprobenmittel

$$\bar{\upsilon} = \frac{1}{n} \Upsilon 1_n, \tag{5}$$

- für die Stichprobenkovarianzmatrix

$$C = \frac{1}{n-1} \left(\Upsilon \left(I_n - \frac{1}{n} 1_{nn} \right) \Upsilon^T \right), \tag{6}$$

- und mit

$$D := \text{diag} \left(\sqrt{(C)_{ii}}^{-1}, i = 1, \ldots, m \right) \tag{7}$$

für die Stichprobenkorrelationsmatrix

$$R = DCD \tag{8}$$

Bemerkungen

- Das Theorem erlaubt eine mathematisch konzise Darstellung von $\bar{\upsilon}, C$ und R.
- Das Theorem erlaubt eine programmatisch effiziente Berechung von $\bar{\upsilon}, C$ und R.

Beweis

Die Darstellung des Stichprobenmittels ergibt sich nach durch

$$\bar{\upsilon} := \frac{1}{n}\sum_{i=1}^{n}\upsilon_i = \frac{1}{n}\begin{pmatrix}\sum_{i=1}^{n}\upsilon_{i1}\\ \vdots \\ \sum_{i=1}^{n}\upsilon_{im}\end{pmatrix} = \frac{1}{n}\left(\begin{pmatrix}\upsilon_{11} & \cdots & \upsilon_{n1}\\ \vdots & \ddots & \vdots \\ \upsilon_{1m} & \cdots & \upsilon_{nm}\end{pmatrix}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right) = \frac{1}{n}\Upsilon 1_n. \tag{9}$$

Hinsichtlich der Darstellung der Stichprobenkovarianzmatrix halten wir zunächst fest, dass gilt, dass

$$\begin{aligned}
C &:= \frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})(\upsilon_i - \bar{\upsilon})^T \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\upsilon_i\upsilon_i^T - \upsilon_i\bar{\upsilon}^T - \bar{\upsilon}\upsilon_i^T + \bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - \sum_{i=1}^{n}\upsilon_i\bar{\upsilon}^T - \sum_{i=1}^{n}\bar{\upsilon}\upsilon_i^T + \sum_{i=1}^{n}\bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - n\bar{\upsilon}\bar{\upsilon}^T - n\bar{\upsilon}\bar{\upsilon}^T + n\bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - n\bar{\upsilon}\bar{\upsilon}^T\right)
\end{aligned} \tag{10}$$

Beweis (fortgeführt)

Mit $1_n 1_n^T = 1_{nn}$ ergibt sich dann weiterhin

$$\begin{aligned}
\Upsilon \left(I_n - \frac{1}{n} 1_{nn} \right) \Upsilon^T &= \left(\Upsilon I_n - \frac{1}{n} \Upsilon 1_{nn} \right) \Upsilon^T \\
&= \Upsilon \Upsilon^T - \frac{1}{n} \Upsilon 1_{nn} \Upsilon^T \\
&= \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix} \begin{pmatrix} \upsilon_1^T \\ \vdots \\ \upsilon_n^T \end{pmatrix} - \frac{1}{n} \Upsilon 1_n 1_n^T \Upsilon^T \\
&= \sum_{i=1}^{n} \upsilon_i \upsilon_i^T - n \left(\frac{1}{n} \Upsilon 1_n \right) \left(\frac{1}{n} 1_n^T \Upsilon^T \right) \qquad (11) \\
&= \sum_{i=1}^{n} \upsilon_i \upsilon_i^T - n \bar{\upsilon} \bar{\upsilon}^T \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (\upsilon_i - \bar{\upsilon})(\upsilon_i - \bar{\upsilon})^T \\
&= C.
\end{aligned}$$

Schließlich ergibt sich für die Korrelationsmatrix für ein beliebiges Indexpaar i, j mit $1 \leq i, j \leq m$, dass

$$R_{y_{ij}} = \frac{(C)_{ij}}{\sqrt{(C)_{ii}} \sqrt{(C)_{jj}}} = \frac{1}{\sqrt{(C)_{ii}}} (C)_{ij} \frac{1}{\sqrt{(C)_{jj}}} = (DCD)_{ij}. \qquad (12)$$

Definition (Mahalanobis Distanz)

ξ_1 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel, ξ_2 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel und Ξ sei eine Kovarianzmatrix oder eine Stichprobenkovarianzmatrix. Dann heißt

$$D = (\xi_1 - \xi_2)^T \, \Xi^{-1} \, (\xi_1 - \xi_2) \tag{13}$$

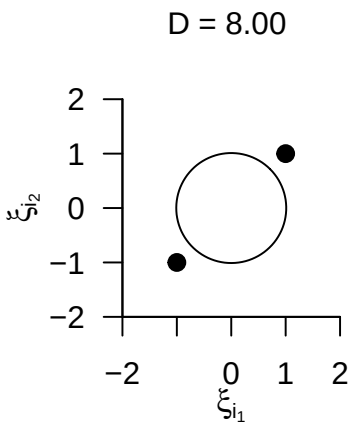
Mahalanobis Distanz von ξ_1 und ξ_2 hinsichtlich Ξ.

Bemerkungen

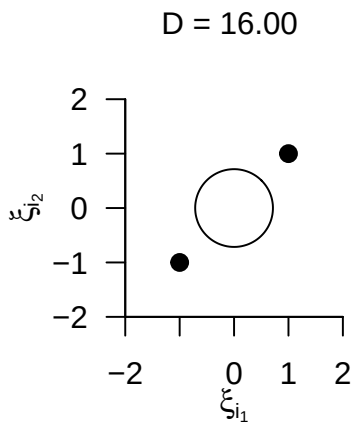
- Eine Mahalanobis Distanz ist eine Kovarianzmatrix-normalisierte quadrierte Euklidische Distanz.
- Ähnliche Maße in der univariaten Statistik sind die z-Transformation $z = \frac{y-\mu}{\sigma}$ und Cohen's $d = \frac{\bar{v}_1 - \bar{v}_2}{s_{12}}$.
- Ähnlich wie bei z-Werten wird bei der Mahalanobis Distanz in "Einheiten von Kovarianzen" gemessen.
- Stark variante Komponenten von ξ_1 und ξ_2 tragen weniger zur Distanz bei.
- Stark kovariante Komponenten von ξ_1 und ξ_2 tragen weniger zur Distanz bei.

Mahalanobis Distanzen als Funktion von Komponentenvarianzen

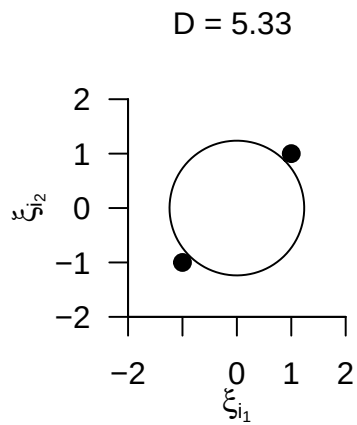
$$\Sigma := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$



$$\Sigma := \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix}$$

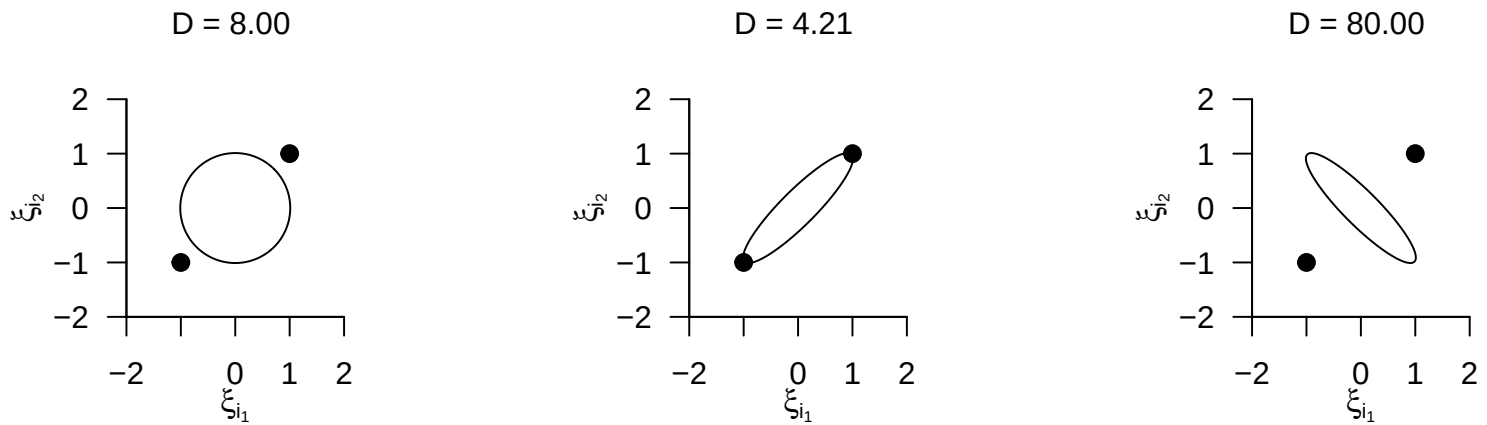


$$\Sigma := \begin{pmatrix} 1.5 & 0.0 \\ 0.0 & 1.5 \end{pmatrix}$$



Mahalanobis Distanzen als Funktion von Komponentenkovarianzen

$$\Sigma := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \qquad \Sigma := \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix} \qquad \Sigma := \begin{pmatrix} 1.0 & -0.9 \\ -0.9 & 1.0 \end{pmatrix}$$

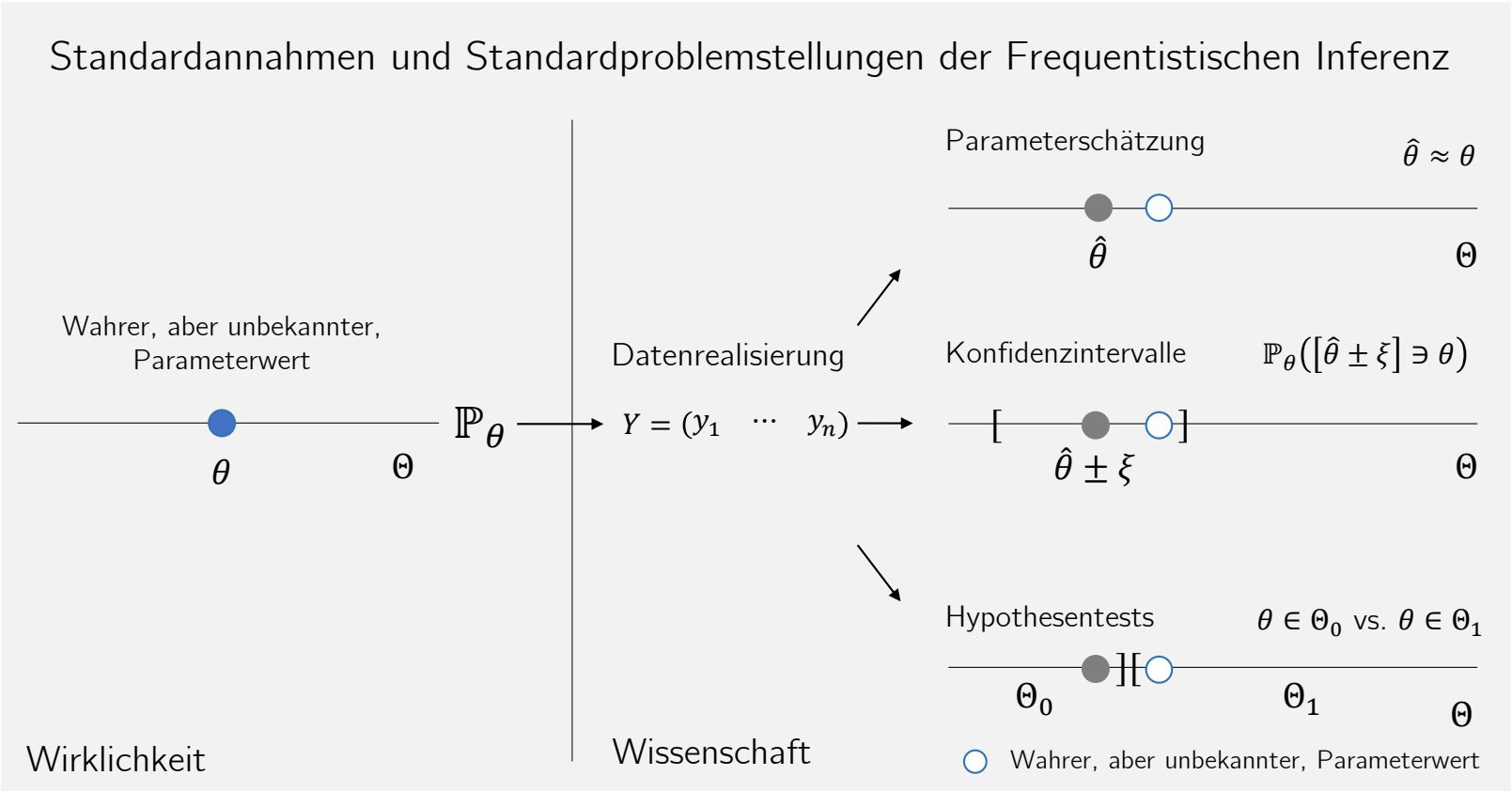


Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen


Standardannahmen und Standardproblemstellungen der Frequentistischen Inferenz
Wahrer, aber unbekannter, Parameterwert
θ
Θ
$\mathbb{P}_\theta$
Datenrealisierung
$Y = (y_1 \quad \cdots \quad y_n)$
Parameterschätzung
$\hat{\theta} \approx \theta$
$\hat{\theta}$
Θ
Konfidenzintervalle
$\mathbb{P}_\theta([\hat{\theta} \pm \xi] \ni \theta)$
$\hat{\theta} \pm \xi$
Θ
Hypothesentests
$\theta \in \Theta_0$ vs. $\theta \in \Theta_1$
Θ_0
Θ_1
Θ
Wirklichkeit
Wissenschaft
Wahrer, aber unbekannter, Parameterwert

Standardannahmen Frequentistischer Inferenz

- $\mathcal{M}$ sei ein Frequentistisches Inferenzmodell mit $\upsilon_1, ..., \upsilon_n \sim p_\theta$. Es wird angenommen, dass eine konkrete Datenmatrix $Y \in \mathbb{R}^{mn}$ eine der möglichen Realisierungen von $\Upsilon = \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix}$ ist.
- Aus Frequentistischer Sicht kann man eine Studie unendlich oft wiederholen und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. das Stichprobenmittel:

$$\text{Datensatz (1)} : Y^{(1)} = \begin{pmatrix} y_1^{(1)} & \cdots & y_n^{(1)} \end{pmatrix} \text{ mit } \bar{y}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(1)}$$

$$\text{Datensatz (2)} : Y^{(2)} = \begin{pmatrix} y_1^{(2)} & \cdots & y_n^{(2)} \end{pmatrix} \text{ mit } \bar{y}^{(2)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(2)}$$

$$\text{Datensatz (3)} : Y^{(3)} = \begin{pmatrix} y_1^{(3)} & \cdots & y_n^{(3)} \end{pmatrix} \text{ mit } \bar{y}^{(3)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(3)}$$

$$\text{Datensatz (4)} : Y^{(4)} = \begin{pmatrix} y_1^{(4)} & \cdots & y_n^{(4)} \end{pmatrix} \text{ mit } \bar{y}^{(4)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(4)}$$

$$\text{Datensatz (5)} : Y^{(5)} = ...$$

- Um die Qualität statistischer Methoden zu beurteilen betrachtet die Frequentistische Statistik deshalb die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme von $\upsilon_1, ..., \upsilon_n \sim p_\theta$. Was zum Beispiel ist die Verteilung der $\bar{y}^{(1)}$, $\bar{y}^{(2)}$, $\bar{y}^{(3)}$, $\bar{y}_n^{(4)}$, ... also die Verteilung der Zufallsvariable $\bar{\upsilon}$?
- Wenn eine statistische Methode im Sinne der Frequentistischen Standardannahmen "gut" ist, dann heißt das also, dass sie bei häufiger Anwendung "im Mittel gut" ist. Im Einzelfall, also im Normalfall nur eines vorliegenden Datensatzes, kann sie auch "schlecht" sein.

Standardproblemstellungen Frequentistischer Inferenz

(1) Parameterschätzung

Ziel der Parameterschätzung ist es, einen möglichst guten Tipp für wahre, aber unbekannte, Parameterwerte oder Funktionen dieser abzugeben, typischerweise mithilfe der Daten.

(2) Konfidenzintervalle

Ziel der Bestimmung von Konfidenzintervallen ist es, basierend auf der angenommenen Verteilung der Daten eine quantitative Aussage über die mit Schätzwerten assoziierte Unsicherheit zu treffen.

(3) Hypothesentests

Ziel des Hypothesentestens ist es, basierend auf der angenommenen Verteilung der Daten in einer möglichst zuverlässigen Form zu entscheiden, ob ein wahrer, aber unbekannter Parameterwert in einer von zwei sich gegenseitig ausschließenden Untermengen des Parameterraumes liegt.

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Selbstkontrollfragen

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(5) Einstichproben-T^2-Tests

Wie im univariaten Fall unterscheidet man im multivariaten Fall

- Z-Tests
- Einstichproben-T^2-Tests
- Zweistichproben-T^2-Tests bei unabhängigen Stichproben
- Zweistichproben-T^2-Tests bei abhängigen Stichproben

Wir betrachten hier exemplarisch Einstichproben-T^2-Tests.

Anwendungsszenario

Hypothesenszenarien

Einfache Nullhypothese $H_0 : \mu = \mu_0$, Einfache Alternativhypothese $H_1 : \mu = \mu_1$

- Theoretisch wichtiges Szenario (Neymann-Pearson Lemma)
- Praktische Relevanz eher gering

Einfache Nullhypothese $H_0 : \mu = \mu_0$, Zusammengesetzte Alternativhypothese $H_1 : \mu \neq \mu_0$

- Zweiseitiger Einstichproben-T^2-Test mit ungerichteter Hypothese
- Ungerichtete Fragestellung nach einem Unterschied

Zusammengesetzte Nullhypothese $H_0 : \mu \leq \mu_0$, Zusammengesetzte Alternativhypothese $H_1 : \mu > \mu_0$

- Einseitiger Einstichproben-T^2-Test mit gerichteter Hypothese
- Gerichtete Fragestellung nach einem positiven Unterschied

Zusammengesetzte Nullhypothese $H_0 : \mu \geq \mu_0$, Zusammengesetzte Alternativhypothese $H_1 : \mu < \mu_0$

- Gerichtete Fragestellung nach einem negativen Unterschied
- Qualitativ äquivalente Theorie zum umgekehrten Fall

Wir betrachten hier exemplarisch Einstichproben-T^2-Tests

mit einfacher Nullhypothese und zusammengesetzter Alternativhypothese.

Anwendungsszenario

Modellformulierung und Modellschätzung

Modellevaluation

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

- Eine Stichprobe/Gruppe n experimenteller Einheiten mit Datendimension $m > 1$.
- Annahme unabhängiger und identisch multivariat normalverteilter Daten.
- Erwartungswertparameter μ und Kovarianzmatrixparameter Σ unbekannt.
- Quantifizieren der Unsicherheit beim Inferenzvergleich von μ mit μ_0 beabsichtigt.

Anwendungsbeispiel

- Gruppenanalyse von BDI und Glukokortikoid Daten
    - $\mu \neq \mu_0$ als Evidenz für eine multivariate Abweichung von einem Normwert μ_0.

Anwendungsszenario

Abweichung des wahren, aber unbekannten, Erwartungswertparameters

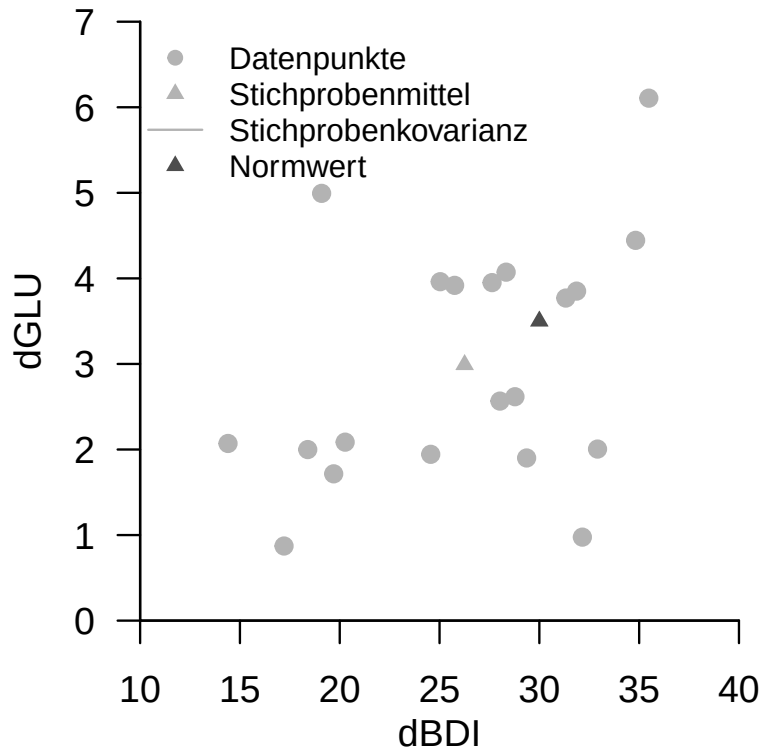
vom Therapieerfolgsnormwert $\mu_0 := (30, 3.5)^T$?

dBDI	dGLU
35	6.1
25	4.0
20	1.7
29	2.6
29	1.9
17	0.9
33	2.0
28	4.1
26	3.9
31	3.8
14	2.1
18	2.0
19	5.0
28	2.6
20	2.1
35	4.4
28	4.0
32	3.9
32	1.0
25	1.9

Anwendungsszenario

Abweichung des wahren, aber unbekannten, Erwartungswertparameters

vom Therapieerfolgsnormwert $\mu_0 := (30, 3.5)^T$?



Anwendungsszenario

Modellformulierung und Modellschätzung

Modellevaluation

Anwendungsbeispiel

Selbstkontrollfragen

Definition (Einstichproben-T^2-Test-Modell)

Für $i = 1, ..., n$ seien v_i m-dimensionale Zufallsvektoren, die die n Datenpunkte eines Einstichproben-T^2-Test Szenarios modellieren. Dann hat das *Einstichproben-T^2-Test-Modell* die strukturelle Form

$$v_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0_m, \Sigma) \text{ u.i.v. für } i = 1, ..., n \text{ mit } \mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m} \text{ pd} \quad (1)$$

und die Datenverteilungsform

$$v_i \sim N(\mu, \Sigma) \text{ u.i.v. für } i = 1, ..., n \text{ mit } \mu \in \mathbb{R}^m, \Sigma \in \mathbb{R}^{m \times m} \text{ pd.} \quad (2)$$

Bemerkungen

- Die Äquivalenz von struktureller Form und Datenverteilungsform des Einstichproben-T^2-Test-Modells folgt direkt aus dem Theorem zu linear-affiner Transformation multivariat normalverteilter Zufallsvektoren.

MODELLSCHÄTZUNG THEOREM OHNE BEWEIS

Anwendungsszenario

Modellformulierung und Modellschätzung

Modellevaluation

Anwendungsbeispiel

Selbstkontrollfragen

Modellevaluation

Zu Gliederung und Wiederholung des univariaten Falls, siehe (12) Hypothesentests

(1) Teststatistik und Test

(2) Analyse der Testgütefunktion

(3) Testumfangkontrolle

(4) p-Werte

(5) Analyse der Powerfunktion

(6) Bestimmung einer optimalen Stichprobengröße

Definition (Einstichproben-T^2-Teststatistik)

Gegeben seien das Einstichproben-T^2-Test-Modell und ein Nullhypothesenparameter $\mu_0 \in \mathbb{R}^m$. Dann ist die Einstichproben-T^2-Teststatistik definiert als

$$T^2 := n(\bar{v} - \mu_0)^T C^{-1} (\bar{v} - \mu_0), \tag{3}$$

wobei $\bar{v}$ und C das Stichprobenmittel und die Stichprobenkovarianzmatrix der $v_1, ..., v_n$ bezeichnen.

Bemerkungen

- T^2 ist die Stichprobenumfang-skalierte Mahalanobis Distanz von $\bar{v}$ und μ_0 hinsichtlich C.
- $T^2 \uparrow$ für $||\bar{v} - \mu_0|| \uparrow$, $C \downarrow$ und $n \uparrow$.

Theorem (Verteilung der skalierten Einstichproben-T^2-Teststatistik)

Es seien $\upsilon_1, ..., \upsilon_n \sim N(\mu, \Sigma)$ mit $\mu \in \mathbb{R}^m$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd,

$$\nu := \frac{n-m}{(n-1)m} \tag{4}$$

und für $\mu \in \mathbb{R}^m$ sei die Einstichproben-T^2-Teststatistik wie oben definiert. Dann gilt

$$\nu T^2 \sim f(\delta, m, n-m) \tag{5}$$

wobei $f(\delta, m, n-m)$ die nichtzentrale f-Verteilung mit Nichtzentralitätsparameter

$$\delta := n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) \tag{6}$$

sowie mit Freiheitsgradparametern m und $n-m$ bezeichnet.

Bemerkungen

- Für einen Beweis von (5) verweisen wir auf Hotelling (1931) und Anderson (2003).
- Für $\mu = \mu_0$ und damit $\delta = 0$ entspricht $f(m, n-m, \delta)$ der f-Verteilung $f(m, n-m)$
- Für $m := 1$ ist $\nu = (n-1)/(n-1) \cdot 1 = 1$ und mit der Stichprobenvarianz S^2 gilt

$$T^2 = n \frac{(\bar{\upsilon} - \mu_0)^2}{S^2} = \left(\sqrt{n} \frac{\bar{\upsilon} - \mu_0}{S} \right)^2 \tag{7}$$

- Das Quadrat der univariaten Einstichproben-T-Teststatistik $T := \sqrt{n} \frac{\bar{\upsilon} - \mu_0}{S}$ ist also $f(\delta, 1, n-1)$ verteilt.
- Wir erinnern nachfolgend an die Begriffe der f-Verteilung und der nichtzentralen f-Verteilung.

Definition (f-Zufallsvariable)

ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und Wahrscheinlichkeitsdichtefunktion

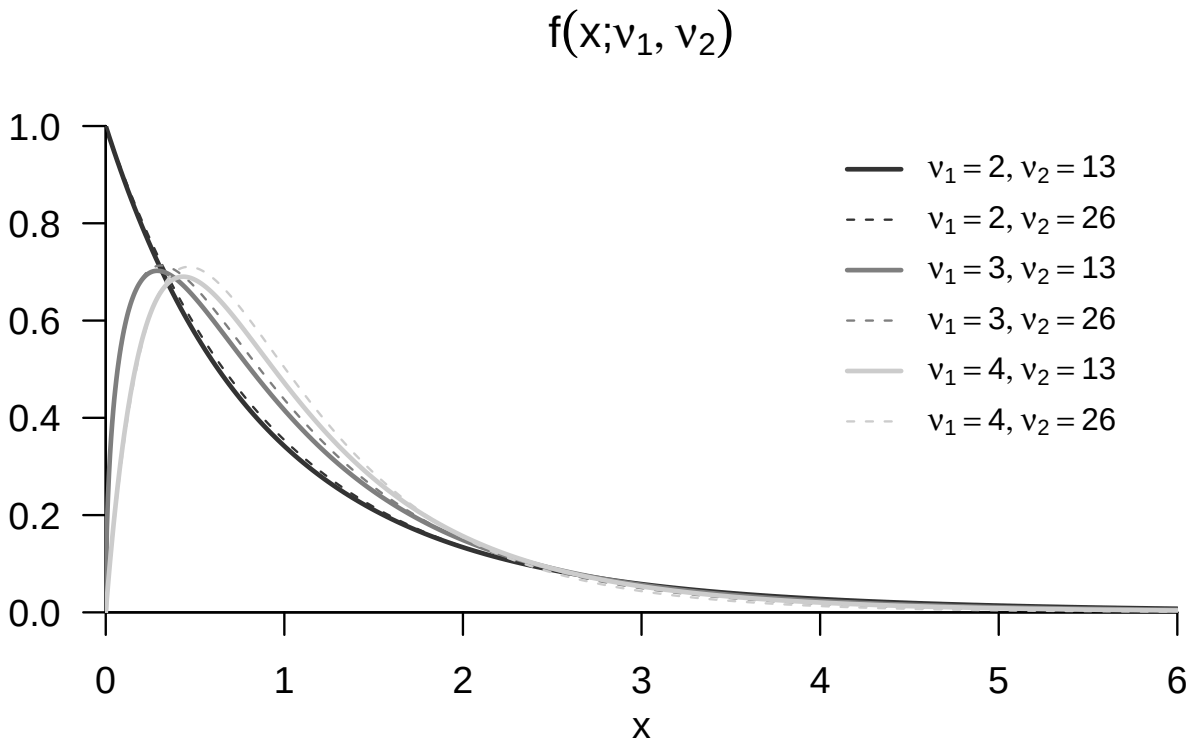
$$p_\xi : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p_\xi(x) := {\nu_1}^{\frac{\nu_1}{2}} {\nu_2}^{\frac{\nu_2}{2}} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{\frac{\nu_1}{2}-1}}{(\nu_1 x+\nu_2)^{\frac{\nu_1+\nu_2}{2}}}, \tag{8}$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass ξ einer f-Verteilung mit Freiheitsgradparametern ν_1 und ν_2 unterliegt und nennen ξ eine f-Zufallsvariable mit Freiheitsgradparametern ν_1 und ν_2. Wir kürzen dies mit $\xi \sim f(\nu_1, \nu_2)$ ab. Die Wahrscheinlichkeitsdichtefunktion (WDF) einer f-Zufallsvariable bezeichnen wir mit $f(x; \nu_1, \nu_2)$, die kumulative Verteilungsfunktion (KVF) einer f-Zufallsvariable bezeichnen wir mit $F(x; \nu_1, \nu_2)$, und die inverse kumulative Verteilungsfunktion einer f-Zufallsvariable bezeichnen wir mit $F^{-1}(x; \nu_1, \nu_2)$.

Bemerkungen

- Im univariaten Fall ist die F-Statistik der Varianzanalyse bei Zutreffen der Nullhypothese f-verteilt
- Im multivariaten Fall ist z.B. die T^2-Statistik bei Zutreffen der Nullhypothese f-verteilt.

Wahrscheinlichkeitsdichtefunktionen von f-Verteilungen



Definition (Nichtzentrale f-Zufallsvariable)

ξ sei eine Zufallsvariable mit Ergebnisraum $\mathbb{R}_{>0}$ und Wahrscheinlichkeitsdichtefunktion

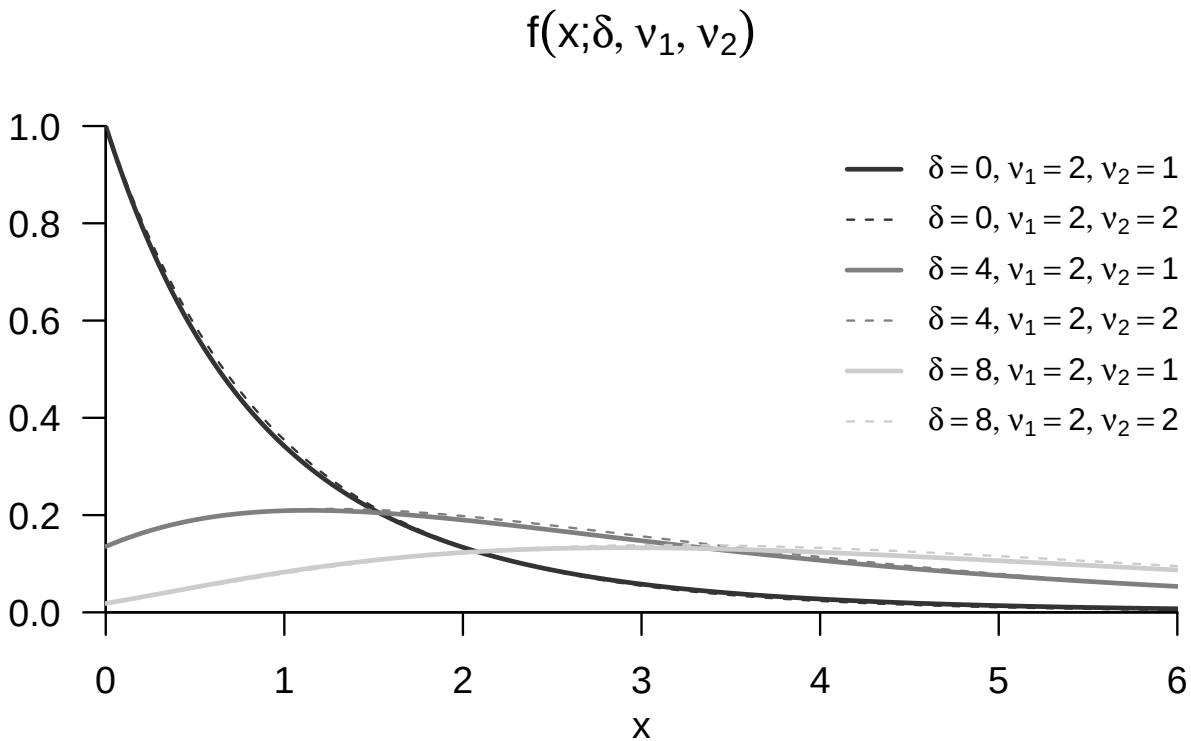
$$p_\xi : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p_\xi(x) := \sum_{k=0}^{\infty} \frac{e^{-\delta/2}(\delta/2)^k}{\frac{\Gamma(\nu_2/2)\Gamma(\nu_1/2+k)}{\Gamma(\nu_2/2+\nu_1/2+k)}\, k!} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2+k} \left(\frac{\nu_2}{\nu_2+\nu_1 x}\right)^{(\nu_1+\nu_2)/2+k} x^{\nu_1/2-1+k} \quad (9)$$

wobei Γ die Gammafunktion bezeichne. Dann sagen wir, dass ξ einer nichtzentralen f-Verteilung mit Nichtzentralitätsparameter δ und Freiheitsgradparametern ν_1 und ν_2 unterliegt und nennen ξ eine nichtzentrale f-Zufallsvariable mit Nichtzentralitätsparameter δ und Freiheitsgradparametern ν_1 und ν_2. Wir kürzen dies mit $\xi \sim f(\delta, \nu_1, \nu_2)$ ab. Die Wahrscheinlichkeitsdichtefunktion (WDF) einer f-Zufallsvariable bezeichnen wir mit $f(x; \delta, \nu_1, \nu_2)$, die kumulative Verteilungsfunktion (KVF) einer nichtzentralen f-Zufallsvariable bezeichnen wir mit $F(x; \delta, \nu_1, \nu_2)$, und die inverse kumulative Verteilungsfunktion einer nichtzentralen f-Zufallsvariable bezeichnen wir mit $F^{-1}(x; \delta, \nu_1, \nu_2)$.

Bemerkungen

- Es gilt $f(0, \nu_1, \nu_2) = f(\nu_1, \nu_2)$.
- Im univariaten Fall ist die F-Statistik bei Nichtzutreffen der Nullhypothese nichtzentral f-verteilt
- Im multivariaten Fall ist z.B. die T^2-Statistik bei Nichtzutreffen der Nullhypothese nichtzentral f-verteilt.

Wahrscheinlichkeitsdichtefunktionen von nichtzentralen f-Verteilungen



Theorem (WDF und KDF der Einstichproben-T^2-Teststatistik)

Im Einstichproben-T^2-Testszenario sei

$$\nu := \frac{n-m}{(n-1)m} \tag{10}$$

Dann ist eine WDF der Einstichproben-T^2-Teststatistik gegeben durch

$$p_{T^2} : \mathbb{R}_{\geq 0} \to \mathbb{R}, t^2 \mapsto p_{T^2}(t^2) := \nu f(\nu t^2; \delta, m, n-m) \tag{11}$$

und eine KDF der Einstichproben-T^2-Teststatistik ist gegeben durch

$$P_{T^2} : \mathbb{R}_{\geq 0} \to [0,1], t^2 \mapsto P_{T^2}(t^2) := F(\nu t^2; \delta, m, n-m) \tag{12}$$

Bemerkungen

- νT^2 hat die WDF $f(\delta, m, n-m)$, T^2 dagegen hat die WDF $\nu f(\nu t^2; \delta, m, n-m)$.

Beweis

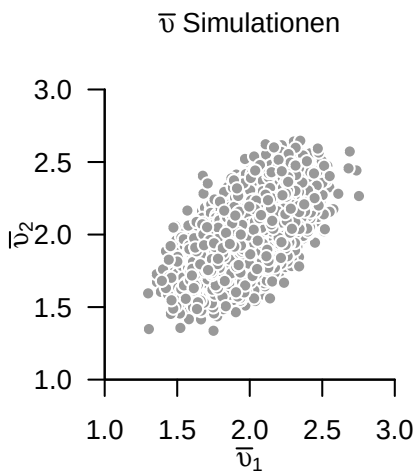
Wir halten zunächst fest, dass das Theorem zur univariate WDF Transformation bei linear-affinen Abbildungen besagt, dass für eine Zufallsvariable ξ mit WDF p_ξ und der Definition $\upsilon = f(\xi)$ mit $f(\xi) := a\xi + b$ für $a \neq 0$ eine WDF von υ definiert ist durch $p_\upsilon(y) := (1/|a|)p_\xi((y-b)/a)$. Im vorliegenden Fall ist $\xi = \nu T^2$ mit WDF $f(\delta, m, n-m)$ und $\upsilon := T^2 = \frac{1}{\nu}\nu T^2$, also $a = 1/\nu$ und $b = 0$. Mit $\nu > 0$ ergibt sich (11) also aus

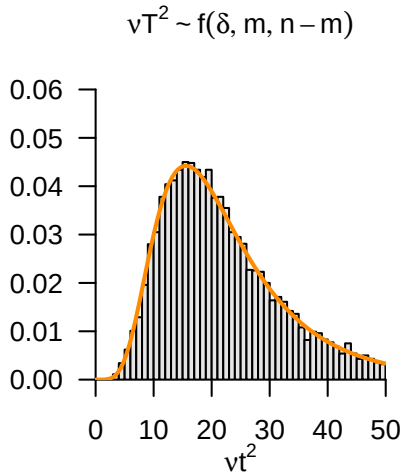
$$p_{T^2}(t^2) = \frac{1}{a} p_{\nu T^2}\left(\frac{t^2}{a}\right) = \nu f(\nu t^2; m, n-m) \tag{13}$$

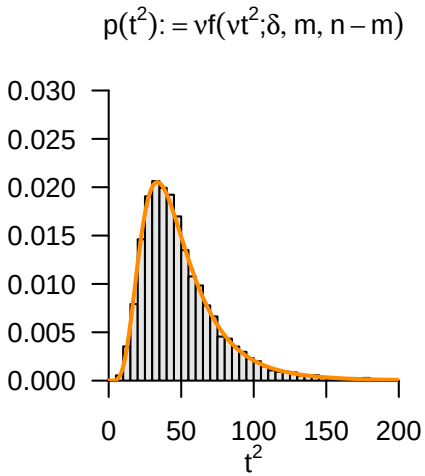
(12) folgt dann mit der Tatsache, dass WDFen bei kontinuierlichen Zufallsvariablen die Ableitungen der entsprechenden KVF sind, sowie der Kettenregel der Differentiation

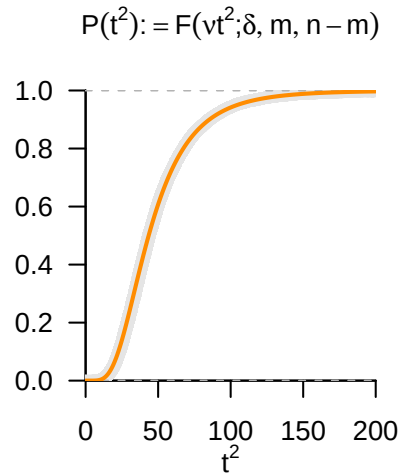
$$\begin{aligned} \frac{d}{dt^2} P_{T^2}\left(t^2\right) &= \frac{d}{dt^2}\left(F(\nu t^2; m, n-m, \delta)\right) \\ &= \frac{d}{dt^2} F(\nu t^2; m, n-m, \delta) \frac{d}{dt^2}\left(\nu t^2\right) \\ &= \nu f(\nu t^2; m, n-m, \delta) \\ &= p_{T^2}(t^2). \end{aligned} \tag{14}$$

□


$\bar{\upsilon}$ Simulationen
3.0
2.5
2.0
1.5
1.0
$\bar{\upsilon}_2$
1.0 1.5 2.0 2.5 3.0
$\bar{\upsilon}_1$


$\nu T^2 \sim f(\delta, m, n-m)$
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0 10 20 30 40 50
νt^2


$p(t^2) := \nu f(\nu t^2; \delta, m, n-m)$
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0 50 100 150 200
t^2


$P(t^2) := F(\nu t^2; \delta, m, n-m)$
1.0
0.8
0.6
0.4
0.2
0.0
0 50 100 150 200
t^2

Definition (Einstichproben-T^2-Test)

Gegeben seien das Einstichproben-T^2-Test-Modell und die Einstichproben-T^2-Teststatistik. Dann ist für $\Upsilon = \begin{pmatrix} v_1 & \cdots & v_n \end{pmatrix}$ und einen kritischen Wert $k \geq 0$ der Einstichproben-T^2-Test definiert als der kritische Wert-basierte Test

$$\phi(\Upsilon) := 1_{\{T^2 > k\}} := \begin{cases} 1 & T^2 > k \\ 0 & T^2 \leq k \end{cases}. \tag{15}$$

Bemerkung

- Wie üblich repräsentiert $\phi(\Upsilon) = 1$ das Ablehnen von H_0.
- Wie üblich repräsentiert $\phi(\Upsilon) = 0$ das Nichtablehnen von H_0.

Definition (Testgütefunktion)

Für einen Test ϕ ist die *Testgütefunktion* definiert als

$$q_\phi : \Theta \to [0,1], \theta \mapsto q_\phi(\theta) := \mathbb{P}_\theta(\phi = 1). \tag{16}$$

Für $\theta \in \Theta_1$ heißt q_ϕ auch *Powerfunktion* oder *Trennschärfefunktion*.

Bemerkung

- Der Wert der Testgütefunktion ist die Wahrscheinlichkeit dafür, dass der Test die Nullhypothese ablehnt.

Theorem (Testgütefunktion)

ϕ sei der im obigen Testszenario definiert Test. Dann ist die Testgütefunktion von ϕ gegeben durch

$$q_\phi : \mathbb{R}^m \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - F(\nu k; \delta_\mu, m, n-m) \tag{17}$$

wobei $F(\cdot; \delta_\mu, m, n-m)$ die KVF der nichtzentralen f-Verteilung mit Freiheitsgradparameter$m$ und $n-m$ sowie mit Nichtzentralitätsparameter

$$\delta_\mu := n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) \tag{18}$$

bezeichnet.

Bemerkungen

- q_ϕ kann zur Bestimmung kritischer Werte für einen erwünschten Testumfang genutzt werden.
- q_ϕ kann zur Bestimmung der Testpower genutzt werden.

Beweis

Die Testgütefunktion des betrachteten Tests im vorliegenden Testszenario ist definiert als

$$q_\phi : \mathbb{R}^m \to [0,1], \mu \mapsto q_\phi(\mu) := \mathbb{P}_\mu(\phi = 1) \tag{19}$$

Da die Wahrscheinlichkeiten für $\phi = 1$ und dafür, dass die zugehörige Teststatistik im Ablehnungsbereich des Tests liegt, gleich sind, benötigen wir also zunächst die Verteilung der Teststatistik. Wir haben oben aber bereits gesehen, dass

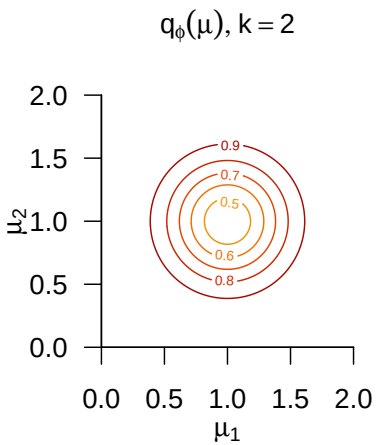
$$\frac{n-m}{m(n-1)} T^2 \sim f(\delta_\mu, m, n-m) \text{ mit } \delta_\mu := n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0) \tag{20}$$

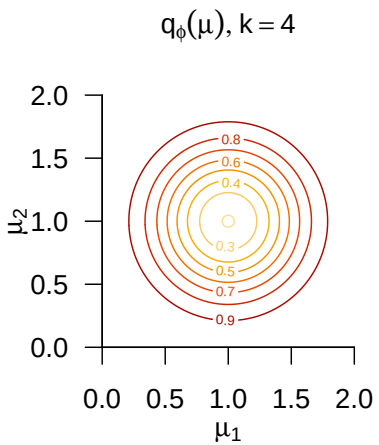
gilt. Der Ablehnungsbereich des betrachteten Tests ist $A :=]k, \infty[$. Also ergibt sich

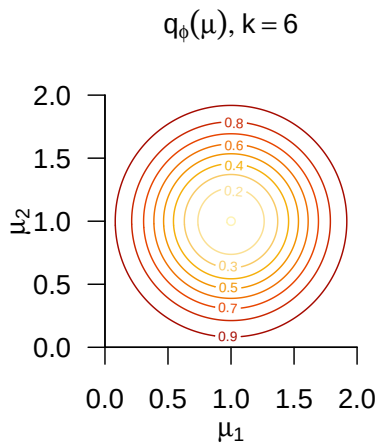
$$\begin{aligned} q_\phi(\mu) &= \mathbb{P}_\mu(\phi = 1) \\ &= \mathbb{P}_\mu\left(T^2 \in]k, \infty[\right) \\ &= \mathbb{P}_\mu\left(T^2 > k\right) \\ &= 1 - \mathbb{P}_\mu\left(T^2 \leq k\right) \\ &= 1 - F(\nu k; \delta_\mu, m, m-n). \end{aligned} \tag{21}$$

Beispiele

$$m := 2, n := 15, \Sigma := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}, \mu_0 := \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}$$

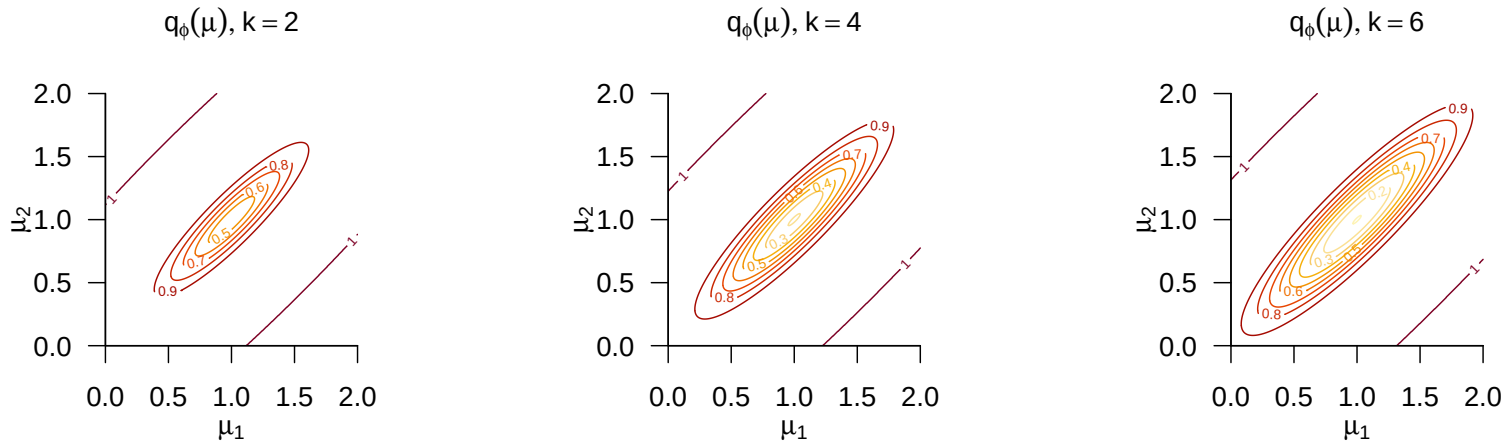






Beispiele

$$m := 2, n := 15, \Sigma := \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix}, \mu_0 := \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}$$



Definition (Level-α_0-Test, Signifikanzlevel α_0, Testumfang α)

q_ϕ sei die Testgütefunktion eines Tests ϕ und es sei $\alpha_0 \in [0, 1]$. Dann heißt ein Test ϕ, für den gilt, dass

$$q_\phi(\theta) \leq \alpha_0 \text{ für alle } \theta \in \Theta_0 \tag{22}$$

ein *Level-α_0-Test* und man sagt, dass der Test das *Signifikanzlevel* α_0 hat. Die Zahl

$$\alpha := \max_{\theta \in \Theta_0} q_\phi(\theta) \in [0, 1] \tag{23}$$

heißt der *Testumfang* von ϕ.

Bemerkungen

- α ist die größtmögliche Wahrscheinlichkeit für einen Typ I Fehler.
- Ein Test ist dann, und nur dann, ein Level-α_0-Test, wenn $\alpha \leq \alpha_0$ gilt.
- Bei einer einfachen Nullhypothese gilt für den Testumfang, dass $\alpha = q_\phi(\theta_0) = \mathbb{P}_{\theta_0}(\phi = 1)$.

Theorem (Testumfangkontrolle)

ϕ sei der im obigen Testszenario definierte Test. Dann ist ϕ ein Level-α_0-Test mit Testumfang α_0, wenn der kritische Wert definiert ist durch

$$k_{\alpha_0} := \nu^{-1} F^{-1}\left(1 - \alpha_0; m, n - m\right) \tag{24}$$

wobei $\nu := (n - m)/((n - 1)m)$ und $F^{-1}(\cdot; m, n - m)$ die inverse KVF der f-Verteilung mit Freiheitsgradparametern m und $n - m$ ist.

Beweis

Damit der betrachtete Test ein Level-α_0-Test ist, muss bekanntlich $q_\phi(\mu) \leq \alpha_0$ für alle $\mu \in \{\mu_0\}$, also hier $q_\phi(\mu_0) \leq \alpha_0$ gelten. Weiterhin ist der Testumfang des betrachteten Tests durch $\alpha = \max_{\mu \in \{\mu_0\}} q_\phi(\mu)$, also hier durch $\alpha = q_\phi(\mu_0)$ gegeben. Wir müssen also zeigen, dass die Wahl von k_{α_0} garantiert, dass ϕ ein Level-α_0-Test mit Testumfang α_0 ist. Dazu merken wird zunächst an, dass für $\mu = \mu_0$ gilt, dass

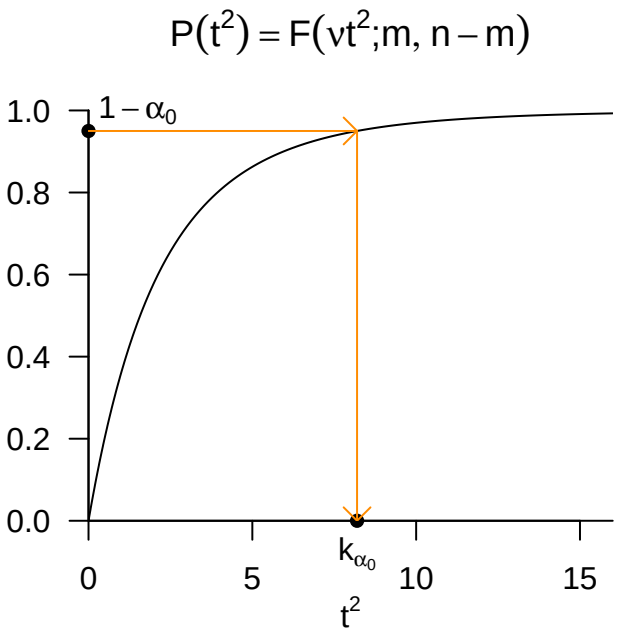
$$q_\phi(\mu_0) = 1 - F(\nu k; \delta_{\mu_0}, m, n-m) = 1 - F(\nu k; 0, m, n-m) = 1 - F(\nu k; m, n-m) \tag{25}$$

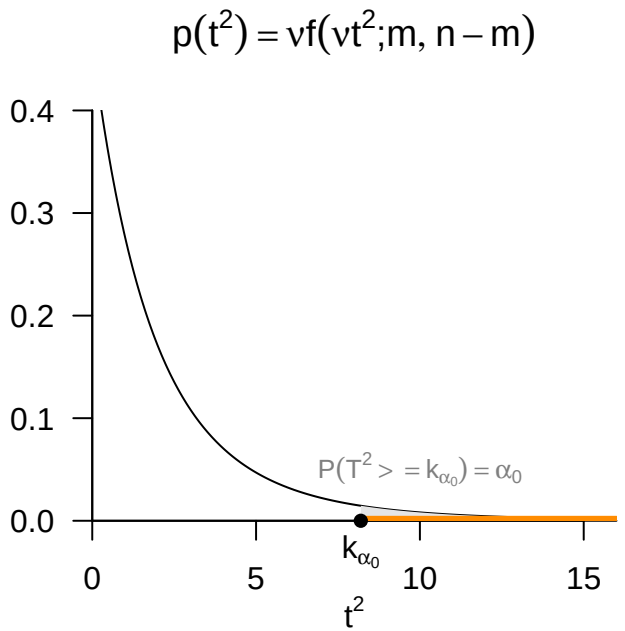
wobei$F(\nu k; \delta, m, n-m)$ und $F(\nu k; m, n-m)$ die KVF der nichtzentralen f-Verteilung mit Nichtzentralitätsparameter δ und Freiheitsgradparametern m und $n-m$ sowie der f-Verteilung mit Freiheitsgradparametern m und $n-m$, respektive, bezeichnen. Sei nun also $k := k_{\alpha_0}$. Dann gilt

$$\begin{aligned} q_\phi(\mu_0) &= 1 - F(\nu k_{\alpha_0}; m, n-m) \\ &= 1 - F\left(\nu \nu^{-1} F^{-1}(1-\alpha_0; m, n-m); m, n-m\right) \\ &= 1 - F\left(F^{-1}(1-\alpha_0; m, n-m); m, n-m\right) \\ &= 1 - (1-\alpha_0) = \alpha_0. \end{aligned} \tag{26}$$

Es folgt also direkt, dass bei der Wahl von $k = k_{\alpha_0}, q_\phi(\mu_0) \leq \alpha_0$ ist der betrachtete Test somit ein Level-α_0-Test ist. Weiterhin folgt direkt, dass der Testumfang des betrachteten Tests bei der Wahl von $k = k_{\alpha_0}$ gleich α_0 ist.

Wahl von $k_{\alpha_0} := \nu^{-1} F^{-1}\left(1 - \alpha_0; m, n - m\right)$ mit $m = 2, n = 15$ und $\alpha_0 := 0.05$





Praktisches Vorgehen

- Man nimmt an, dass ein vorliegender Datensatz $\Upsilon = \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix}$ eine Realisation von $\upsilon_1, ..., \upsilon_n \sim N(\mu, \Sigma)$ mit unbekannten Parametern $\mu \in \mathbb{R}^m$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd ist.
- Man möchte entscheiden ob für ein $\mu_0 \in \mathbb{R}^m$ eher $H_0 : \mu = \mu_0$ oder $H_1 : \mu \neq \mu_0$ zutrifft.
- Man wählt ein Signifikanzlevel α_0 und bestimmt den zugehörigen Freiheitsgradparameter-abhängigen kritischen Wert k_{α_0}. Zum Beispiel gilt bei Wahl von $\alpha_0 := 0.05, m = 2$ und $n = 15$, also Freiheitsgradparametern 2 und 13, dass $k_{0.05} = \nu^{-1} F^{-1}(1 - 0.05; 2, 13) \approx 8.2$.
- Anhand von $m, n, \mu_0, \bar{\upsilon}$ und C berechnet man die Realisierung der Einstichproben-T^2-Teststatistik

$$T^2 := n(\bar{\upsilon} - \mu_0)^T C^{-1} (\bar{\upsilon} - \mu_0) \tag{27}$$

- Wenn T^2 größer als k_{α_0} ist, lehnt man die Nullhypothese ab, andernfalls nicht.
- Die oben entwickelte Theorie garantiert dann, dass man in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt.

Theorem (p-Wert)

Für den p-Wert den Einstichproben-T^2-Test gilt

$$\text{p-Wert } = \mathbb{P}\left(T^2 \geq t^2\right) = 1 - F(\nu t^2; m, n-m). \tag{28}$$

Bemerkungen

Wir erinnern daran, dass per Definition der p-Wert das kleinste Signifikanzlevel α_0 ist, bei welchem man die Nullhypothese basierend auf einem vorliegenden Wert der Teststatistik ablehnen würde. Zum Beispiel ergeben sich

- Bei $m = 2$ und $n = 15$ der p-Wert für $t^2 = 7.00$ zu 0.071
- Bei $m = 2$ und $n = 15$ der p-Wert für $t^2 = 9.00$ zu 0.040
- Bei $m = 2$ und $n = 99$ der p-Wert für $t^2 = 7.00$ zu 0.035
- Bei $m = 4$ und $n = 15$ der p-Wert für $t^2 = 7.00$ zu 0.304

Beweis

Bei einem beobachteten Wert t^2 der Einstichproben-T^2-Teststatistik T^2 würde H_0 für jedes α_0 mit $t^2 \geq \nu^{-1} F^{-1}(1-\alpha_0; m, n-m)$ abgelehnt werden. Für diese α_0 gilt, wie unten gezeigt

$$\alpha_0 \geq \mathbb{P}\left(T^2 \geq t^2\right) \tag{29}$$

Das kleinste $\alpha_0 \in [0,1]$ mit $\alpha_0 \geq \mathbb{P}\left(T^2 \geq t^2\right)$ ist dann $\alpha_0 = \mathbb{P}(T^2 \geq t^2)$, also folgt

$$\text{p-Wert } = \mathbb{P}\left(T^2 \geq t^2\right) = 1 - F(\nu t^2; m, n-m). \tag{30}$$

Es bleibt zu zeigen, dass gilt

$$\begin{aligned} t^2 &\geq \nu^{-1} F^{-1}(1-\alpha_0; m, n-m) \\ \Leftrightarrow \nu t^2 &\geq F^{-1}(1-\alpha_0; m, n-m) \\ \Leftrightarrow \alpha_0 &\geq \mathbb{P}\left(T^2 \geq t^2\right). \end{aligned} \tag{31}$$

Dies aber folgt aus

$$\begin{aligned} t^2 &\geq \nu^{-1} F^{-1}(1-\alpha_0; m, n-m) \\ \nu t^2 &\geq F^{-1}(1-\alpha_0; m, n-m) \\ F(\nu t^2; m, n-m) &\geq F\left(F^{-1}(1-\alpha_0; m, n-m); m, n-m\right) \\ F(\nu t^2; m, n-m) &\geq 1-\alpha_0 \\ \mathbb{P}\left(T^2 \leq t^2\right) &\geq 1-\alpha_0 \\ \alpha_0 &\geq 1 - \mathbb{P}\left(T^2 \leq t^2\right). \end{aligned} \tag{32}$$

Wir betrachten die Testgütefunktion

$$q_\phi : \mathbb{R}^m \to [0,1], \mu \mapsto q_\phi(\mu) := 1 - F(\nu k; \delta_\mu, m, n-m) \tag{33}$$

bei kontrolliertem Testumfang, also für

$$k_{\alpha_0} := \nu^{-1} F^{-1}\left(1 - \alpha_0; m, n-m\right) \tag{34}$$

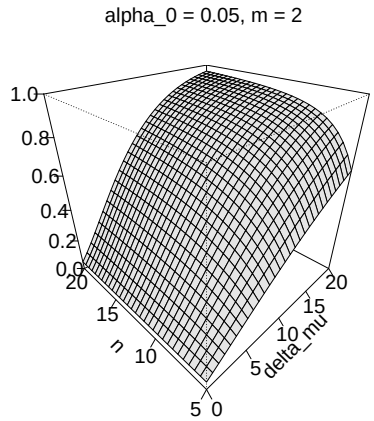
mit festem α_0 als Funktion des Nichtzentralitätsparameters und des Stichprobenumfangs. Namentlich hängt hier k_{α_0} auch von n ab.

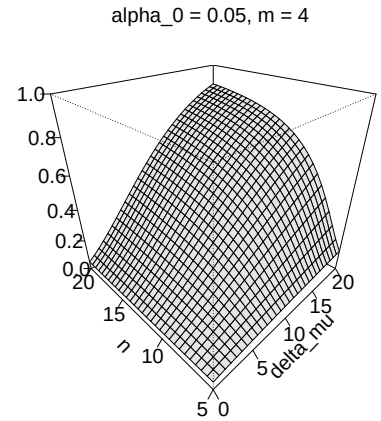
Es ergibt sich die bivariate reellwertige Funktion

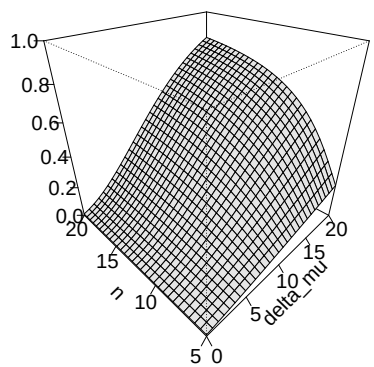
$$\pi : \mathbb{R} \times \mathbb{N} \to [0,1], (\delta_\mu, n) \mapsto \pi(\delta_\mu, n) := 1 - F(\nu k_{\alpha_0}; \delta_\mu, m, n-m) \tag{35}$$

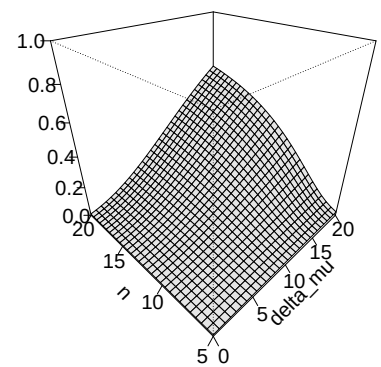
Bei festgelegtem α_0 hängt die Powerfunktion des Einstichproben-T^2-Tests also vom unbekannten Wert δ_μ, von der Datendimensionalität m und von der Stichprobengröße n ab. Wir evaluieren und visualisieren diese Abhängigkeiten untenstehend.

Modellevaluation | (5) Analyse der Powerfunktion









Praktisches Vorgehen

Mit größerem n steigt die Powerfunktion des Tests an

- Ein großer Stichprobenumfang ist besser als ein kleiner Stichprobenumfang.
- Kosten für die Erhöhung des Stichprobenumfangs werden aber nicht berücksichtigt.

$\Rightarrow$ Die Theorie statistischer Hypothesentests ist nicht besonders lebensnah.

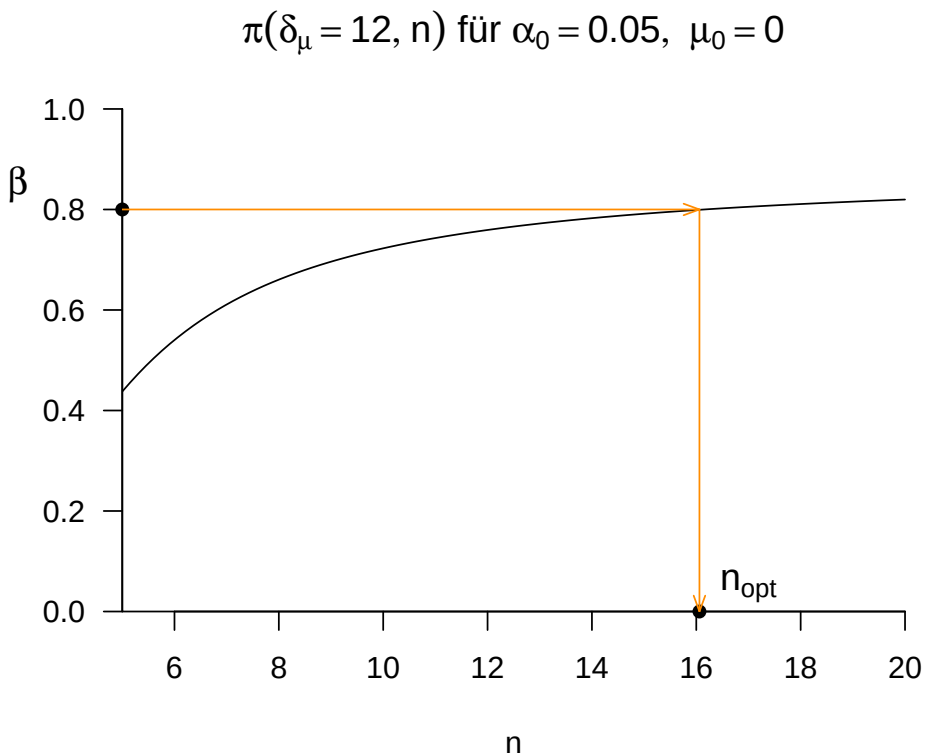
Die Powerfunktion hängt vom wahren, aber unbekannten, Parameterwert $\delta_\mu = n(\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0)$ ab.

$\Rightarrow$ Wenn man δ_μ schon kennen würde, würde man den Test nicht durchführen.

Generell wird folgendes Vorgehen favorisiert

- Man legt das Signifikanzlevel α_0 fest und evaluiert die Powerfunktion.
- Man wählt einen Mindestparameterwert δ_μ^*, den man mit $\pi(\delta_\mu, n) = \beta$ detektieren möchte.
- Ein konventioneller Wert ist $\beta = 0.8$.
- Man liest die für $\pi(\delta_\mu = \delta_\mu^*, n) = \beta$ nötige Stichprobengröße n ab.

Praktisches Vorgehen



Anwendungsszenario

Modellformulierung und Modellschätzung

Modellevaluation

Anwendungsbeispiel

Selbstkontrollfragen

Praktisches Vorgehen

```
D         = read.csv("./5_Daten/5_Einstichproben_T2_Tests.csv") # Datensatzeinlesen
Y         = rbind(D$y_1i, D$y_2i)                               # Datenmatrix
m         = nrow(Y)                                             # Dimensionalität der Zufallsvektoren/Daten
n         = ncol(Y)                                             # Anzahl der Datenpunkte
nu        = (n-m)/(m*(n-1))                                     # Parameter
mu_0      = matrix(c(30,3.5) , nrow = 2)                        # H0 Hypothesenparameter ("Normwert")
alpha_0   = 0.05                                                # Signifikanzlevel
k_alpha_0 = (1/nu)*qf(1-alpha_0,m,n-m)                          # kritischer Wert
j_n       = matrix(rep(1,n), nrow = n)                          # 1_n
I_n       = diag(n)                                             # I_n
J_n       = matrix(rep(1,n^2), nrow = n)                        # 1_{nn}
y_bar     = (1/n)*(Y %*% j_n)                                   # Stichprobenmittel
C         = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))          # Stichprobenkovarianzmatrix
T2        = n*t(y_bar - mu_0) %*% solve(C) %*% (y_bar - mu_0)   # Einstichproben-T^2-Statistik
if(T2 > k_alpha_0){phi = 1} else {phi = 0}                      # Test 1_{T^2 >= k_alpha_0}
p         = 1 - pf(nu*T2,m,n-m)                                 # p-Wert
cat("Y_bar   = ", y_bar,                                        # Ausgabe
    "\nC       = ", C,
    "\nT^2     = ", T2,
    "\nalpha_0 = ", alpha_0,
    "\nk       = ", k_alpha_0,
    "\nphi     = ", phi,
    "\np       = ", p)
```

```
Y_bar   =  26.25615 2.991039
C       =  38.8981 3.549813 3.549813 1.972143
T^2     =  7.546368
alpha_0 =  0.05
k       =  7.504065
phi     =  1
p       =  0.04928746
```

Black-Box-Verfahren

```
library(MVTests)                                                    # R Paket
D         = read.csv("./5_Daten/5_Einstichproben_T2_Tests.csv")     # Datensatzeinlesen
Y         = rbind(D$y_1i, D$y_2i)                                   # Datenmatrix
mu_0      = matrix(c(30,3.5) , nrow = 2)                            # H0 Hypothesenparameter ("Normwert")
alpha_0   = 0.05                                                    # Signifikanzlevel
phi       = OneSampleHT2(t(Y), mu_0, alpha_0)                       # Einstichproben-T^2-Test
cat("Y_bar   = "  , phi$Descriptive[2,],                           # Ausgabe
    "\nT^2     = ", phi$HT2,
    "\nalpha_0 = ", phi$alpha,
    "\np       = ", phi$p.value)
```

```
Y_bar   =  26.25615 2.991039
T^2     =  7.546368
alpha_0 =  0.05
p       =  0.04928746
```

Anwendungsszenario

Modellformulierung und Modellschätzung

Modellevaluation

Anwendungsbeispiel

Selbstkontrollfragen

Selbstkontrollfragen

1. Beschreiben Sie das Anwendungsszenario für einen Einstichproben-T^2-Test.
2. Geben Sie die Definition des Einstichproben-T^2-Test Modells wieder
3. Geben Sie die Definition der Einstichproben-T^2-Teststatistik wieder.
4. Erläutern Sie, wann die Einstichproben-T^2-Teststatistik hohe Werte annimmt.
5. Geben Sie das Theorem zu WDF und KDF der Einstichproben-T^2-Teststatistik wieder.
6. Geben Sie das Theorem zur Testumfangkontrolle eins Einstichproben-T^2-Tests wieder.
7. Erläutern Sie das praktische Vorgehen bei der Durchführung eines Einstichproben-T^2-Tests.
8. Geben Sie das Theorem zum p-Wert eines Einstichproben-T^2-Test an und erläutern Sie die Komponenten des entsprechenden Ausdrucks.

Referenzen

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley-Interscience.

Hotelling, Harold. 1931. "The Generalization of Student's Ratio." *The Annals of Mathematical Statistics* 2 (3): 360–78. https://doi.org/10.1214/aoms/1177732979.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(6) Einfaktorielle Varianzanalyse

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Anwendungsszenario und Datendeskription

Anwendungsszenario

- Zwei oder mehr Gruppen experimenteller Einheiten mit Datendimension $m > 1$
- Annahme der unabhängigen und identischen Normalverteilung $N(\mu_i, \Sigma)$ der Daten
- $\mu_i \in \mathbb{R}^m, i = 1, ..., p$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd unbekannt
- Absicht des inferentiellen Testens der Nullhypothese identischer Gruppenerwartungswerte

Generalisierung des Zweistichproben-T^2-Tests bei unabhängigen Stichproben mit einfacher Nullhypothese für mehr als zwei Stichproben

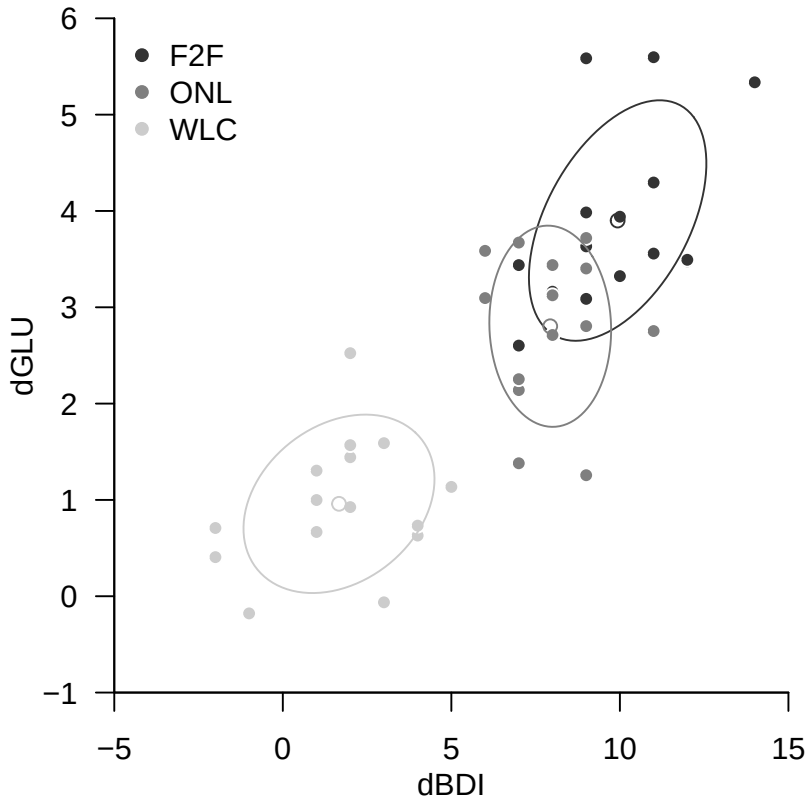
Anwendungsbeispiel

- Analyse von Daten dreier Therapiegruppen (F2F, ONL, WLC) von `dBDI` und `dGLU` Daten
- Testen der Nullhypothese $\mu_1 = \mu_2 = \mu_3$

Table 1: Prä-Post-Interventions-BDI-II-Score und -Glukokortikoidplasmalevel Differenzenwerte von drei Studiengruppen (F2F: Face-to-face, ONL: online, WLC: waitlist control) jeweils 15 Patient:innen. Die Tabelle zeigt exemplarisch die ersten fünf Datenpunkte jeder Gruppe.

	COND	dBDI	dGLU
1	F2F	11	4.3
2	F2F	10	3.9
3	F2F	12	3.5
4	F2F	7	2.6
5	F2F	10	3.3
16	ONL	6	3.1
17	ONL	8	2.7
18	ONL	7	2.1
19	ONL	8	3.1
20	ONL	11	2.8
31	WLC	-2	0.7
32	WLC	2	1.4
33	WLC	1	1.0
34	WLC	2	0.9
35	WLC	3	1.6

Anwendungsbeispiel



Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Definition (Modell der einfaktoriellen multivariaten Varianzanalyse)

Für $i = 1, ..., p$ und $j = 1, ..., n_i$ seien υ_{ij} m-dimensionale Zufallsvektoren, die die $n := \sum_{i=1}^{p} n_i$ m-dimensionalen Datenpunkte eines einfaktoriellen multivariaten Varianzanalyseszenarios modellieren. Dann hat das Modell der einfaktoriellen multivariaten Varianzanalyse die strukturelle Form

$$\upsilon_{ij} = \mu_i + \varepsilon_{ij} \text{ mit } \varepsilon_{ij} \sim N(0_m, \Sigma) \text{ u.i.v. mit } \mu_i \in \mathbb{R}^m \text{ und } \Sigma \in \mathbb{R}^{m \times m} \text{ pd} \tag{1}$$

und die Datenverteilungsform

$$\upsilon_{ij} \sim N(\mu_i, \Sigma) \text{ u.v. mit } \mu_i \in \mathbb{R}^m \text{ und } \Sigma \in \mathbb{R}^{m \times m} \text{ pd.} \tag{2}$$

Bemerkungen

- Der Einfachheit halber setzen wir meist identische Gruppengrößen $k := n_i$ voraus.
- Die Gesamtheit aller Datenzufallsvektoren bezeichnen wir mit Υ.

Theorem (Parameterschätzer)

Gegeben sei das Modell der einfaktoriellen multivariaten Varianzanalyse. Dann ist für $i = 1, \ldots, p$

$$\hat{\mu}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} \upsilon_{ij} \tag{3}$$

ein unverzerrte Schätzer des gruppenspezifischen Erwartungswertparameters μ_i *und*

$$\hat{\Sigma} := \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(\upsilon_{ij} - \hat{\mu}_i\right) \left(\upsilon_{ij} - \hat{\mu}_i\right)^T \tag{4}$$

ein unverzerrter Schätzer des Kovarianzmatrixparameters Σ.

Bemerkungen

- $\hat{\mu}_i$ ist das Stichprobenmittel der $i$ten Gruppe.
- $\hat{\Sigma}$ ist die mit $1/(n-p)$ skalierte Within-Group Sum-of-Squares Matrix (siehe unten).
- Anstelle eines Beweis validieren wir die Aussage des Theorems mithilfe einer Simulation.

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Modellevaluation mit der Wilks'-Λ-Statistik

Überblick

- Ziel einer einfaktoriellen multivariaten Varianzanalyse ist meist das Testen der Nullhypothese

$$H_0 : \mu_1 = \cdots = \mu_p. \tag{5}$$

- Es ergibt sich damit die Alternativhypothese

$$H_1 : \mu_{i_l} \neq \mu_{j_l} \text{ für mindestens ein Paar } i, j \text{ mit } i \neq j, 1 \leq i, j \leq p \text{ und mindestens ein } l \text{ mit } 1 \leq l \leq m. \tag{6}$$

- Zur Evaluation von H_0 wurden eine Reihe von Teststatistiken vorgeschlagen.
- Alle Teststatistiken beruhen auf der Kreuzproduktsummenmatrixzerlegung.
- Wir betrachten hier exemplarisch die Wilks'-Λ-Statistik.
- Die Verteilungen der Wilks'-Λ-Statistik unter der Nullhypothese sind zum Teil analytisch angebbar, zum Teil müssen sie approximiert werden und haben die Form von f-Verteilungen.

Theorem (Kreuzproduktsummenmatrizenzerlegung)

Für $i = 1, ..., p$ und $j = 1, ..., n_i$ bezeichne v_{ij} den $j$ten Stichprobenvektor der $i$ten Stichprobengruppe eines einfaktoriellen multivariaten Varianzanalysemodells. Weiterhin seien

$$\bar{v} := \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} v_{ij} \text{ und } \bar{v}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} v_{ij} \tag{7}$$

das *Gesamtstichprobenmittel* und das i*te Gruppenstichprobenmittel*, respektive. Schließlich seien

$T := \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(v_{ij} - \bar{v}\right) \left(v_{ij} - \bar{v}\right)^T$ die *Totale Sum-of-Squares Matrix*

$B := \sum_{i=1}^{p} n_i \left(\bar{v}_i - \bar{v}\right) \left(\bar{v}_i - \bar{v}\right)^T$ die *Between-Group Sum-of-Squares Matrix*

$W := \sum_{i=1}^{p} \sum_{j=1}^{n_i} \left(v_{ij} - \bar{v}_i\right) \left(v_{ij} - \bar{v}_i\right)^T$ die *Within-Group Sum-of-Squares Matrix*.

Dann gilt

$$T = B + W. \tag{8}$$

Modellevaluation mit der Wilks'-Λ-Statistik

Bemerkungen

$T \in \mathbb{R}^{m \times m}$ misst die totale Variabilität der Datenvektoren um das Gesamtstichprobenmittel.

$B \in \mathbb{R}^{m \times m}$ misst die Variabilität der Gruppenstichprobenmittel um das Gesamtstichprobenmittel.

$W \in \mathbb{R}^{m \times m}$ misst die Variabilität der Datenvektoren um ihre jeweiligen Gruppenstichprobenmittel.

Die totale Variabilität wird hier also in zwei unabhängige Beiträge von Variabilität zerlegt.

W heißt auch *Residualvariabiliät*, weil sie die verbleibende Variabilität nach Schätzung der Gruppenerwartungswertparameter quantifiziert und gilt das

$$W = (n - p)\hat{\Sigma}. \tag{9}$$

Ein Beweis gibt sich durch algebraische Umformung.

Modellevaluation mit der Wilks'-Λ-Statistik

Beweis

$$
\begin{aligned}
T &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(v_{ij}-\bar{v})(v_{ij}-\bar{v})^T \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i+\bar{v}_i-\bar{v})(v_{ij}-\bar{v}_i+\bar{v}_i-\bar{v})^T \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\big((v_{ij}-\bar{v}_i)+(\bar{v}_i-\bar{v})\big)\big((v_{ij}-\bar{v}_i)+(\bar{v}_i-\bar{v})\big)^T \\
&= \sum_{i=1}^{p}\sum_{j=1}^{n_i}\big((v_{ij}-\bar{v}_i)(v_{ij}-\bar{v}_i)^T+2(v_{ij}-\bar{v}_i)(\bar{v}_i-\bar{v})^T+(\bar{v}_i-\bar{v})(\bar{v}_i-\bar{v})^T\big) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i)(v_{ij}-\bar{v}_i)^T+\sum_{j=1}^{n_i}2(v_{ij}-\bar{v}_i)(\bar{v}_i-\bar{v})^T+\sum_{j=1}^{n_i}(\bar{v}_i-\bar{v})(\bar{v}_i-\bar{v})^T\right) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i)(v_{ij}-\bar{v}_i)^T+2\left(\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i)\right)(\bar{v}_i-\bar{v})^T+n_i(\bar{v}_i-\bar{v})(\bar{v}_i-\bar{v})^T\right) \qquad (10) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i)(v_{ij}-\bar{v}_i)^T+2\left(\sum_{j=1}^{n_i}\left(v_{ij}-\frac{1}{n_i}\sum_{j=1}^{n_i}v_{ij}\right)\right)(\bar{v}_i-\bar{v})^T+n_i(\bar{v}_i-\bar{v})(\bar{v}_i-\bar{v})^T\right) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i)(v_{ij}-\bar{v}_i)^T+2\left(\sum_{j=1}^{n_i}v_{ij}-\sum_{j=1}^{n_i}v_{ij}\right)(\bar{v}_i-\bar{v})^T+n_i(\bar{v}_i-\bar{v})(\bar{v}_i-\bar{v})^T\right) \\
&= \sum_{i=1}^{p}\left(\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i)(v_{ij}-\bar{v}_i)^T+n_i(\bar{v}_i-\bar{v})(\bar{v}_i-\bar{v})^T\right) \\
&= \sum_{i=1}^{p}n_i(\bar{v}_i-\bar{v})(\bar{v}_i-\bar{v})^T+\sum_{i=1}^{p}\sum_{j=1}^{n_i}(v_{ij}-\bar{v}_i)(v_{ij}-\bar{v}_i)^T \\
&= B+W.
\end{aligned}
$$

Definition (Wilks'-Λ-Statistik)

Es seien das Modell der einfaktoriellen Varianzanalyse sowie die Between Sum of Squares Matrix B und die Within Sum-of-Squares Matrix W definiert wie oben. Dann ist die Wilks'-Λ-Statistik Teststatistik definiert als

$$\Lambda := \frac{|W|}{|B+W|}, \tag{11}$$

wobei $|\cdot|$ die Determinante bezeichnet.

Bemerkungen

- Intuitiv misst Λ das Verhältnis von Residualvariabilität und Gesamtvariabilität.
- Ohne Beweis halten wir fest, dass $\Lambda \in [0, 1]$
- Für $\bar{v}_1 = \cdots = \bar{v}_p = \bar{v}$ gilt $B = 0_{mm}$ und damit $\Lambda = 1$.
- Für steigende Unterschiede zwischen den $\bar{v}_i$ nimmt $|B+W|$ gegenüber $|W|$ zu, Λ also ab.
- Kleine Werte von Λ sprechen also für eine Abweichung von der Nullhypothese.

Theorem (Spezielle H_0 Verteilungen von Λ Transformationen)

Es seien das Modell der einfaktoriellen Varianzanalyse und Wilks'-Λ-Statistik definiert wie oben und es gelte außerdem

$$H_0 : \mu_1 = \cdots = \mu_p. \tag{12}$$

Dann sind für die in den ersten beiden Tabellenspalten aufgeführten Spezialfällen die in der dritten Tabellenspalte aufgeführten Statistiken f-Zufallsvariablen und zwar mit den in der vierten Tabellenspalte aufgeführten Parametern.

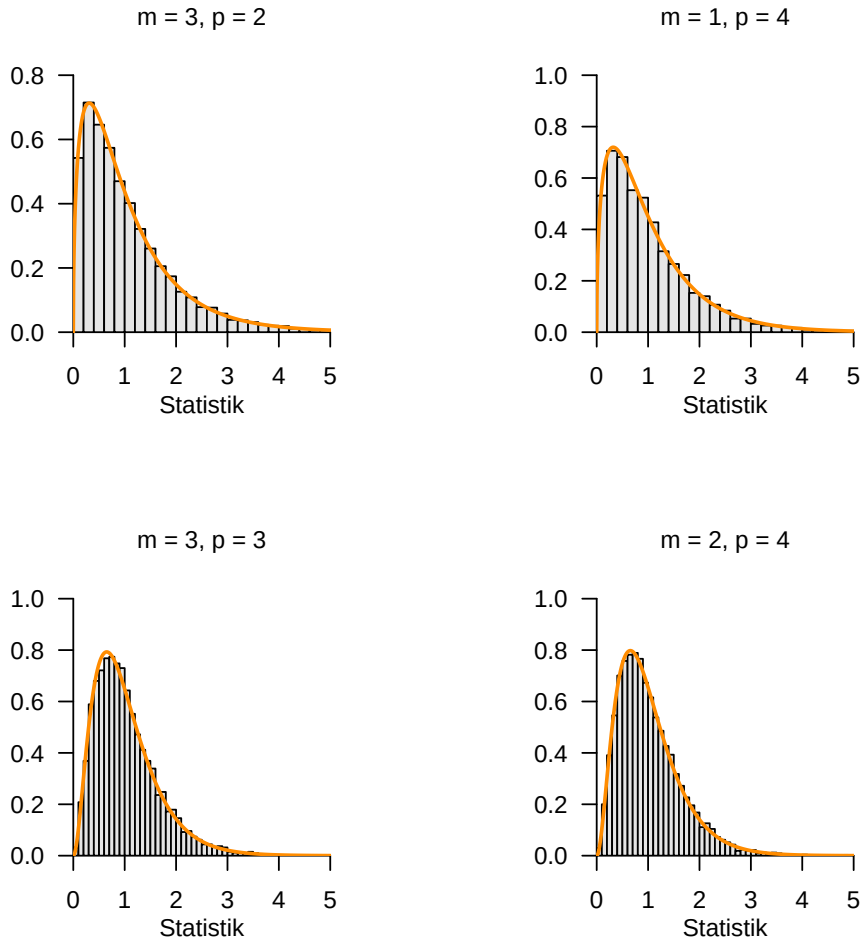
Datendimension m	Gruppenanzahl p	Statistik	f-Verteilungsparameter
Beliebig	2	$\frac{1-\Lambda}{\Lambda}\,\frac{n-p-m+1}{m}$	$m, n-p-m+1$
Beliebig	3	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\,\frac{n-p-m+1}{m}$	$2m, 2(n-p-m+1)$
1	Beliebig	$\frac{1-\Lambda}{\Lambda}\,\frac{n-p}{p-1}$	$p-1, n-p$
2	Beliebig	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\,\frac{n-p-1}{p-1}$	$2(p-1), 2(n-p-1)$

Bemerkungen

- Die Verteilungen gehen zurück auf Wilks (1932).

Modellevaluation mit der Wilks'-Λ-Statistik

Simulation spezieller H_0 Verteilungen von Wilks'-Λ-Statistik Transformationen



Theorem (Approximative H_0 Verteilungen von Λ Transformationen)

Es seien das Modell der einfaktoriellen Varianzanalyse und Wilks'-Λ-Statistik definiert wie oben und es gelte außerdem

$$H_0 : \mu_1 = \cdots = \mu_p. \tag{13}$$

Dann ist die Statistik

$$\tau := \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{\nu_2}{\nu_1} \tag{14}$$

mit

$$\nu_1 := m(p-1) \text{ und } \nu_2 := wt - \frac{1}{2}(m(p-1) - 2) \tag{15}$$

sowie

$$w := n - 1 - \frac{1}{2}(m + k) \text{ und } t := \sqrt{\frac{m^2(p-1)^2 - 4}{m^2 + (p-1)^2 - 5}} \tag{16}$$

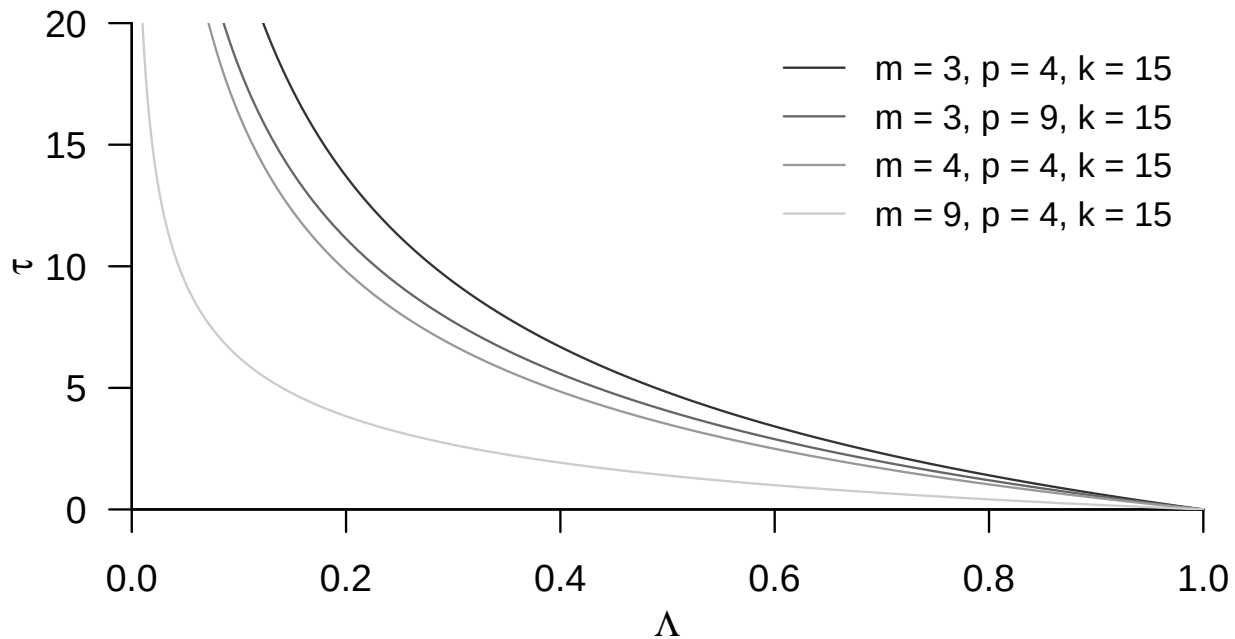
approximativ f-verteilt mit Freiheitsgradparametern ν_1 und ν_2.

Bemerkungen

- Die Approximation geht zurück auf Rao (1951).

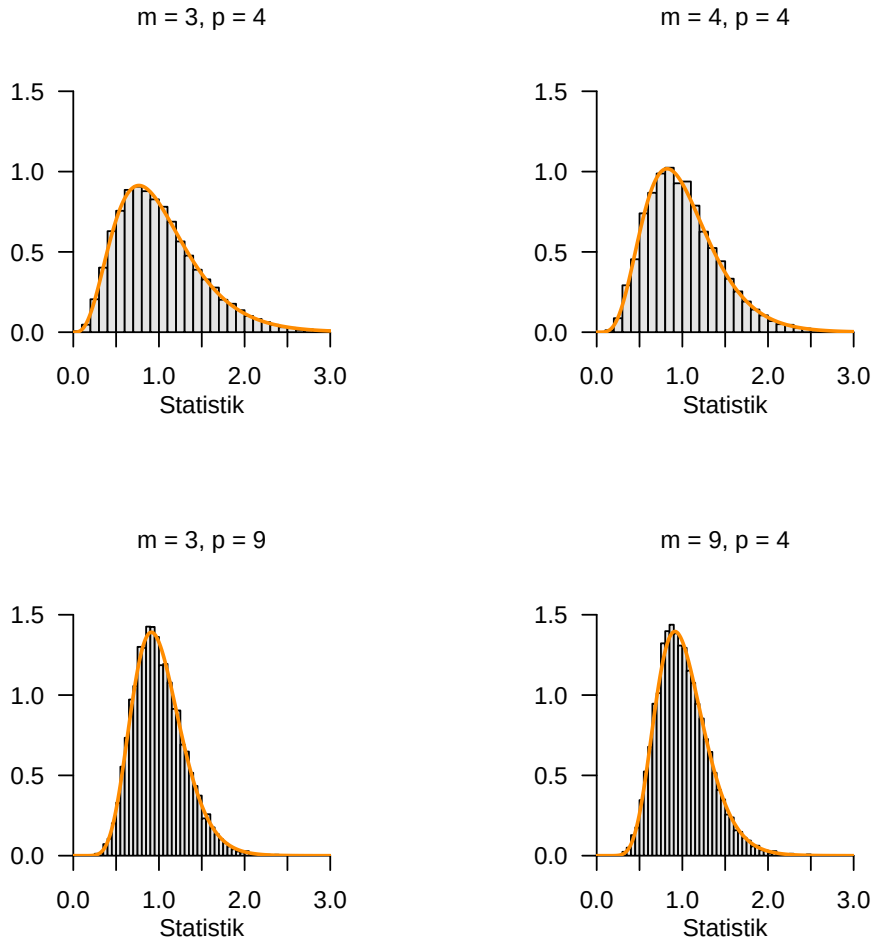
Modellevaluation mit der Wilks'-Λ-Statistik

τ als Funktion von Λ : $\Lambda \downarrow \Rightarrow \tau \uparrow$



Modellevaluation mit der Wilks'-Λ-Statistik

Simulation approximativer H_0 Verteilungen von Wilks'-Λ-Statistik Transformationen



Theorem (Wilks'-Λ-Statistik-basierter Test)

Es seien das Modell der einfaktoriellen Varianzanalyse und die Wilks'-Λ-Statistik basierte Teststatistik τ mit Verteilungsparametern ν_1, ν_2 wie oben definiert. Weiterhin sei der kritische Wert-basierte Test

$$\phi(\Upsilon) := 1_{\{\tau > k\}} := \begin{cases} 1 & \tau > k \\ 0 & \tau \leq k \end{cases} \tag{17}$$

definiert. ϕ ist genau dann ein Level-α_0-Test mit Testumfang α, wenn

$$k := k_{\alpha_0} := F^{-1}(1 - \alpha_0; \nu_1, \nu_2) \tag{18}$$

ist und der p-Wert einer realisiertern τ-Teststatistik $\tilde{\tau}$ ergibt sich zu

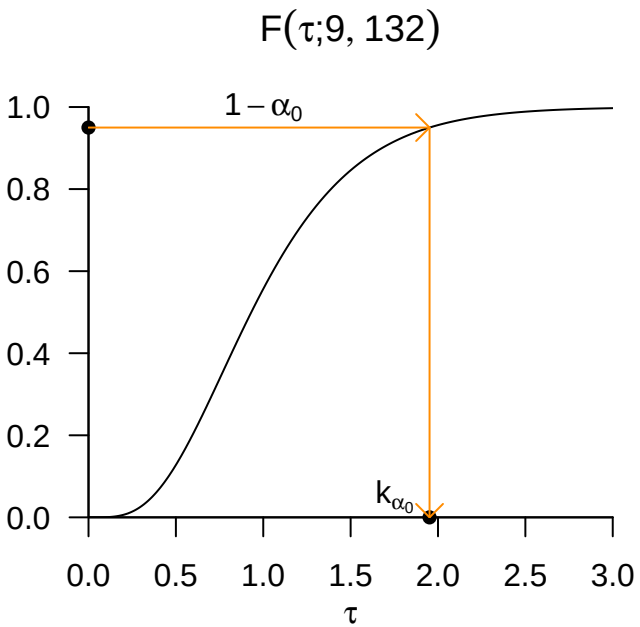
$$\text{p-Wert} = \mathbb{P}(\tau \geq \tilde{\tau}) = 1 - F(\tilde{\tau}; \nu_1, \nu_2) \tag{19}$$

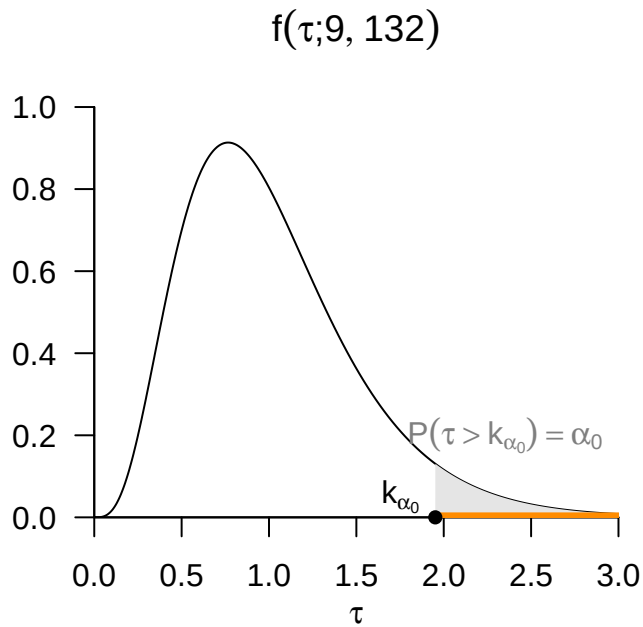
Bemerkungen

- Ein Beweis kann in Analogie zum Einstichproben-T^2-Test Fall geführt werden.
- Wir validieren die Testumfangkontrolle mithilfe von k_{α_0} in untenstehender Simulation.

Testumfangkontrolle

Wahl von k_{α_0} bei $m = 3, p = 4, k = 15 \Rightarrow \nu_1 = 9, \nu_2 = 132$ und $\alpha_0 = 0.05$.





Praktisches Vorgehen

- Man nimmt an, dass ein vorliegender Datensatz von p Gruppen von m-dimensionalen Datenvektoren für jeweils $j = 1, \ldots, n_i$ Realisationen von $\upsilon_{ij} \sim N(\mu_i, \Sigma)$ mit unbekannten Parametern $\mu_i \in \mathbb{R}^m, i = 1, \ldots, p$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd sind.
- Man möchte entscheiden ob $H_0 : \mu_1 = \cdots = \mu_p$ eher zutrifft oder eher nicht.
- Man wählt ein Signifikanzniveau α_0 und bestimmt den zugehörigen Freiheitsgradparameter-abhängigen kritischen Wert k_{α_0}. Zum Beispiel gilt bei Wahl von $\alpha_0 := 0.05, m = 3, p = 4, k = 15$ für $i = 1, \ldots p$ und somit $n = 60$ sowie $\nu_1 = 9, \nu_2 = 132$, dass $k_{\alpha_0} = F^{-1}(1 - 0.05; 9, 132) \approx 1.95$ ist.
- Anhand der Wilk's Λ Statistik sowie m, p und k berechnet man man den realisierten Wert der τ-Teststatistik, den wir hier mit $\tilde{\tau}$ bezeichnen.
- Wenn $\tilde{\tau}$ größer als k_{α_0} ist, lehnt man die Nullhypothese ab, andernfalls nicht.
- Die oben entwickelte Theorie garantiert dann, dass man im Mittel in höchstens $\alpha_0 \cdot 100$ von 100 Fällen die Nullhypothese fälschlicherweise ablehnt.
- Schließlich ergibt sich der assoziierte p-Wert der realisiertern τ-Teststatistik $\tilde{\tau}$ zu

$$\text{p-Wert} = \mathbb{P}(\tau \geq \tilde{\tau}) = 1 - F(\tilde{\tau}; \nu_1, \nu_2) \tag{20}$$

Anwendungsszenario und Datendeskription

Modellformulierung und Modellschätzung

Modellevaluation mit der Wilks'-Λ-Statistik

Selbstkontrollfragen

Selbstkontrolfragen

1. Erläutern Sie das Anwendungsszenario einer einfaktoriellen multivariaten Varianzanalyse.
2. Geben Sie die Definition des Modells der einfaktoriellen multivariaten Varianzanalyse wieder.
3. Geben Sie das Theorem zu den Parameterschätzern der einfaktoriellen multivariaten Varianzanalyse wieder.
4. Erläutern Sie die Null- und Alternativhypothesen einer einfaktoriellen multivariaten Varianzanalyse.
5. Geben Sie das Theorem zur Kreuzproduktsummenmatrizenzerlegung wieder.
6. Was messen die Totale, Between-Group und die Within-Group Sum-of-Squares Matrizen, respektive?
7. Geben Sie die Definition der Wilks'-Λ-Statistik wieder.
8. Erläutern Sie Gemeinsamkeiten und Unterschiede zwischen speziellen und approximativen H_0 Verteilungen von Wilks-Λ-Transformationen bei der einfaktoriellen multivariaten Varianzanalyse.
9. Geben Sie das Theorem zum Wilks-Λ-Statistik-basierten Test im Rahmen der einfaktoriellen multivariaten Varianzanalyse wieder.
10. Erläutern Sie das praktische Vorgehen zur Durchführung eines Wilks-Λ-Statistik-basierten Tes tim Rahmen der einfaktoriellen multivariaten Varianzanalyse.

Referenzen

Rao, C. Radhakrishna. 1951. "An Asymptotic Expansion of the Distribution of Wilk's Criterion." *Bulletin of the International Statistical Institute* 33 (2): 177–80.

Wilks, S. S. 1932. "Certain Generalizations in the Analysis of Variance." *Biometrika* 24 (3-4): 471–94. https://doi.org/10.1093/biomet/24.3-4.471.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(7) Kanonische Korrelationsanalyse

Datenanalyseszenarien

UV	AV	Datenanalysemethoden
Univariat	Univariat	Korrelation, Einfache Regression,T-Tests
Multivariat	Univariat	Multiple Korrelation, Multiple Regression, Allgemeines Lineares Modell
Univariat	Multivariat	Einstichproben-T^2-Tests, Einfaktorielle Varianzanalyse
Multivariat	Multivariat	Kanonische Korrelation, Multivariates Allgemeines Lineares Modell

Datenanalyseszenarien

UV	AV
x_1	y_1
x_{11}	y_{11}
x_{12}	y_{12}
x_{13}	y_{13}
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_{1n}	y_{1n}

Korrelation
Einfache Regression
T-Tests

UV			AV
x_1	$\cdots$	x_m	y_1
x_{11}	$\cdots$	x_{m1}	y_{11}
x_{12}	$\cdots$	x_{m2}	y_{12}
x_{13}	$\cdots$	x_{m3}	y_{13}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	x_{mn}	y_{1n}

Multiple Korrelation
Multiple Regression
Allgemeines Lineares Modell

Datenanalyseszenarien

UV	AV		
x_1	y_1	$\cdots$	y_m
x_{11}	y_{12}	$\cdots$	y_{m1}
x_{12}	y_{13}	$\cdots$	y_{m2}
x_{13}	y_{14}	$\cdots$	y_{m3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	y_{1n}	$\cdots$	y_{mn}

Einstichproben-T^2-Tests

Einfaktorielle Varianzanalyse

UV			AV		
x_1	$\cdots$	x_{m_x}	y_1	$\cdots$	y_{m_y}
x_{11}	$\cdots$	x_{m_x1}	y_{11}	$\cdots$	y_{m_y1}
x_{12}	$\cdots$	x_{m_x2}	y_{12}	$\cdots$	y_{m_y2}
x_{13}	$\cdots$	x_{m_x3}	y_{13}	$\cdots$	y_{m_y3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	x_{m_xn}	y_{1n}	$\cdots$	y_{m_yn}

Kanonische Korrelationsanalyse

Multivariates Allgemeines Lineares Modell

Korrelation

Modellformulierung

Modellschätzung

Selbstkontrollfragen

Korrelation

Modellformulierung

Modellschätzung

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie

Mehr Therapiestunden

⇒ Höhere Wirksamkeit?

Unabhängige Variable

- Anzahl Therapiestunden

Abhängige Variable

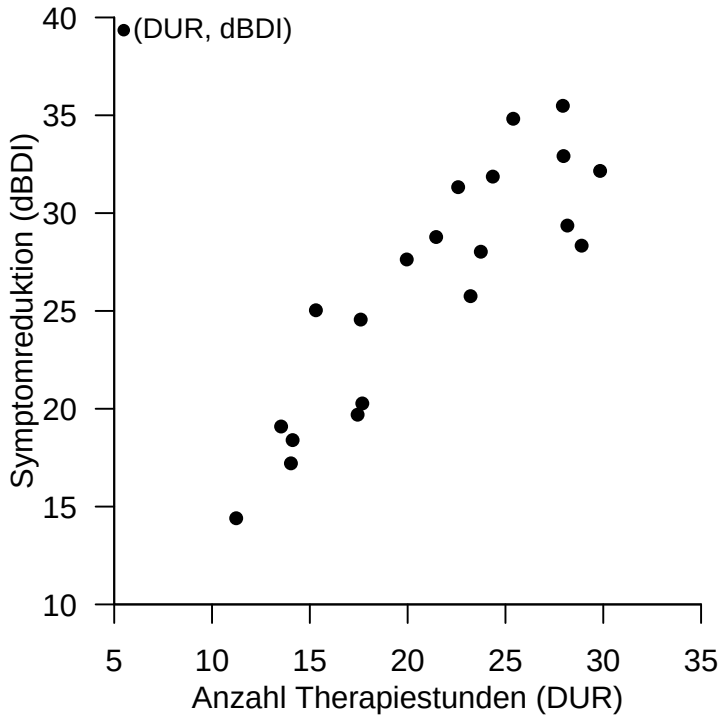
- BDI Score Reduktion

Beispieldatensatz

$i = 1, \ldots, 20$ Patient:innen, `dBDI` Symptomreduktion bei Patient:in i, `DUR` Anzahl Therapiestunden von Patient:in i

DUR	dBDI
27.9	35.5
15.3	25.0
17.4	19.7
21.5	28.8
28.2	29.4
14.0	17.2
28.0	32.9
28.9	28.3
23.2	25.8
22.6	31.3
11.2	14.4
14.1	18.4
13.5	19.1
23.7	28.0
17.7	20.3
25.4	34.8
20.0	27.6
24.4	31.9
29.8	32.2
17.6	24.6

Beispieldatensatz



Wie stark hängen Anzahl Therapiestunden und Symptomreduktion zusammen?

Definition (Korrelation)

Die *Korrelation* zweier Zufallsvariablen ξ und υ ist definiert als

$$\rho(\xi, \upsilon) := \frac{\mathbb{C}(\xi, \upsilon)}{\sqrt{\mathbb{V}(\xi)}\sqrt{\mathbb{V}(\upsilon)}}, \tag{1}$$

wobei $\mathbb{C}(\xi, \upsilon)$ die Kovarianz von ξ und υ und $\mathbb{V}(\xi)$ und $\mathbb{V}(\upsilon)$ die Varianzen von ξ und υ bezeichnen.

Für eine Einführung zur Korrelation siehe die entsprechenden BSc Lehreinheiten

- Erwartungswert, Varianz, Kovarianz
- Korrelation

Bemerkungen

- $\rho(\xi, \upsilon)$ wird auch *Korrelationskoeffizient* von ξ und υ genannt.
- Wir haben bereits gesehen, dass $-1 \leq \rho(\xi, \upsilon) \leq 1$ gilt.
- Wenn $\rho(\xi, \upsilon) = 0$ ist, werden ξ und υ *unkorreliert* genannt.
- Aus der Unabhängigkeit von ξ und υ folgt $\rho(\xi, \upsilon) = 0$.
- Aus $\rho(\xi, \upsilon) = 0$ folgt die Unabhängigkeit von ξ und υ im Allgemeinen nicht.

Definition (Stichprobenkorrelation)

$(x_1, y_1), ..., (x_n, y_n)$ seien unabhängigie Realisierungen eines Zufallsvektors (ξ, υ). Weiterhin seien:

- Die Stichprobenmittel der x_i und y_i definiert als

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \text{ und } \bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{2}$$

- Die Stichprobenstandardabweichungen x_i und y_i definiert als

$$s_x := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ und } s_y := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{3}$$

- Die Stichprobenkovarianz der $(x_1, y_1), ..., (x_n, y_n)$ definiert als

$$c_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{4}$$

Dann ist die *Stichprobenkorrelation* der $(x_1, y_1), ..., (x_n, y_n)$ definiert als

$$r_{xy} := \frac{c_{xy}}{s_x s_y} \tag{5}$$

und wird auch *Stichprobenkorrelationskoeffizient* genannt.

Beispiel

```
# Laden des Beispieldatensatzes
fname = "./7_Daten/7_Kanonische_Korrelationsanalyse.csv"                    # Dateipfad
D     = read.csv("./7_Daten/7_Kanonische_Korrelationsanalyse.csv")          # Laden als Dataframe
x_i   = D$DUR                                                               # x_i Werte
y_i   = D$dBDI                                                              # y_i Werte
n     = length(x_i)                                                         # n

# "Manuelle" Berechnung der Stichprobenkorrelation
x_bar = (1/n)*sum(x_i)                                                      # \bar{x}
y_bar = (1/n)*sum(y_i)                                                      # \bar{y}
s_x   = sqrt(1/(n-1)*sum((x_i - x_bar)^2))                                  # s_x
s_y   = sqrt(1/(n-1)*sum((y_i - y_bar)^2))                                  # s_y
c_xy  = 1/(n-1) * sum((x_i - x_bar) * (y_i - y_bar))                        # c_{xy}
r_xy  = c_xy/(s_x * s_y)                                                    # r_{xy}
print(r_xy)                                                                 # Ausgabe
```

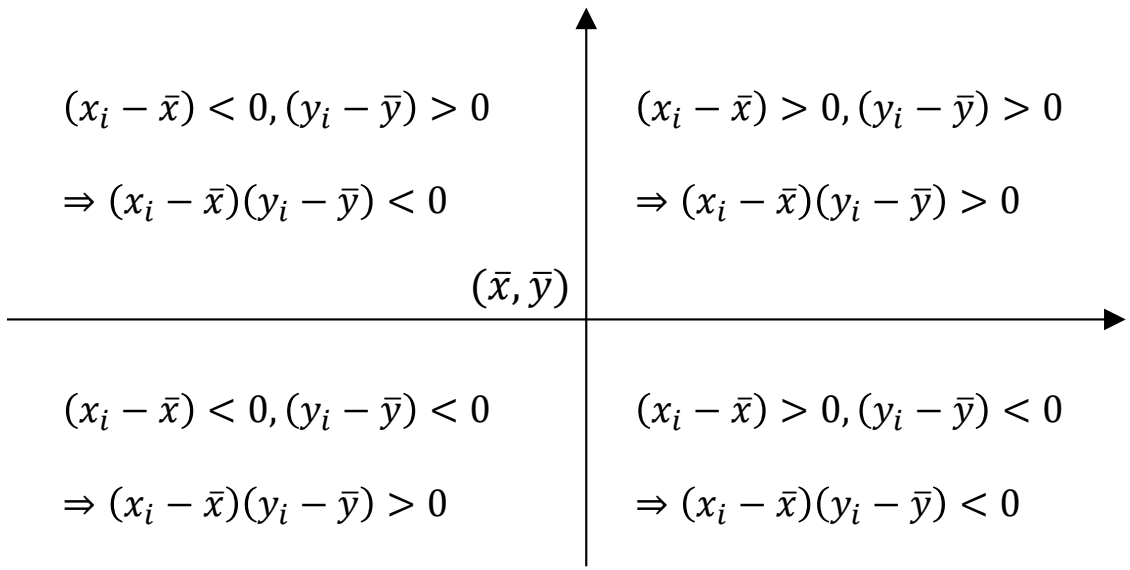
```
[1] 0.8829308
```

```
# Automatische Berechnung mit cor()
r_xy  = cor(x_i,y_i)                                                        # r_{xy}
print(r_xy)                                                                 # Ausgabe
```

```
[1] 0.8829308
```

⇒ Anzahl Therapiestunden und Symptomreduktion sind hochkorreliert.

Mechanik der Kovariationsterme

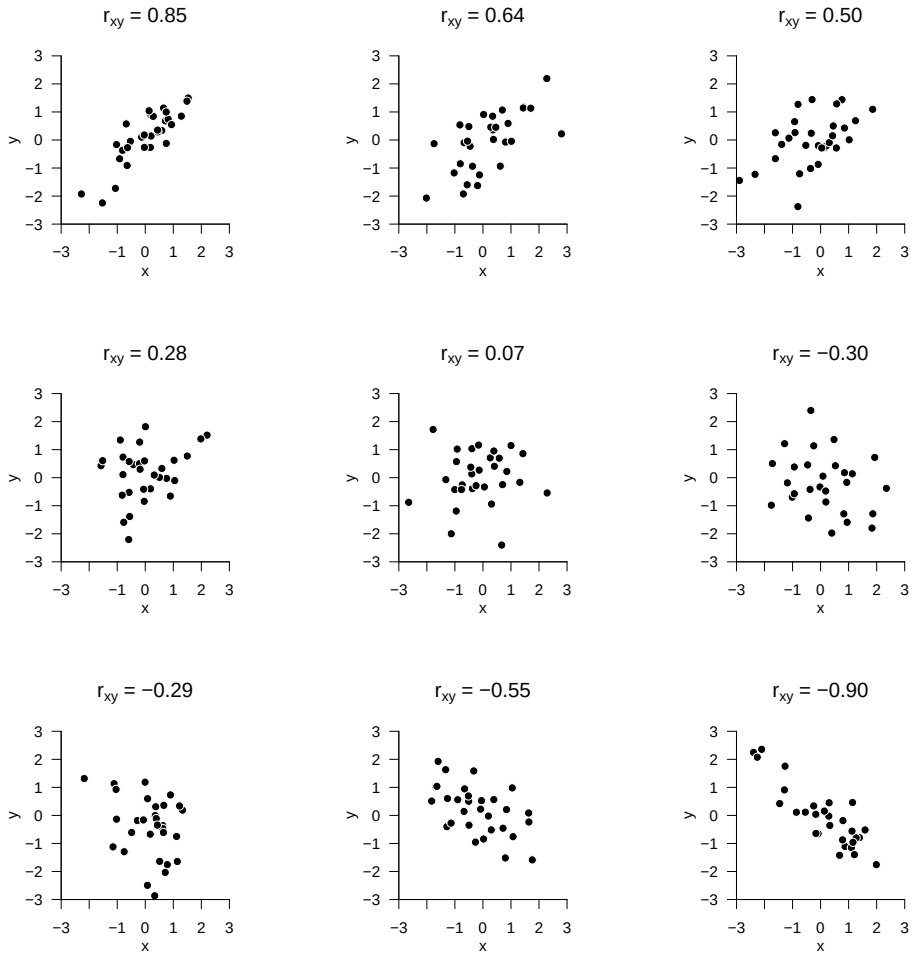


Häufige richtungsgleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ Positive Korrelation

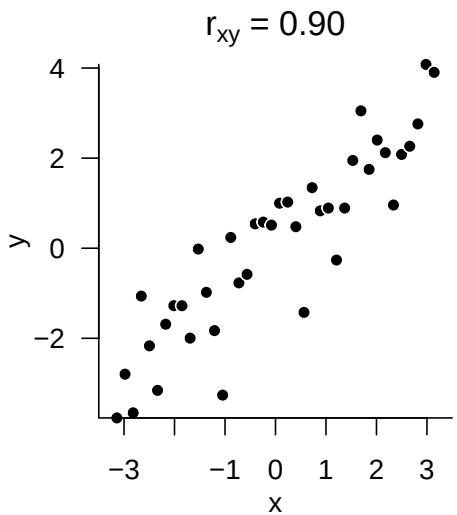
Häufige richtungsungleiche Abweichung der x_i und y_i von ihren Mittelwerten $\Rightarrow$ Negative Korrelation

Keine häufigen richtungsgleichen oder -entgegengesetzten Abweichungen $\Rightarrow$ Keine Korrelation

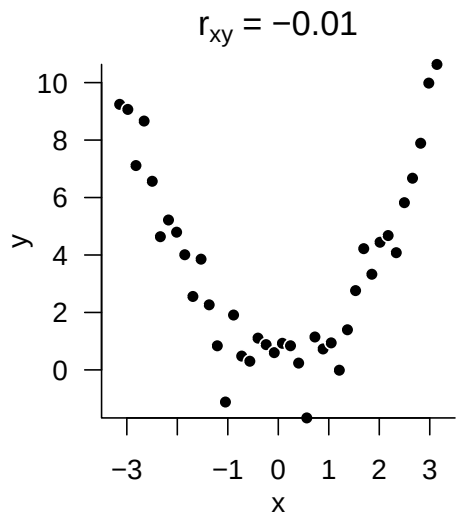
Beispiele


r_{xy} = 0.85
r_{xy} = 0.64
r_{xy} = 0.50
r_{xy} = 0.28
r_{xy} = 0.07
r_{xy} = −0.30
r_{xy} = −0.29
r_{xy} = −0.55
r_{xy} = −0.90
x
y

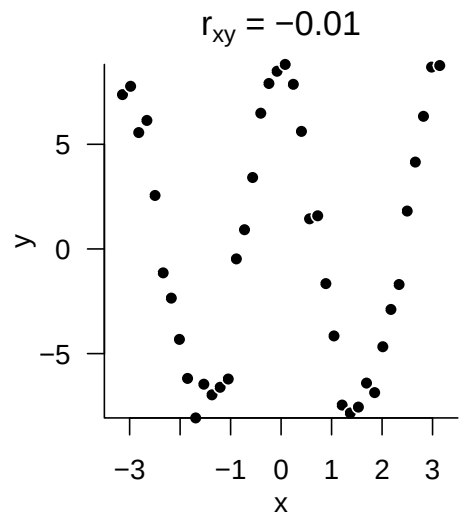
Funktionale Abhängigkeiten und Stichprobenkorrelation



$$y_i = x_i + \varepsilon_i$$



$$y_i = x_i^2 + \varepsilon_i$$



$$y_i = 8\cos(2x_i) + \varepsilon_i$$

$$\varepsilon_i \sim N(0, 1)$$

Theorem (Kovarianz und Korrelation bei linear-affinen Transformationen)

ξ und υ seien Zufallsvariablen und es seien $\alpha, \beta, \gamma, \delta \in \mathbb{R}$. Dann gelten

$$\mathbb{C}(\alpha\xi + \beta, \gamma\upsilon + \delta) = \alpha\gamma\mathbb{C}(\xi, \upsilon) \tag{6}$$

und

$$\rho(\alpha\xi + \beta, \gamma\upsilon + \delta) = \rho(\xi, \upsilon). \tag{7}$$

Bemerkungen

- Wir benötigen diese Aussage im Kontext der Kanonischen Korrelationsanalyse.
- Die Kovarianz zweier Zufallsvariablen ändert sich bei linear-affiner Transformation der Zufallsvariablen.
- Die Korrelation zweier Zufallsvariablen ändert sich bei linear-affiner Transformation der Zufallsvariablen nicht.

Korrelation

Beweis

Es gilt zunächst

$$\begin{aligned}
\mathbb{C}(\alpha\xi+\beta,\gamma\upsilon+\delta) &= \mathbb{E}((\alpha\xi+\beta-\mathbb{E}(\alpha\xi+\beta))(\gamma\upsilon+\delta-\mathbb{E}(\gamma\upsilon+\delta))) \\
&= \mathbb{E}((\alpha\xi+\beta-\alpha\mathbb{E}(\xi)-\beta)(\gamma\upsilon+\delta-\gamma\mathbb{E}(\gamma\upsilon)-\delta)) \\
&= \mathbb{E}(\alpha(\xi-\mathbb{E}(\xi))(\gamma(\upsilon-\mathbb{E}(\upsilon))) \\
&= \mathbb{E}(\alpha\gamma((\xi-\mathbb{E}(\xi))(\upsilon-\mathbb{E}(\upsilon)))) \\
&= \alpha\gamma\mathbb{C}(\xi,\upsilon)
\end{aligned} \tag{8}$$

Also folgt

$$\begin{aligned}
\rho(\alpha\xi+\beta,\gamma\upsilon+\delta) &= \frac{\mathbb{C}(\alpha\xi+\beta,\gamma\upsilon+\delta)}{\sqrt{\mathbb{V}(\alpha\xi+\beta)}\sqrt{\mathbb{V}(\gamma\upsilon+\delta)}} \\
&= \frac{\alpha\gamma\mathbb{C}(\xi,\upsilon)}{\sqrt{\alpha^2\mathbb{V}(\xi)}\sqrt{\gamma^2\mathbb{V}(\upsilon)}} \\
&= \frac{\alpha\gamma\mathbb{C}(\xi,\upsilon)}{\alpha\mathbb{S}(\xi)\gamma\mathbb{S}(\upsilon)} \\
&= \frac{\mathbb{C}(\xi,\upsilon)}{\mathbb{S}(\xi)\mathbb{S}(\upsilon)} \\
&= \rho(\xi,\upsilon).
\end{aligned} \tag{9}$$

Anwendungsszenario zur Kanonischen Korrelationsanalyse

Therapiegüte als Therapieerfolgsfaktor?

Unabhängige Variablen

- Anzahl Therapiestunden
- Erfahrung Therapeut:in

⇒ Maß für Therapiegüte

Abhängige Variablen

- BDI Score Reduktion
- Glucocorticoid Reduktion

⇒ Maß für Therapieerfolg

Beispieldatensatz zur Kanonischen Korrelationsanalyse

$i = 1, ..., n$ Patient:innen

`DUR` (x_1) Therapiedauer, `EXP` (x_2) Erfahrung Psychotherapeut:in

`dBDI` (y_1) BDI Score Reduktion, `dGLU` (y_2) Glukokortikoidplasmalevel Reduktion

DUR	EXP	dBDI	dGLU
27.9	7.8	35.5	6.1
15.3	9.3	25.0	4.0
17.4	2.1	19.7	1.7
21.5	6.5	28.8	2.6
28.2	1.3	29.4	1.9
14.0	2.7	17.2	0.9
28.0	3.9	32.9	2.0
28.9	0.1	28.3	4.1
23.2	3.8	25.8	3.9
22.6	8.7	31.3	3.8
11.2	3.4	14.4	2.1
14.1	4.8	18.4	2.0
13.5	6.0	19.1	5.0
23.7	4.9	28.0	2.6
17.7	1.9	20.3	2.1
25.4	8.3	34.8	4.4
20.0	6.7	27.6	4.0
24.4	7.9	31.9	3.9
29.8	1.1	32.2	1.0
17.6	7.2	24.6	1.9

Korrelation

Modellformulierung

Modellschätzung

Selbstkontrollfragen

Grundzüge der Kanonischen Korrelationsanalyse

Zufallsvektoren

u.i.v. Realisierungen

x			y		
x_{11}	$\cdots$	x_{m_x1}	y_{11}	$\cdots$	y_{m_y1}
x_{12}	$\cdots$	x_{m_x2}	y_{12}	$\cdots$	y_{m_y2}
$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
x_{1n}	$\cdots$	x_{m_xn}	y_{1n}	$\cdots$	y_{m_yn}

Parameter $a \in \mathbb{R}^{m_x}, b \in \mathbb{R}^{m_y}$

Linearkombinationen von Zufallsvektorkomponenten

$a^T x$	$b^T y$
$\sum_{i=1}^{m_x} a_i x_{1i}$	$\sum_{i=1}^{m_y} b_i y_{1i}$
$\sum_{i=1}^{m_x} a_i x_{2i}$	$\sum_{i=1}^{m_y} b_i y_{2i}$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\sum_{i=1}^{m_x} a_i x_{ni}$	$\sum_{i=1}^{m_y} b_i y_{ni}$

Definition $\xi := a^T x, \upsilon = b^T$

Korrelation von Zufallsvariablen

ξ	υ
ξ_1	υ_1
ξ_2	υ_2
$\vdots$	$\vdots$
$\vdots$	$\vdots$
ξ_n	υ_n

$\max_{a,b} \rho(\xi, \upsilon)$ u. d. N. $\mathbb{V}(\xi) = \mathbb{V}(\upsilon) = 1$

Grundzüge der Kanonischen Korrelationsanalyse

Man betrachtet multivariate unabhängige Variablen und multivariate abhängige Variablen. Die unabhängigen Variablen werden dabei auch *Prädiktoren*, die abhängigen Variablen auch *Kriterien* genannt.

Beobachtete Werte der Prädiktoren werden als u.i.v. Realisierungen eines m_x-dimensionalen Zufallsvektors x interpretiert, beobachtete Werte der Kriterien werden als u.i.v. Realisierungen eines m_y-dimensionalen Zufallsvektors y interpretiert.

Man fragt nach der maximal möglichen Korrelation von Linearkombinationen von x und y. Wir bezeichnen dazu Linearkombinationen von x und y mit Vektoren $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ mit

$$\xi = a^T x = a_1 x_1 + \cdots + a_{m_x} x_{m_x} \text{ und } \upsilon = b^T y = b_1 y_1 + \cdots + b_{m_y} y_{m_y}. \tag{10}$$

Insbesondere ξ und υ sind dann als Linearkombinationen von Zufallsvariablen selbst Zufallsvariablen, die Korrelation von ξ und υ bezeichnen wir mit $\rho(\xi, \upsilon)$.

Die spezifische Linearkombinationen $\xi = a^T x$ und $\upsilon = b^T y$ für die $\rho(\xi, \upsilon)$ maximal ist, können dann als "bester Prädiktor" und als "am besten prädizierbares Kriterium" interpretiert werden. Um diese zu bestimmenm fragt die Kanonische Korrelationsanalyse also nach Parametern $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ für die $\rho(\xi, \upsilon)$ maximal ist.

Für Skalare $\alpha, \beta \in \mathbb{R}$ sind die Korrelationen $\rho\left(a^T x, b^T y\right)$ und $\rho\left(\left(\alpha a^T\right) x, \left(\beta b^T\right) y\right)$ allerdings, wie im Theorem zu Kovarianz und Korrelation bei linear-affinen Transformationen gesehen, identisch. Man sucht deshalb Parameter $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ für die $\rho(\xi, \upsilon) = \rho\left(a^T x, b^T y\right)$ maximal ist und für die $a^T x$ und $b^T y$ jeweils eine Varianz von 1 haben, also $\mathbb{V}(\xi) = \mathbb{V}(\upsilon) = 1$ gilt.

Grundzüge der Kanonischen Korrelationsanalyse

Da mit dem Theorem zu Kovarianz und Korrelation bei linear-affinen Transformationen die Varianzen zu verschiedenen skalaren Vielfachen von $a^T x$ und $b^T y$ verschieden sind, legt $\mathbb{V}(\xi) = \mathbb{V}(\upsilon) = 1$ die $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$, für die $\rho(\xi, \upsilon)$ maximal ist, eindeutig fest. Zur Bestimmung von $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ ist man also auf ein restringiertes Optimierungsproblem geführt.

In der folgenden Entwicklung der Kanonischen Korrelationsanalyse folgen wir Mardia et al. (1979). Dabei werden die Zufallsvektoren x und y in einem Zufallsvektor

$$z := \begin{pmatrix} x \\ y \end{pmatrix} \tag{11}$$

zusammengefasst, für den wir durchgängig annehmen, dass $\mathbb{E}(z) = 0_m$ mit $m = m_x + m_y$. Dies entspricht auf der Anwendungsebene der Subtraktion des Stichprobenmittels von den beobachteten Daten vor Durchführung der Kanonischen Korrelationsanalyse

Der mathematische Fokus der Entwicklung nach Mardia et al. (1979) ist auf der Kovarianzmatrix $\mathbb{C}(z)$. Speziell ergeben sich die Kovarianzen von Linearkombinationen von x und y aus Matrixprodukten von $\mathbb{C}(z)$ und es können einige Matrixtheoreme, die im Folgenden diskutiert werden, auf diese Matrixprodukte angewendet werden. Generell wird in der Entwicklung nach Mardia et al. (1979) ein restringierter Optimierungsansatz mithilfe der Lagrangefunktion zugunsten der Eigenanalyse von Matrixprodukten supprimiert. Für die Entwicklung mit einem Lagrangeansatz, siehe zum Beispiel Anderson (2003) und die Originalarbeiten von Hotelling (1935) und Hotelling (1936).

Theorem (Kovarianzmatrizen von Zufallsvektoren)

Es seien

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \text{ mit } \mathbb{E}(z) := 0_m \tag{12}$$

ein $m_x + m_y$-dimensionaler Zufallsvektor und sein Erwartungswertvektor, respektive. Dann kann die $m \times m$ Kovarianzmatrix z geschrieben werden als

$$\mathbb{C}(z) = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{13}$$

wobei

$$\begin{aligned} \Sigma_{xx} &:= \mathbb{E}\left(xx^T\right) \in \mathbb{R}^{m_x \times m_x} \\ \Sigma_{xy} &:= \mathbb{E}\left(xy^T\right) \in \mathbb{R}^{m_x \times m_y} \\ \Sigma_{yx} &:= \mathbb{E}\left(yx^T\right) \in \mathbb{R}^{m_y \times m_x} \\ \Sigma_{yy} &:= \mathbb{E}\left(yy^T\right) \in \mathbb{R}^{m_x \times m_y} \end{aligned} \tag{14}$$

Beweis

Nach Definition der Kovarianzmatrix eines Zufallsvektors gilt

$$
\begin{aligned}
\mathbb{C}(z) &= \mathbb{E}\left((z-\mathbb{E}(z))(z-\mathbb{E}(z))^T\right) \\
&= \mathbb{E}\left((z-0_m)(z-0_m)^T\right) \\
&= \mathbb{E}\left(zz^T\right) \\
&= \mathbb{E}\left(\begin{pmatrix} x \\ y \end{pmatrix}\begin{pmatrix} x^T & y^T \end{pmatrix}\right) \\
&= \mathbb{E}\left(\begin{pmatrix} xx^T & xy^T \\ yx^T & yy^T \end{pmatrix}\right) \\
&= \begin{pmatrix} \mathbb{E}\left(xx^T\right) & \mathbb{E}\left(xy^T\right) \\ \mathbb{E}\left(yx^T\right) & \mathbb{E}\left(yy^T\right) \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}
\end{aligned}
\tag{15}
$$

□

Theorem (Linearkombinationen von Zufallsvektorpartitionen)

Es sei

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \text{ mit } \mathbb{E}(z) = 0_m \text{ und } \mathbb{C}(z) = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \tag{16}$$

ein m-dimensionaler partitionierter Zufallsvektor sowie sein Erwartungswertvektor und seine Kovarianzmatrix, respektive. Weiterhin seien für $a \in \mathbb{R}^{m_x}$ und $b \in \mathbb{R}^{m_y}$ die Zufallsvariablen

$$\xi := a^T x \text{ und } \upsilon := b^T y \tag{17}$$

als Linearkombinationen der Komponenten von x und y definiert. Dann gelten

(1) $\mathbb{V}(\xi) = a^T \Sigma_{xx} a$

(2) $\mathbb{V}(\upsilon) = b^T \Sigma_{yy} b$

(2) $\rho(\xi, \upsilon) = a^T \Sigma_{xy} b$, wenn $\mathbb{V}(\xi) = 1$ und $\mathbb{V}(\upsilon) = 1$.

Bemerkungen

- Die Varianz der Zufallsvariable $a^T x$ ergibt sich als "quadrierte Linearkombination" von Σ_{xx}.
- Die Varianz der Zufallsvariable $b^T y$ ergibt sich als "quadrierte Linearkombination" von Σ_{yy}.
- Die Korrelation der Zufallsvariablen $a^T x$ und $b^T y$ ergibt sich "quadrierte Linearkombination" von Σ_{xy}.

Beweis von (1) und (2)

Wir betrachten zunächst die Varianz von ξ. Mit dem Varianzverschiebungssatz gilt

$$\begin{aligned}
\mathbb{V}(\xi) &= \mathbb{E}\left(\xi\xi\right) - \mathbb{E}(\xi)\mathbb{E}(\xi) \\
&= \mathbb{E}\left((a^T x)(a^T x)\right) - \mathbb{E}\left(a^T x\right)\mathbb{E}\left(a^T x\right) \\
&= \mathbb{E}\left((a^T x)(a^T x)^T\right) - \mathbb{E}\left(a^T x\right)\mathbb{E}\left(a^T x\right) \\
&= \mathbb{E}\left(a^T x x^T a\right) - \mathbb{E}\left(a^T x\right)\mathbb{E}\left(a^T x\right) \\
&= a^T \mathbb{E}\left(x x^T\right) a - a^T \mathbb{E}(x) a^T \mathbb{E}(x) \\
&= a^T \mathbb{E}\left(x x^T\right) a - a^T 0_{m_x} a^T 0_{m_x} \\
&= a^T \Sigma_{xx} a.
\end{aligned} \tag{18}$$

Der Beweis zur Varianz von υ folgt dann analog.

Beweis von (3)

Mit der Definition der Korrelation von Zufallsvariablen, $\mathbb{V}(\xi) = \mathbb{V}(\upsilon) = 1$ und dem Kovarianzverschiebungssatz gilt

$$
\begin{aligned}
\rho(\xi, \upsilon) &= \frac{\mathbb{C}(\xi, \upsilon)}{\sqrt{\mathbb{V}(\xi)}\sqrt{\mathbb{V}(\upsilon)}} \\
&= \frac{\mathbb{C}(\xi, \upsilon)}{\sqrt{1}\sqrt{1}} \\
&= \mathbb{C}(\xi, \upsilon) \\
&= \mathbb{E}(\xi\upsilon) - \mathbb{E}(\xi)\mathbb{E}(\upsilon) \\
&= \mathbb{E}\left((a^T x)(b^T y)\right) - \mathbb{E}(a^T x)\mathbb{E}(b^T y) \\
&= \mathbb{E}\left((a^T x)(b^T y)^T\right) - \mathbb{E}(a^T x)\mathbb{E}(b^T y) \\
&= \mathbb{E}\left(a^T x y^T b\right) - \mathbb{E}(a^T x)\mathbb{E}(b^T y) \\
&= a^T \mathbb{E}\left(x y^T\right) b - a^T \mathbb{E}(x) b^T \mathbb{E}(y) \\
&= a^T \mathbb{E}\left(x y^T\right) b - a^T 0_{m_x} b^T 0_{m_y} \\
&= a^T \Sigma_{xy} b.
\end{aligned}
\tag{19}
$$

□

Definition (Kanonische Koeffizientenvektoren, Variate, Korrelationen)

Es seien

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \text{ mit } \mathbb{E}(z) := 0_m \text{ und } \mathbb{C}(z) := \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{20}$$

ein m-dimensionaler partitionierter Zufallsvektor sowie sein Erwartungswert und seine Kovarianzmatrix, respektive. Weiterhin sei

$$K := \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \in \mathbb{R}^{m_x \times m_y} \tag{21}$$

mit der Singulärwertzerlegung

$$K = A \Lambda B^T, \tag{22}$$

wobei

$$A := \begin{pmatrix} \alpha_1 & \cdots & \alpha_k \end{pmatrix} \in \mathbb{R}^{m_x \times m_y} \text{ und } B := \begin{pmatrix} \beta_1 & \cdots & \beta_k \end{pmatrix} \in \mathbb{R}^{m_y \times m_y} \tag{23}$$

die orthogonalen Matrix der Eigenvektoren von KK^T und die orthogonale Matrix der Eigenvektoren von $K^T K$, respektive, bezeichnen und

$$\Lambda := \text{diag}\left(\lambda_1^{1/2}, ..., \lambda_k^{1/2}\right) \in \mathbb{R}^{m_y \times m_y}, \tag{24}$$

die Diagonalmatrix der Quadratwurzeln der zugehörigen absteigend geordneten Eigenwerte bezeichnet. Schließlich seien für $i = 1, ..., k$

$$a_i := \Sigma_{xx}^{-1/2} \alpha_i \in \mathbb{R}^{m_x} \text{ und } b_i := \Sigma_{yy}^{-1/2} \beta_i \in \mathbb{R}^{m_y}. \tag{25}$$

Dann heißen für $i = 1, ..., k$

(1) $a_i \in \mathbb{R}^{m_x}$ und $b_i \in \mathbb{R}^{m_y}$ die *$i$ten kanonischen Koeffizientenvektoren*,

(2) die Zufallsvektoren $\xi_i := a_i^T x$ und $\upsilon_i := b_i^T y$ die $i$ten *$i$ten kanonischen Variaten* und

(3) $\rho_i := \lambda_i^{1/2}$ die *$i$te kanonische Korrelation*.

Theorem (Eigenschaften kanonischer Korrelationen und Variaten)

Es seien

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \text{ mit } \mathbb{E}(z) := 0_m \text{ und } \mathbb{C}(z) := \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{26}$$

ein m-dimensionaler partitionierter Zufallsvektor sowie sein Erwartungswert und seine Kovarianzmatrix, respektive. Weiterhin seien für $i = 1, \ldots, k$ die kanonischen Koeffizientenvektoren a_i, b_i, die kanonischen Variaten ξ, υ_i und die kanonischen Korrelationen ρ_i definiert wie oben. Dann gilt, dass für $1 \leq r \leq k$ das Maximum des $r$ten restringierten Optimierungsproblems

$$\phi_r = \max_{a,b} a^T \Sigma_{xy} b \tag{27}$$

unter den Nebenbedingungen

$$a^T \Sigma_{xx} a = 1, \quad b^T \Sigma_{yy} b = 1, \quad a_i^T \Sigma_{xx} a = 0 \text{ für } i = 1, \ldots, r-1 \tag{28}$$

(1) den Wert $\phi_r = \rho_r$ hat und (2) bei $a = a_r$ und $b = b_r$ angenommen wird.

Bemerkungen

- ϕ_1 ist die größtmögliche Korrelation von $\xi = a^T x$ und $\upsilon = b^T y$ unter den Nebenbedingungen
  - $\mathbb{V}(\xi) = a^T \Sigma_{xx} a = 1$ und $\mathbb{V}(\upsilon) = b^T \Sigma_{yy} b = 1$
- ϕ_r ist die größtmögliche Korrelation von $\xi = a^T x$ und $\upsilon = b^T y$ unter den Nebenbedingungen
  - $\mathbb{V}(\xi) = a^T \Sigma_{xx} a = 1$ und $\mathbb{V}(\upsilon) = b^T \Sigma_{yy} b = 1$
  - $\mathbb{C}(\xi_i, \xi) = a_i^T \Sigma_{xx} a = 0$ für die ersten $i = 1, \ldots, r-1$ kanonischen Variaten ξ_i

Simulationsbeispiel

Wir betrachten das Beispiel (vgl. Uurtio et al. (2018))

$$p(x) = N(x; 0_4, I_4) \text{ und } p(y|x) = N(y; Lx, G) \tag{29}$$

mit

$$L := \begin{pmatrix} 0.0 & 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -1.0 \end{pmatrix} \text{ und } G := \begin{pmatrix} 0.2 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.3 \end{pmatrix} \tag{30}$$

Hier gilt offenbar $m_x = 4, m_y = 3, m = 7$ und

$$\begin{aligned} y_1 &= \quad x_3 + \varepsilon_1 \\ y_2 &= \quad x_1 + \varepsilon_2 \\ y_3 &= -x_4 + \varepsilon_3 \end{aligned} \tag{31}$$

mit

$$x_1 \sim N(0,1), x_3 \sim N(0,1), x_4 \sim N(0,1) \tag{32}$$

und

$$\varepsilon_1 \sim N(0, 0.2), \varepsilon_2 \sim N(0, 0.4), \varepsilon_3 \sim N(0, 0.3) \tag{33}$$

Simulationsbeispiel

Mit dem Theorem zu gemeinsamen Normalverteilungen (vgl. Einheit (3) Zufallsvektoren) ergibt sich, dass

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N(0_7, \Sigma) \tag{34}$$

mit

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \tag{35}$$

wobei

$$\Sigma_{xx} = I_4, \quad \Sigma_{xy} = L^T, \quad \Sigma_{yx} = L \text{ und } \Sigma_{yy} = G + LL^T. \tag{36}$$

Explizit ergibt sich also

$$\Sigma = \begin{pmatrix} I_4 & L^T \\ L & G + LL^T \end{pmatrix} = \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & -1.0 \\ 0.0 & 0.0 & 1.0 & 0.0 & 1.2 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.4 & 0.0 \\ 0.0 & 0.0 & 0.0 & -1.0 & 0.0 & 0.0 & 1.3 \end{pmatrix} \tag{37}$$

Simulationsbeispiel

```
# R Pakete für Matrizenrechnung
library(expm)

# Modellparameter
L = matrix(c(0,0,1, 0,
             1,0,0, 0,
             0,0,0,-1),
           nrow  = 3,
           byrow = T)
G = diag(c(0.2,0.4,0.3))

# Kovarianzmatrixpartition
Sigma_xx = diag(4)
Sigma_xy = t(L)
Sigma_yx = L
Sigma_yy = G + L %*% t(L)
Sigma    = rbind(cbind(Sigma_xx, Sigma_xy), cbind(Sigma_yx, Sigma_yy))
print(Sigma)
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1    0    0    0  0.0  1.0  0.0
[2,]    0    1    0    0  0.0  0.0  0.0
[3,]    0    0    1    0  1.0  0.0  0.0
[4,]    0    0    0    1  0.0  0.0 -1.0
[5,]    0    0    1    0  1.2  0.0  0.0
[6,]    1    0    0    0  0.0  1.4  0.0
[7,]    0    0    0   -1  0.0  0.0  1.3
```

Simulationsbeispiel

```
# Evaluation der iten kanonischen Koeffizientenvektoren und Korrelationen
K      = sqrtm(solve(Sigma_xx)) %*% Sigma_xy %*% sqrtm(solve(Sigma_yy)) # K
ALB    = svd(K)                                                        # K = A\LambdaV
A      = ALB$u                                                         # A
Lambda = ALB$d                                                         # Lambda
B      = ALB$v                                                         # B
rho    = Lambda                                                        # \rho_i = \lambda_i^{1/2}
a      = sqrtm(solve(Sigma_xx)) %*% A                                    # a_i = \Sigma_{xx}^{-1/2}\alpha_i
b      = sqrtm(solve(Sigma_yy)) %*% B                                    # b_i = \Sigma_{yy}^{-1/2}\beta_i
```

Die kanonische Korrelationen und kanonischen Koeffizientenvektoren ergeben sich zu

```
rho_1 =  0.9128709 , a_1^T = ( 0 0 -1 0 ), b_1^T = ( -0.9128709 0 0 )
rho_2 =  0.877058  , a_2^T = ( 0 0 0 1 ) , b_2^T = ( 0 0 -0.877058 )
rho_3 =  0.8451543 , a_3^T = ( -1 0 0 0 ), b_3^T = ( 0 -0.8451543 0 )
```

Korrelation

Modellformulierung

Modellschätzung

Selbstkontrollfragen

Definition (Schätzer der kanonischen Korrelationsanalyse)

Für $i = 1, \ldots, n$ seien

$$z_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix} \text{ mit } \mathbb{E}(z_i) := 0_m \text{ und } \mathbb{C}(z_i) := \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{38}$$

unabhängig und identisch verteilte m-dimensionale partitionierte Zufallsvektoren sowie ihr Erwartungswert und ihre Kovarianzmatrix, respektive, und

$$C := \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \in \mathbb{R}^{m \times m} \tag{39}$$

sei ihre Stichprobenkovarianzmatrix. Dann sind für $i = 1, \ldots, k := \min\{m_x, m_y\}$

$$\hat{a}_i := C_{xx}^{-1/2} \hat{\alpha}_i \in \mathbb{R}^{m_x}, \quad \hat{b}_i := C_{yy}^{-1/2} \hat{\beta}_i \in \mathbb{R}^{m_y} \text{ und } \hat{\rho}_i := \hat{\lambda}_i^{1/2} \tag{40}$$

Schätzer der $i$ten kanonischen Koeffizientenvektoren und kanonischen Korrelationen, respektive. Dabei sind mit

$$\hat{K} := C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2} \in \mathbb{R}^{m_x \times m_y} \tag{41}$$

$\hat{\alpha}_i$ und $\hat{\lambda}_i$ der $i$te Eigenvektor und sein zugehöriger Eigenwert von $\hat{K}\hat{K}^T$ und $\hat{\beta}_i$ der entsprechende Eigenkvektor von $\hat{K}^T\hat{K}$.

Bemerkungen

- Zur Modellschätzung wird $\mathbb{C}(z)$ also durch C ersetzt.

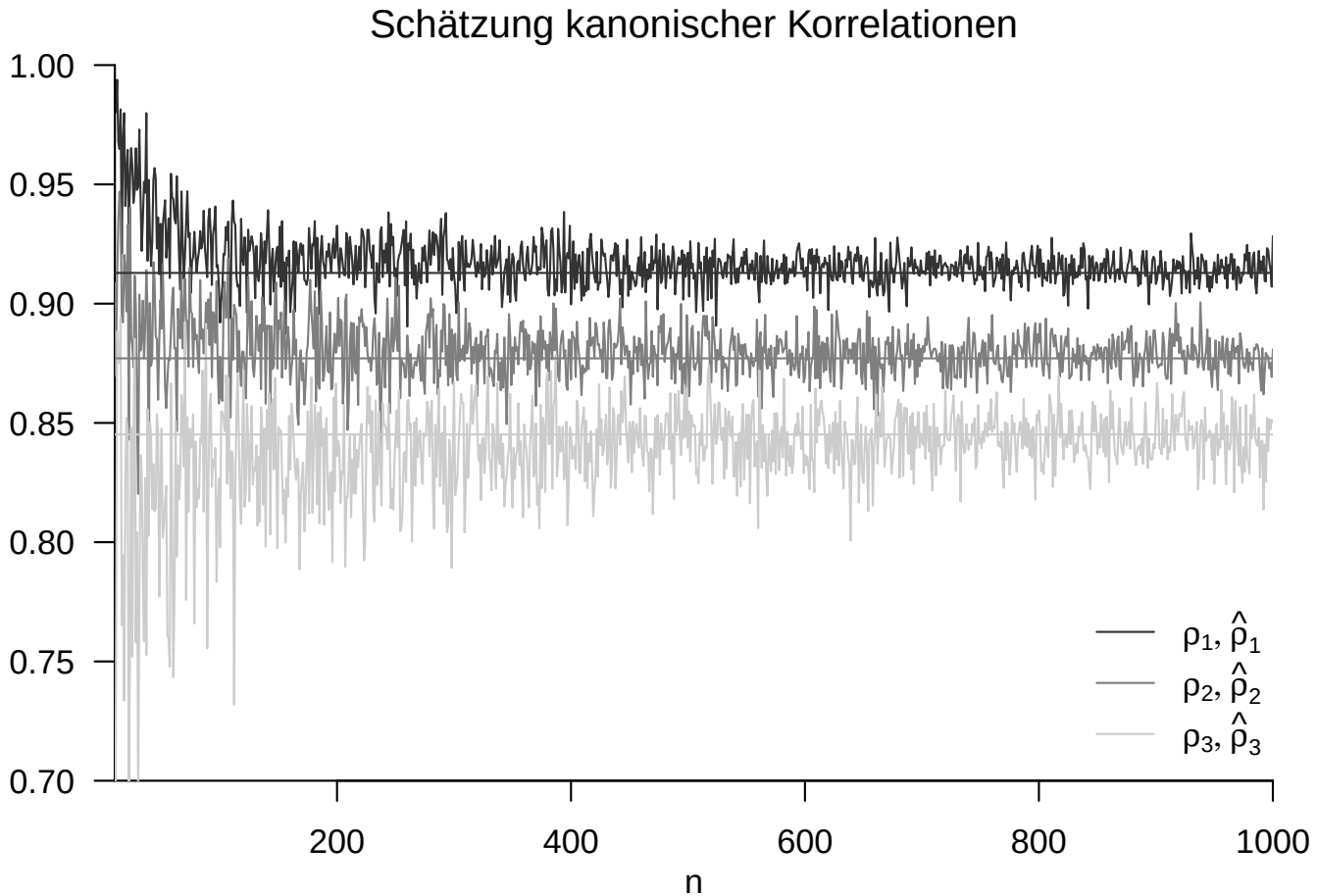
Simulationsbeispiel

```
# R Pakete
library(MASS)
library(expm)

# Modellparameter
m_x      = 4
m_y      = 3
k        = min(m_x,m_y)
L        = matrix(c(0,0,1,0,1,0,0,0,0,0,0,-1), nrow = 3,byrow = 3)
G        = diag(c(0.2,0.4,0.3))
Sigma_xx = diag(4)
Sigma_xy = t(L)
Sigma_yx = L
Sigma_yy = G + L %*% t(L)
Sigma    = rbind(cbind(Sigma_xx, Sigma_xy), cbind(Sigma_yx, Sigma_yy))
K        = sqrtm(solve(Sigma_xx)) %*% Sigma_xy %*% sqrtm(solve(Sigma_yy))
ALB      = svd(K)
A        = ALB$u
Lambda   = ALB$d
B        = ALB$v
rho      = Lambda
a        = sqrtm(solve(Sigma_xx)) %*% A
b        = sqrtm(solve(Sigma_yy)) %*% B
```

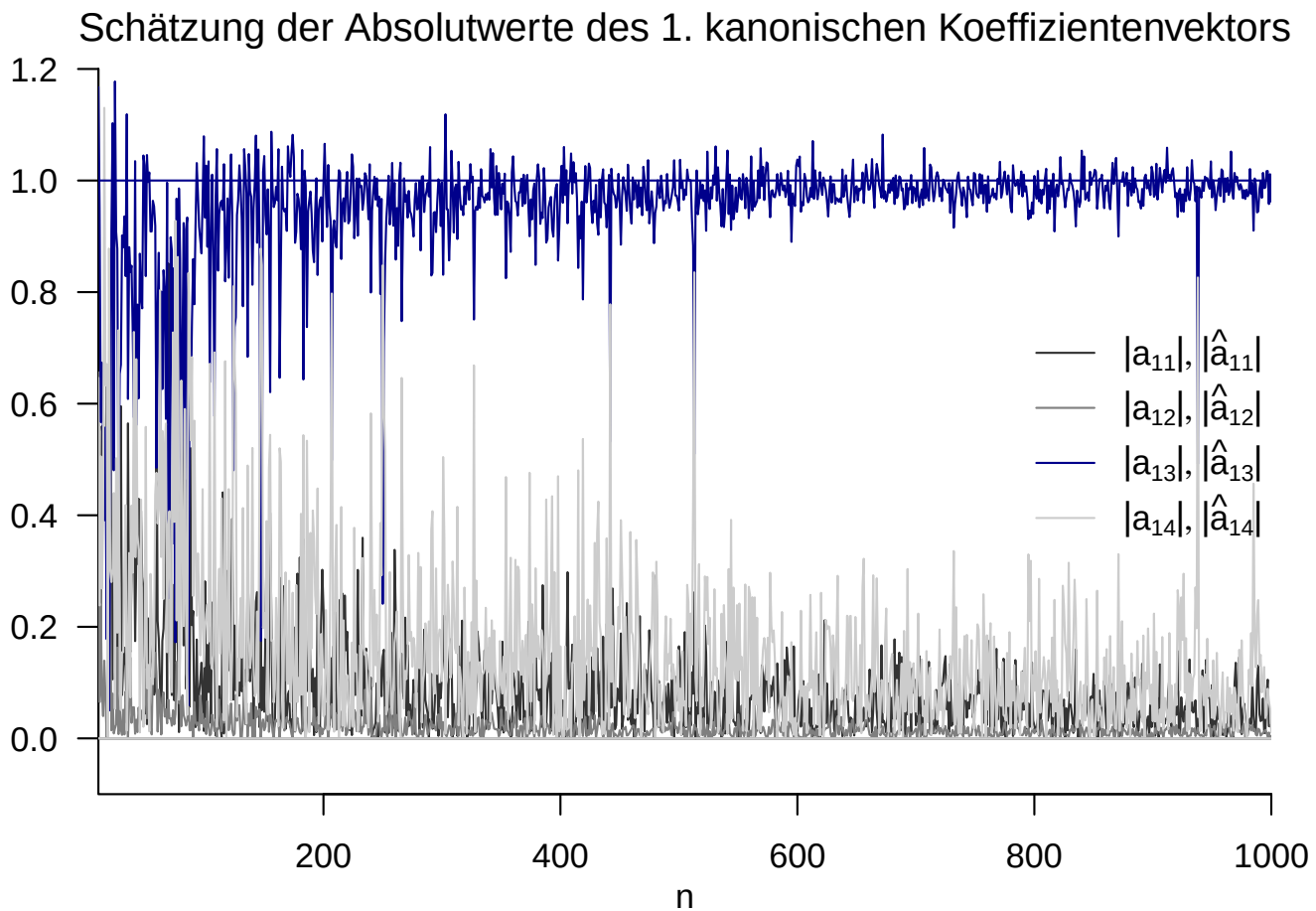
Simulationsbeispiel

```
# Simulationen
n       = 1e1:1e3
rho_hat = matrix(rep(NaN, length(n)*k)  , nrow = k)
a_1_hat = matrix(rep(NaN, length(n)*m_x), nrow = m_x)
for(i in 1:length(n)){

    # Datengeneration
    Y           = t(mvrnorm(n[i],rep(0, m_x+m_y),Sigma))
    I_n         = diag(n[i])
    J_n         = matrix(rep(1,n[i]^2), nrow = n[i])

    # Stichprobenkovarianzmatrixpartition
    C           = (1/(n[i]-1))*(Y %*% (I_n-(1/n[i])*J_n) %*% t(Y))
    C_xx        = C[1:m_x,1:m_x]
    C_xy        = C[1:m_x,(m_x+1):(m_x+m_y)]
    C_yx        = C[(m_x+1):(m_x+m_y),1:m_x]
    C_yy        = C[(m_x+1):(m_x+m_y),(m_x+1):(m_x+m_y)]

    # Kanonische Korrelationsanalyse
    K_hat       = sqrtm(solve(C_xx)) %*% C_xy %*% sqrtm(solve(C_yy))
    ALB_hat     = svd(K_hat)
    A_hat       = ALB_hat$u
    Lambda_hat  = ALB_hat$d
    B_hat       = ALB_hat$v
    a_hat       = sqrtm(solve(C_xx)) %*% A_hat
    b_hat       = sqrtm(solve(C_yy)) %*% B_hat
    rho_hat[,i] = as.matrix(Lambda_hat)
    a_1_hat[,i] = a_hat[,1]
}
```

Simulationsbeispiel



Simulationsbeispiel



Anwendungsbeispiel

Therapiegüte als Therapieerfolgsfaktor?

Unabhängige Variablen

- Anzahl Therapiestunden
- Erfahrung Therapeut:in

⇒ Maß für Therapiegüte

Abhängige Variablen

- BDI Score Reduktion
- Glucocorticoid Reduktion

⇒ Maß für Therapieerfolg

Anwendungsbeispiel

$i = 1, \ldots, n$ Patient:innen

`DUR` (x_1) Therapiedauer, `EXP` (x_2) Erfahrung Psychotherapeut:in

`dBDI` (y_1) BDI Score Reduktion, `dGLU` (y_2) Glukokortikoidplasmalevel Reduktion

DUR	EXP	dBDI	dGLU
27.9	7.8	35.5	6.1
15.3	9.3	25.0	4.0
17.4	2.1	19.7	1.7
21.5	6.5	28.8	2.6
28.2	1.3	29.4	1.9
14.0	2.7	17.2	0.9
28.0	3.9	32.9	2.0
28.9	0.1	28.3	4.1
23.2	3.8	25.8	3.9
22.6	8.7	31.3	3.8
11.2	3.4	14.4	2.1
14.1	4.8	18.4	2.0
13.5	6.0	19.1	5.0
23.7	4.9	28.0	2.6
17.7	1.9	20.3	2.1
25.4	8.3	34.8	4.4
20.0	6.7	27.6	4.0
24.4	7.9	31.9	3.9
29.8	1.1	32.2	1.0
17.6	7.2	24.6	1.9

Anwendungsbeispiel

```
# R Pakete
library(expm)

# Datenpräprozessierung
D           = read.csv("./7_Daten/7_Kanonische_Korrelationsanalyse.csv")
x           = as.matrix(cbind(D$DUR, D$EXP))
y           = as.matrix(cbind(D$dBDI, D$dGLU))
n           = nrow(x)
m_x         = ncol(x)
m_y         = ncol(y)
Y           = t(cbind(x,y))

# Stichprobenkovarianzmatrixpartition
I_n         = diag(n)
J_n         = matrix(rep(1,n^2), nrow = n)
C           = (1/(n-1))*(Y %*% (I_n-(1/n)*J_n) %*% t(Y))
C_xx        = C[1:m_x,1:m_x]
C_xy        = C[1:m_x,(m_x+1):(m_x+m_y)]
C_yx        = C[(m_x+1):(m_x+m_y),1:m_x]
C_yy        = C[(m_x+1):(m_x+m_y),(m_x+1):(m_x+m_y)]

# Kanonische Korrelationsanalyse
K_hat       = sqrtm(solve(C_xx)) %*% C_xy %*% sqrtm(solve(C_yy))
ALB_hat     = svd(K_hat)
A_hat       = ALB_hat$u
Lambda_hat  = ALB_hat$d
B_hat       = ALB_hat$v
a_hat       = sqrtm(solve(C_xx)) %*% A_hat
b_hat       = sqrtm(solve(C_yy)) %*% B_hat
rho_hat     = as.matrix(Lambda_hat)
```

```
rho_hat_1 :  0.9950575
a_hat_1   :  -0.1623409 -0.173979
b_hat_1   :  -0.1554175 -0.05025419
rho_hat_2 :  0.5010358
a_hat_2   :  -0.06026274 0.3118808
b_hat_2   :  -0.08128072 0.7773036
```

Anwendungsbeispiel

Kanonische Korrelationsanalyse mir R's `cancor()` Funktion

```
# Datenpräprozessierung
D          = read.csv("./7_Daten/7_Kanonische_Korrelationsanalyse.csv")
x          = as.matrix(cbind(D$DUR , D$EXP))
y          = as.matrix(cbind(D$dBDI, D$dGLU))
cca        = cancor(x,y)
```

```
rho_hat_1 :  0.9950575
rho_hat_2 :  0.5010358
```

Anwendungsbeispiel

Die geschätzte maximale Korrelation einer Linearkombination von (`DUR`, `EXP`) und (`dBDI`, `dGLU`) ist 0.99.

- (`DUR`, `EXP`) und (`dBDI`, `dGLU`) sind multivariat also hochgradig korreliert.

Basierend auf der Schätzung der kanonischen Koeffizientenvektoren ergibt sich

- $\xi = 0.16$ DUR $+ 0.17$ EXP als "bester Prädiktor"
- $\upsilon = 0.15$ dBDI $+ 0.05$ dGLU als "am besten prädizierbares Kriterium"

Therapiedauer `DUR` und Therapeut:innenerfahrung `EXP` scheinen zur bestmöglichen Prädiktion der Therapiegüte also in etwa gleichbedeutend, bei dem bestprädizierbarem Kriterium der Therapiegüte trägt die BDI Score Reduktion `dBDI` etwas mehr bei als die Glukokortikoidplasmalevel Reduktion `dGLU` bei - alles angesichts der betrachteten Datenskalierung.

Korrelation

Modellformulierung

Modellschätzung

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition der Korrelation wieder.
2. Geben Sie die Definition der Stichprobenkorrelation wieder.
3. Geben Sie das Theorem zu Kovarianz und Korrelation bei linear-affinen Transformationen wieder.
4. Erläutern Sie das Anwendungsszenario einer Kanonischen Korrelationsanalyse.
5. Erläutern Sie das Ziel einer Kanonischen Korrelationsanalyse.
6. Erläutern Sie die Begriffe “bester Prädiktor” und “am besten prädizierbares Kriterium”.
7. Geben Sie die Definition Kanonischer Koeffizientenvektoren, Variate und Korrelationen wieder.
8. Geben Sie das Theorem zu den Eigenschaften kanonischer Korrelationen und Variate wieder.
9. Geben Sie die Definition für Schätzer kanonischer Korrelationen und Koeffizientenvektoren wieder.
10. Erläutern Sie die Durchführung einer kanonischen Korrelationsanalyse eines Datensatzes.

Definition (Symmetrische Quadratwurzel einer Matrix)

$A \in \mathbb{R}^{m \times m}$ sei eine invertierbare symmetrische Matrix mit positiven Eigenwerten. Dann sind für $r \in \mathbb{N}^0$ und $s \in \mathbb{N}$ die rationalen Potenzen von A mit der orthonormalen Matrix $Q \in \mathbb{R}^{m \times m}$ der Eigenvektoren von A und der Diagonalmatrix $\Lambda = \text{diag}(\lambda_i) \in \mathbb{R}^{m \times m}$ der zugehörigen Eigenwerte $\lambda_1, ..., \lambda_m$ von A definiert als

$$A^{r/s} = Q\Lambda^{r/s}Q^T \text{ mit } \Lambda^{r/s} = \text{diag}\left(\lambda_i^{r/s}\right). \tag{42}$$

Der Spezialfall $r := 1, s := 2$ wird als symmetrische Quadratwurzel von A bezeichnet und hat die Form

$$A^{1/2} = Q\Lambda^{1/2}Q^T \text{ mit } \Lambda^{1/2} = \text{diag}\left(\lambda_i^{1/2}\right). \tag{43}$$

Bemerkungen

- Offenbar gilt
$$\left(A^{1/2}\right)^2 = Q\Lambda^{1/2}Q^TQ\Lambda^{1/2}Q^T = Q\Lambda^{1/2}\Lambda^{1/2}Q^T = Q\Lambda Q^T = A. \tag{44}$$
- Weiterhin gilt
$$\left(A^{-1/2}\right)^2 = Q\Lambda^{-1/2}Q^TQ\Lambda^{-1/2}Q^T = Q\Lambda^{-1/2}\Lambda^{-1/2}Q^T = Q\Lambda^{-1}Q^T = A^{-1}. \tag{45}$$
- Schließlich gilt
$$\begin{aligned} A^{-1/2}AA^{-1/2} = Q\Lambda^{-1/2}Q^TQ\Lambda Q^TQ\Lambda^{-1/2}Q^T &= Q\Lambda^{-1/2}\Lambda\Lambda^{-1/2}Q^T \\ &= Q\Lambda\Lambda^{-1}Q^T \\ &= I_m \end{aligned} \tag{46}$$

Theorem (Eigenwerte und Eigenvektoren von Matrixprodukten)

Für $A \in \mathbb{R}^{n \times m}$ und $B \in \mathbb{R}^{m \times n}$ sind die Eigenwerte von $AB \in \mathbb{R}^{n \times n}$ und $BA \in \mathbb{R}^{m \times m}$ gleich. Weiterhin gilt, dass für einen Eigenvektor v zu einem von Null verschiedenen Eigenwert λ von AB $w := Bv$ ein Eigenvektor von BA zum Eigenwert λ ist.

Bemerkungen

- Für einen Beweis siehe Mardia et al. (1979), S. 468.

```
A   = matrix(1:6, nrow = 2,  byrow = T)          # Matrix A \in \mathbb{R}^{2 x 3}
B   = matrix(1:6, ncol = 2,  byrow = T)          # Matrix B \in \mathbb{R}^{3 x 2}
EAB = eigen(A %*% B)                             # Eigenanalyse von AB \in \mathbb{R}^{2 \times 2}
EBA = eigen(B %*% A)                             # Eigenanalyse von BA \in \mathbb{R}^{3 \times 3}
w   = B %*% EAB$vectors[,1]                      # Eigenvektor von BA
cat("Eigenwerte von AB :"  , EAB$values[1:2],
    "\nEigenwerte von BA :", EBA$values[1:2],
    "\nBAw mit w = Bv    :", B %*% A %*% w,
    "\nlw mit w = Bv     :", EBA$values[1] * w)
```

```
Eigenwerte von AB : 85.57934 0.4206623
Eigenwerte von BA : 85.57934 0.4206623
BAw mit w = Bv    : -191.1333 -416.7586 -642.3839
lw mit w = Bv     : -191.1333 -416.7586 -642.3839
```

Theorem (Eigenwert und Eigenvektor eines Matrixvektorprodukts)

Für $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{p \times n}, a \in \mathbb{R}^m$ und $b \in \mathbb{R}^p$ gilt, dass der einzige von Null verschiedene Eigenwert von $Aab^T B \in \mathbb{R}^{n \times n}$ gleich $b^T BAa$ mit zugehörigem Eigenvektor Aa ist.

Bemerkungen

- Für einen Beweis siehe Mardia et al. (1979), S. 468.

```
A      = matrix(1:6, nrow = 2,  byrow = T)     # Matrix A \in \mathbb{R}^{2 x 3}
B      = matrix(1:8, ncol = 2,  byrow = T)     # Matrix B \in \mathbb{R}^{4 x 2}
a      = matrix(1:3, nrow = 3,  byrow = T)     # Vektor a \in \mathbb{R}^{3 x 1}
b      = matrix(1:4, nrow = 4,  byrow = T)     # Vektor b \in \mathbb{R}^{4 x 1}
EAabTB = eigen(A %*% a %*% t(b) %*% B)         # Eigenanalyse von Aab^TB \in \mathbb{R}^{4 x 4}
cat("Eigenwerte von AabTB :", EAabTB$values,
    "\nbTBAa                :", t(b) %*% B %*% A %*% a,
    "\nAa                   :", A %*% a,
    "\n(AabTB)Aa            :",(A %*% a %*% t(b) %*% B) %*% A %*% a,            # Mv
    "\n(bTBAa)Aa            :",as.vector((t(b) %*% B %*% A %*% a)) * (A %*% a)) # = \lambda v
```

```
Eigenwerte von AabTB : 2620 0
bTBAa                : 2620
Aa                   : 14 32
(AabTB)Aa            : 36680 83840
(bTBAa)Aa            : 36680 83840
```

Theorem (Maximierung quadratischer Formen mit Nebenbedingungen)

$A \in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{m \times m}$ p.d. seien symmetrische Matrizen und λ_1 sei der größte Eigenwert von $B^{-1}A$ mit assoziertem Eigenvektor $v_1 \in \mathbb{R}^m$. Dann ist λ_1 eine Lösung des Optimierungsproblems

$$\max_x x^T Ax \text{ unter der Nebenbedingung } x^T Bx = 1. \tag{47}$$

Bemerkungen

- Das Theorem ist direkt durch die kanonische Korrelationsanalyse motiviert.
- $\max_x f(x)$ ist das Maximum einer Funktion f, also der Wert der Funktion an der Maximumstelle x
- $\arg\max_x f(x)$ ist die Maximumstelle einer Funktion, also ein Wert in der Definitionsmenge von f.
- Nach Wortlaut des Theorems gilt also für die Funktion

$$f : \mathbb{R}^m \to \mathbb{R}, x \mapsto f(x) := x^T Ax, \tag{48}$$

dass

$$v_1 = \arg\max_x x^T Ax \text{ unter der Nebenbedingung } x^T Bx = 1 \tag{49}$$

und dass

$$\lambda_1 = \max_x x^T Ax \text{ unter der Nebenbedingung } x^T Bx = 1. \tag{50}$$

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten

$B^{1/2}$ sei die symmetrische Quadratwurzel von B und es sei

$$y := B^{1/2}x \Leftrightarrow x = B^{-1/2}y \tag{51}$$

Dann kann mit der symmetrischen Matrix

$$K := B^{-1/2}AB^{-1/2} \in \mathbb{R}^{m\times m} \tag{52}$$

das Optimierungsproblem (47) geschrieben werden als

$$\max_y y^T Ky \text{ unter der Nebenbedingung } y^T y = 1. \tag{53}$$

Dies gilt, weil

$$\max_x x^T Ax \Leftrightarrow \max_y \left(B^{-1/2}y\right)^T A\left(B^{-1/2}y\right) \Leftrightarrow \max_y y^T B^{-1/2}AB^{-1/2}y \Leftrightarrow \max_y y^T Ky \tag{54}$$

und

$$x^T Bx = 1 \Leftrightarrow y^T B^{-1/2}BB^{-1/2}y = 1 \Leftrightarrow y^T y = 1. \tag{55}$$

Weil K eine symmetrische Matrix ist, existiert die Orthonormalzerlegung (vgl. (2) Eigenanalyse)

$$K = Q\Lambda Q^T, \tag{56}$$

wobei die Spalten der orthogonalen Matrix Q die Eigenvektoren von K und die Diagonalemente von Λ die zugehörigen Eigenwerte von K sind.

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten (fortgeführt)

Mit der orthogonalen Matrix Q aus obiger Orthornomalzerlegung sei nun

$$z := Q^T y \Leftrightarrow y := Qz. \tag{57}$$

Dann kann das Optimierungsproblem (53) geschrieben werden als

$$\max_z \sum_{i=1}^{m} \lambda_i z_i^2 \text{ unter der Nebenbedingung } z^T z = 1, \tag{58}$$

weil

$$\max_y y^T K y \Leftrightarrow \max_z (Qz)^T K (Qz) \Leftrightarrow \max_z z^T Q^T Q \Lambda Q^T Q z \Leftrightarrow \max_z z^T \Lambda z \Leftrightarrow \max_z \sum_{i=1}^{m} \lambda_i z_i^2 \tag{59}$$

und

$$y^T y = 1 \Leftrightarrow (Qz)^T Qz = 1 \Leftrightarrow z^T Q^T Q z = 1 \Leftrightarrow z^T z = 1. \tag{60}$$

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten (fortgeführt)

Die Eigenwerte von K seien nun absteigend sortiert, also $\lambda_1 \geq \cdots \geq \lambda_m$. Dann gilt für das Optimierungsproblem (58), dass

$$\max_z \sum_{i=1}^{m} \lambda_i z_i^2 \leq \lambda_1, \tag{61}$$

weil

$$\max_z \sum_{i=1}^{m} \lambda_i z_i^2 \leq \max_z \sum_{i=1}^{m} \lambda_1 z_i^2 = \lambda_1 \max_z \sum_{i=1}^{m} z_i^2 = \lambda_1 \tag{62}$$

wobei sich die letzte Gleichung aus der Nebenbedingung $z^T z = 1$ ergibt. Schließlich gilt

$$\max_z \sum_{i=1}^{m} \lambda_i z_i^2 = \lambda_1, \tag{63}$$

für $z := e_1 = (1, 0, \ldots, 0)^T$. Zusammenfassend heißt das, dass $z = e_1$ eine Lösung des Optimierungsproblem (58) ist und das λ_1 das entsprechende Maximum ist.

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten (fortgeführt)

Damit ergibt sich aber sofort, dass dann

$$y = Qz = Qe_1 = q_1 \text{ und } x = B^{-1/2}q_1 \tag{64}$$

Lösungen der äquivalenten Optimierungsprobleme (53) und (47), respektive, sind. Nach Konstruktion ist q_1 ein Eigenvektor von $B^{-1/2}AB^{-1/2}$ und nach obigem Theorem zu Eigenwerten und Eigenvektoren von Matrixprodukten damit auch ein Eigenvektor von

$$B^{-1/2}B^{-1/2}A = B^{-1}A \tag{65}$$

und die zugehörigen Eigenwerte sind gleich. Damit aber folgt, dass der größte Eigenwert von $B^{-1}A$ und sein assoziierter Eigenvektor eine Lösung von

$$\max_x x^T Ax \text{ unter der Nebenbedingung } x^T Bx = 1. \tag{66}$$

ist. □

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten (fortgeführt)

Wir betrachten das restringierte Optimierungsproblem

$$\phi_r^2 = \max_{a,b} \left(a^T \Sigma_{xy} b\right)^2 \text{ u.d.N. } a^T \Sigma_{xx} a = 1,\, b^T \Sigma_{yy} b = 1,\, a_i^T \Sigma_{xx} a = 0, i = 1, ..., r-1 \tag{67}$$

Wir folgen Mardia et al. (1979), S. 284 und gehen schrittweise vor, d.h. wir lösen das restringierte Optimierungsproblem

$$\phi_r^2 = \max_a \left(\max_b \left(a^T \Sigma_{xy} b\right)^2 \text{ u.d.N.} b^T \Sigma_{yy} b = 1 \right) \text{ u.d.N. } a^T \Sigma_{xx} a = 1,\, a_i^T \Sigma_{xy} a = 0, i = 1, ..., r-1 \tag{68}$$

von innen nach außen.

Schritt (1)

Wir wählen wir zunächst ein festes $a \in \mathbb{R}^m$ und betrachten das restringierte Optimierungsproblem

$$\max_b \left(a^T \Sigma_{xy} b\right)^2 \text{ u.d.N. } b^T \Sigma_{yy} b = 1 \tag{69}$$

Dieses Optimierungsproblem kann geschrieben werden als

$$\max_b b^T \Sigma_{yx} a a^T \Sigma_{xy} b \text{ u.d.N. } b^T \Sigma_{yy} b = 1, \tag{70}$$

weil gilt, dass

$$\left(a^T \Sigma_{xy} b\right)^2 = \left(a^T \Sigma_{xy} b\right)\left(a^T \Sigma_{xy} b\right) = \left(a^T \Sigma_{xy} b\right)^T a^T \Sigma_{xy} b = b^T \Sigma_{yx} a a^T \Sigma_{xy} b. \tag{71}$$

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten (fortgeführt)

Das Optimierungsproblem (70) kann nun mithilfe des Theorems zur Maximierung quadratischer Formen mit Nebenbedingen gelöst werden. Im Sinne dieses Theorems setzen wir dazu

$$A := \Sigma_{yx} a a^T \Sigma_{xy} \text{ und } B := \Sigma_{yy}. \tag{72}$$

Dann hat (70) die Form

$$\max_b b^T A b \text{ unter der Nebenbedingung } b^T B b = 1, \tag{73}$$

Das Maximum von (73) entspricht nach dem Theorem zur Maximierung quadratischer Formen mit Nebenbedingungen dem größten Eigenwert von

$$B^{-1} A = \Sigma_{yy}^{-1} \Sigma_{yx} a a^T \Sigma_{xy} \tag{74}$$

Der größte Eigenwert von $\Sigma_{yy}^{-1} \Sigma_{yx} a a^T \Sigma_{xy}$ wiederum kann mithilfe des Theorems zum Eigenwert und Eigenvektor eines Matrixvektorprodukts bestimmt werden. Im Sinne dieses Theorems setzen wir dazu

$$A := \Sigma_{yy}^{-1} \Sigma_{yx}, \quad b := a, \quad B := \Sigma_{xy} \tag{75}$$

und erhalten für den betreffenden Eigenwert

$$\lambda_a = b^T B A a = a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a. \tag{76}$$

als Lösung (Maximum) des restringierten Optimierungsproblems

$$\max_b \left(a^T \Sigma_{xy} b \right)^2 \text{ u.d.N. } b^T \Sigma_{yy} b = 1 \tag{77}$$

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten (fortgeführt)

Schritt (2)

Basierend auf Schritt (1) verbleibt die Lösung des restringierten Optimierungsproblem

$$\phi_r^2 = \max_a a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a \text{ u.d.N. } a^T \Sigma_{xx} a = 1,\ a_i^T \Sigma_{xx} a = 0, i = 1, ..., r-1 \tag{78}$$

Dazu halten wir zunächst fest, dass (78) mit den Definitionen von α_i und K in der Definition der Kanonischen Koeffizientenvektoren, Variaten, und Korrelationen geschrieben werden kann als

$$\phi_r^2 = \max_\alpha \alpha^T K K^T \alpha \text{ u.d.N. } \alpha^T \alpha = 1,\ \alpha_i^T \alpha = 0, i = 1, ..., r-1, \tag{79}$$

denn

$$\begin{aligned}
&\phi_r^2 = \max_a a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a \text{ u.d.N. } a^T \Sigma_{xx} a = 1, a_i^T \Sigma_{xx} a = 0 \Leftrightarrow \\
&\phi_r^2 = \max_\alpha a^T \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} a \text{ u.d.N. } \alpha^T \Sigma_{xx}^{-1/2} \Sigma_{xx} \Sigma_{xx}^{-1/2} \alpha = 1, \alpha_i^T \Sigma_{xx}^{-1/2} \Sigma_{xx} \Sigma_{xx}^{-1/2} \alpha = 0 \\
&\phi_r^2 = \max_\alpha \alpha^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-1/2} \alpha \text{ u.d.N. } \alpha^T \alpha = 1, \alpha_i^T \alpha = 0 \qquad (80) \\
&\phi_r^2 = \max_\alpha \alpha^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \Sigma_{yy}^{-1/2} \Sigma_{yx} \Sigma_{xx}^{-1/2} \alpha \text{ u.d.N. } \alpha^T \alpha = 1, \alpha_i^T \alpha = 0 \\
&\phi_r^2 = \max_\alpha \alpha^T K K^T \alpha \text{ u.d.N. } \alpha^T \alpha = 1, \alpha_i^T \alpha = 0
\end{aligned}$$

Beweis des Theorems zu den Eigenschaften kanonischer Korrelationen und Variaten (fortgeführt)

Dabei sind nach der betreffenden Definition die α_i die Eigenvektoren von KK^T mit den $i = 1, ..., r-1$ größten Eigenwerten. Nach dem Theorem zur Maximierung quadratischer Formen mit Nebenbedingungen ist die Lösung von (79) der größte Eigenwert von KK^T mit seinem assoziierten Eigenvektor. Die Nebenbedingung $\alpha_i^T \alpha = 0$ schränkt diese Wahl auf den rt-größten Eigenwert und seinen assoziierten Eigenvektor α_r ein. Mit der Definition von Eigenwerten und Eigenvektoren gilt also

$$\phi_r^2 = \alpha_r^T KK^T \alpha_r = \alpha_r^T \lambda_r \alpha_r = \lambda_r \alpha_r^T \alpha_r = \lambda_r. \tag{81}$$

Wir haben also gezeigt, dass das restringierte Optimierungsproblem des Theorems den Maximumwert $\phi_r = \lambda_r^{1/2}$ hat. Es bleibt zu zeigen, dass dieser Maximumwert für a_r und b_r angenommen wird.

Schritt (3)

Einsetzen von a_r und b_r in $a^T \Sigma_{xy} b$ ergibt mit

$$K = A\Lambda B^T \Leftrightarrow KB = A\Lambda B^T B \Leftrightarrow KB = A\Lambda \Leftrightarrow K\beta_r = \alpha_r \lambda_r^{1/2} \tag{82}$$

dass

$$a_r^T \Sigma_{xy} b_r = \alpha_r^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} \beta_r = \alpha_r^T K \beta_r = \alpha_r^T \alpha_r \lambda_r^{1/2} = \rho_r \tag{83}$$

Also nimmt $a^T \Sigma_{xy} b$ bei a_r und b_r seinen restringierten Maximalwert λ_r an.

□

Referenzen

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed). Wiley-Interscience.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, *26*(2), 139–142. https://doi.org/10.1037/h0058165

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, *28*(3/4), 321. https://doi.org/10.2307/2333955

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Academic Press.

Uurtio, V., Monteiro, J. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., & Rousu, J. (2018). A Tutorial on Canonical Correlation Methods. *ACM Computing Surveys*, *50*(6), 1–33. https://doi.org/10.1145/3136624

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

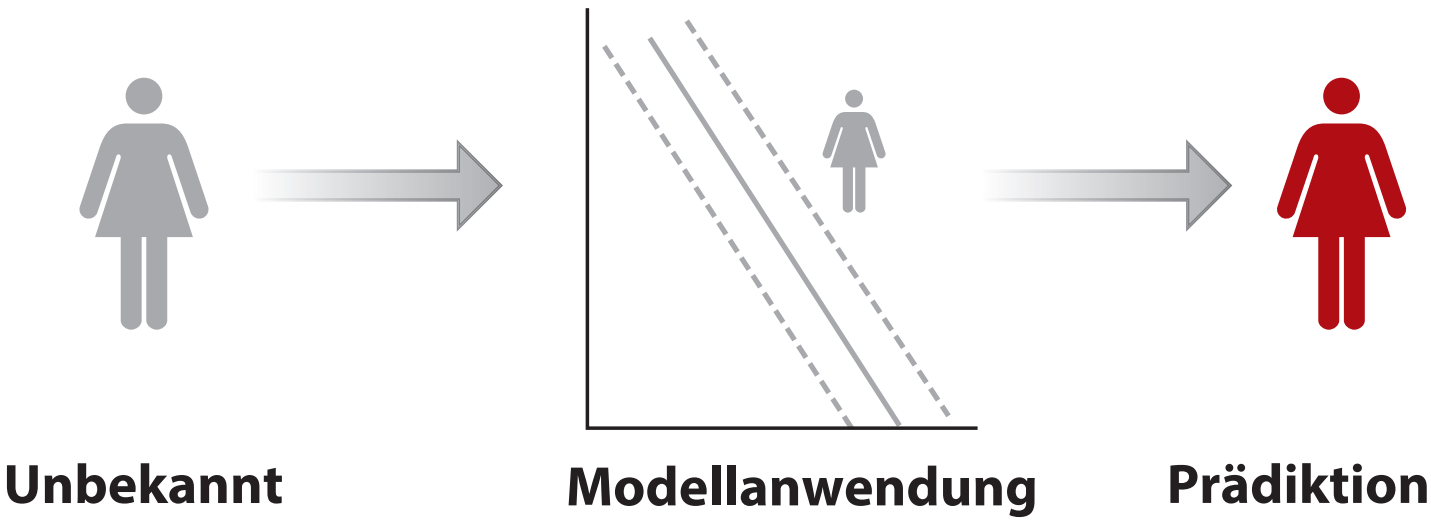
(8) Prädiktion und Kreuzvalidierung

Prädiktive Modellierung und Maschinelles Lernen

Kreuzvalidierung

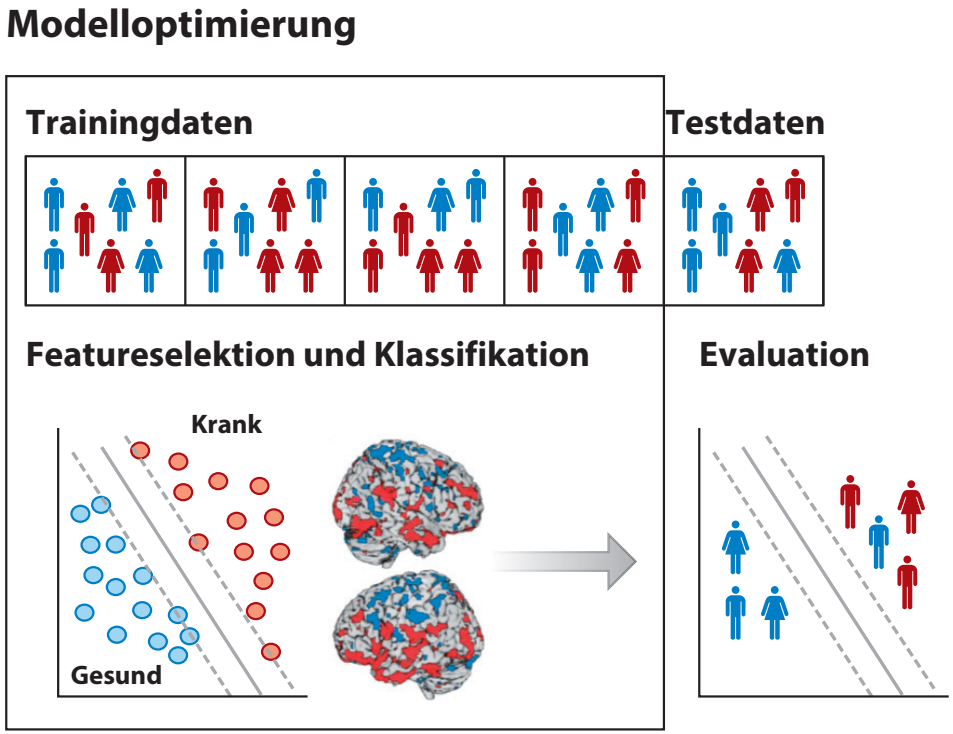
Selbstkontrollfragen

Rhetorik der Prädiktiven Modellierung und des Maschinellen Lernens



Dwyer, Falkai, and Koutsouleris (2018)

Rhetorik der Prädiktiven Modellierung und des Maschinellen Lernens



Dwyer, Falkai, and Koutsouleris (2018)

Daten	Trainingsdaten und Testdaten
Statistisches Modell	Modell, Machine Learning Algorithmus
Schätzen von Parametern	Trainieren des Modells, Parameterlernen, Supervised Learning

Definition (Binärer Klassifikationstrainingdatensatz)

Ein *binärer Klassifikationsdatensatz*

$$\mathcal{D} := \{(x_1, y_1), ..., (x_n, y_n)\} = \{(x_i, y_i)\}_{i=1}^{n} \tag{1}$$

ist eine Menge von n *Trainingsdatenpunkten*

$$(x_i, y_i) \text{ mit } x_i \in \mathbb{R}^m \text{ und } y_i \in \{0, 1\} \text{ for } i = 1, ..., n, \tag{2}$$

wobei x_i m-*dimensionaler Featurevektor* und y_i *Label* genannt wird. Üblicherweise werden die Trainingsdatenpunkte dabei als unabhängige und identische Realisierungen eines Zufallsvektors $m + 1$-dimensionalen Zufallsvektors $\zeta := (\xi, \upsilon)$ verstanden.

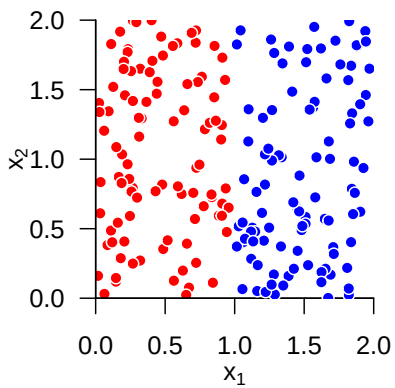
Bemerkungen

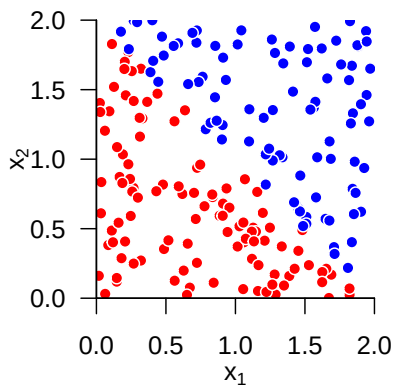
- $y_i \in \{0, 1\}$ bezeichnet die Klassenzugehörigkeit des Featurevektors $x_i \in \mathbb{R}^m$.
- Ein Beispiel für y_i ist “Kein Therapieerfolg” (0)” vs. “Therapieerfolg” (1).
- Beispiele für die m Komponenten der x_i sind Testscores, Biomarker, Soziodemographische Daten.

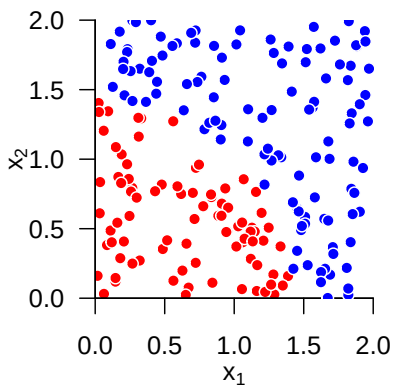
Beispiele bivariater Featureplotszenarien

$$x_i \in \mathbb{R}^2, y_i \in \{0, 1\}, i = 1, ..., n, \bullet\ y_i = 0, \quad \bullet\ y_i = 1$$

$$x_{i_1} > 1 \Leftrightarrow y_i = 1 \qquad x_{i_1} + x_{i_2} > 2 \Leftrightarrow y_i = 1 \qquad x_{i_1}^2 + x_{i_2}^2 > 2 \Leftrightarrow y_i = 1$$







Anwendung der prädiktiven Modellierung

Explanatorische Modellierung ⇔ Grundlagenforschung

Bestimmung von $\hat{\phi} \coloneqq \operatorname{argmin} \left\| \hat{\phi} - \phi \right\|$

Bestimmung von $\hat{f} \coloneqq \operatorname{argmin}_{f \in F} \| \upsilon - f(\xi) \|, F$ beliebig

Prädiktive Modellierung ⇔ Anwendungsorientierte Forschung

Shmueli (2010), Sainani (2014)

Prädiktive Modellierung und Maschinelles Lernen

Kreuzvalidierung

Selbstkontrollfragen

Workflow der prädiktiven Modellierung

Featureselektion

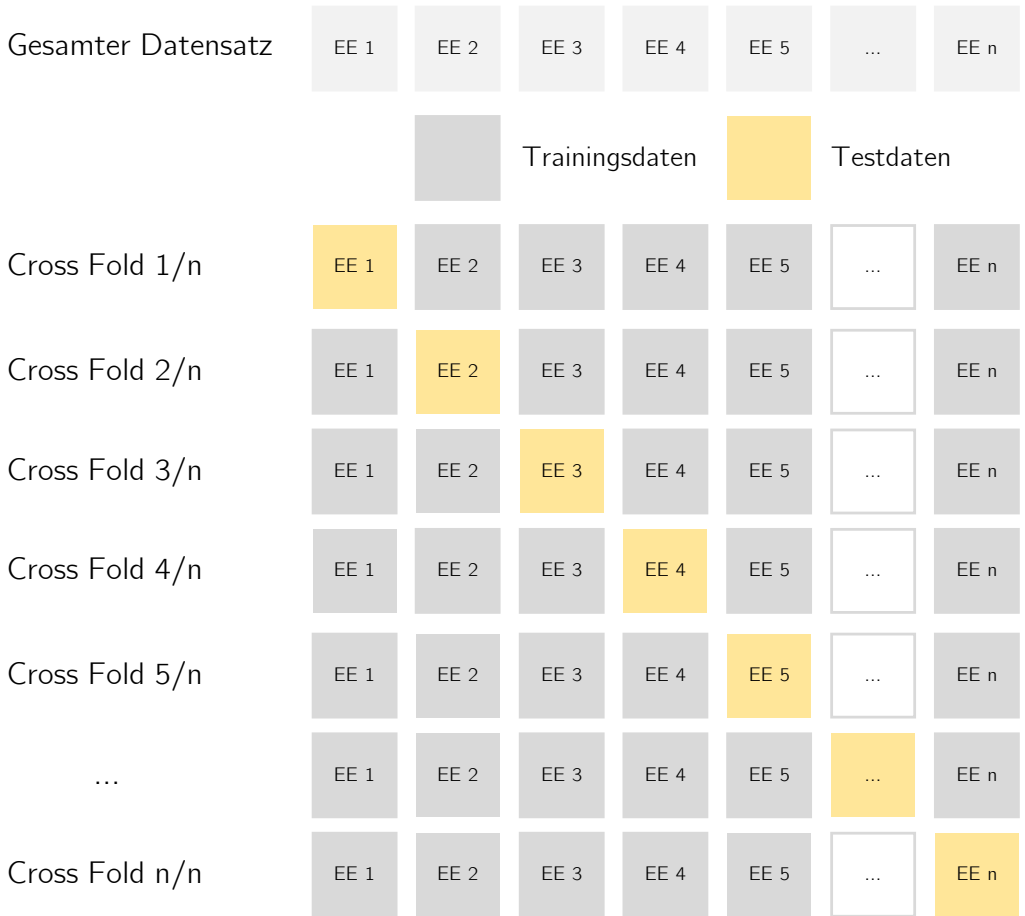
- Auswahl von möglichen prädiktiven Variablen
- Dimensionreduktion zur Verringerung des Curse of Dimensionality

Kreuzvalidierung

- Wiederholtes Trainieren und Testen eines Modells an einem Datensatz
- Einsatz zur Modelloptimierung
- Einsatz zur Messung der probabilistischen Assoziation von Featuren und Labeln

Kreuzvalidierung

Leave-One-Out-Crossvalidation (LOOCV) bei n experimentellen Einheiten (EE)



Kreuzvalidierung

Konfusionsmatrix bei LOOCV mit binärem Label für Testendatenpunkt (x_i, y_i)

Fall	Prädiktion $f(x_i) = 0$	Prädiktion $f(x_i) = 1$	
$y_i = 0$	Richtig Negativ r_n	Falsch Positiv f_p	Gesamt Negativ $r_n + f_p$
$y_i = 1$	Falsch Negativ f_n	Richtig Positiv r_p	Gesamt Positiv $f_n + r_p$
	Negative Prädiktion $r_n + f_n$	Positive Prädiktion $f_p + r_p$	

Exemplarische Performanzmaße bei LOOCV mit binärem Label

- Akkuratheit (Accuracy)

$$\text{ACC} = \frac{\text{Anzahl richtiger Prädiktionen}}{\text{Anzahl aller Prädiktionen}} = \frac{r_n + r_p}{r_n + r_p + f_n + f_p} \tag{3}$$

- Sensitivität (Richtig-positiv-Rate, True Positive Rate, Recall, Hit Rate)

$$\text{SEN} = \frac{\text{Anzahl richtiger Positivprädiktionen}}{\text{Anzahl positiver Fälle}} = \frac{r_p}{r_p + f_n} \tag{4}$$

- Spezifität (Richtig-negativ-Rate, True Negative Rate, Correct Rejection Rate)

$$\text{SPE} = \frac{\text{Anzahl richtiger Negativprädiktionen}}{\text{Anzahl negativer Fälle}} = \frac{r_n}{r_n + f_p} \tag{5}$$

Prädiktive Modellierung und Maschinelles Lernen

Kreuzvalidierung

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie die Rhetorik der Prädiktiven Modellierung.
2. Geben Sie die Definition eines binären Klassifikationstrainingdatensatzes wieder.
3. Erläutern Sie Unterschiede und Gemeinsamkeiten der explanatorischen und prädiktiven Modellierung.
4. Erläutern Sie die Idee der Leave-One-Out-Crossvalidation (LOOCV).
5. Erläutern Sie die Konfusionsmatrix bei LOOCV mit binärem Label.
6. Geben Sie die Definitionen von Akkuratheit, Sensitivität und Spezifität bei LOOCV wieder.

Referenzen

Dwyer, Dominic B., Peter Falkai, and Nikolaos Koutsouleris. 2018. "Machine Learning Approaches for Clinical Psychology and Psychiatry." *Annual Review of Clinical Psychology* 14 (1): 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Sainani, Kristin L. 2014. "Explanatory Versus Predictive Modeling." *PM&R* 6 (9): 841–44. https://doi.org/10.1016/j.pmrj.2014.08.941.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3). https://doi.org/10.1214/10-STS330.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(9) Hauptkomponentenanalyse

Hauptkomponentenanalyse = Principal Component Analysis (PCA).

PCA ist eine Featureselektionsmethode.

- "Features" sind die Komponenten multidimensionaler Zufallsvektoren.
- Korrelierte Features repräsentieren redundante Information.

PCA generiert ein korrelationsfreies Featureset durch lineare Featurekombination.

Die Durchführung einer PCA basiert auf

- einer *Orthonormalzerlegung* der *Stichprobenkovarianzmatrix* und
- einer anschließenden *Vektorkoordinatentransformation*.

PCA kannt zur *Kompression* hochdimensionaler Daten eingesetzt werden.

Vektorraumbasen und Vektorkoordinatentransformationen

Definition und Eigenschaften

Datenkompression

Selbstkontrollfragen

Vektorraumbasen und Vektorkoordinatentransformationen

Definition und Eigenschaften

Datenkompression

Selbstkontrollfragen

Definition (Linearkombination)

$\{v_1, v_2, ..., v_k\}$ sei eine Menge von k Vektoren eines Vektorraums V. Dann ist die *Linearkombination* der Vektoren in $v_1, v_2, ..., v_k$ mit den skalaren Koeffizienten $a_1, a_2, ..., a_k$ definiert als der Vektor

$$w := \sum_{i=1}^{k} a_i v_i \in V. \tag{1}$$

Bemerkung

- Als Beispiel seien in $\mathbb{R}^2$

$$v_1 := \begin{pmatrix} 2 \\ 1 \end{pmatrix}, v_2 := \begin{pmatrix} 1 \\ 1 \end{pmatrix}, v_3 := \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ und } a_1 := 2, a_2 := 3, a_3 := 0. \tag{2}$$

Dann ergibt sich

$$w = a_1 v_1 + a_2 v_2 + a_3 v_3 = 2 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} + 3 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0 \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 7 \\ 5 \end{pmatrix}. \tag{3}$$

Definition (Lineare Hülle und Aufspannen)

V sei ein Vektorraum und es sei $W := \{w_1, ..., w_k\} \subset V$. Dann ist die *lineare Hülle* von W definiert als die Menge aller Linearkombinationen der Elemente von W,

$$\text{Span}(W) := \left\{ \sum_{i=1}^{k} a_i w_i | a_1, ..., a_k \text{ sind skalare Koeffizienten} \right\} \tag{4}$$

Man sagt, dass eine Menge von Vektoren $W \subseteq V$ *einen Vektorraum* V *aufspannt*, wenn jedes $v \in V$ als eine Linearkombination von Vektoren in W geschrieben werden kann.

Bemerkungen

- Wenn eine Menge W von Vektoren einen Vektorraum aufspannt, dann kann jedes Element des Vektorraums durch eine Linearkombination der Elemente der Menge W gebildet werden.

Definition (Lineare Unabhängigkeit)

V sei ein Vektorraum. Eine Menge $W := \{w_1, w_2, ..., w_k\}$ von Vektoren in V heißt *linear unabhängig*, wenn die einzige Repräsentation des Nullelements $0 \in V$ durch eine Linearkombination der $w \in W$ die triviale Repräsentation

$$0 = a_1 w_1 + a_2 w_2 + \cdots + a_k w_k \text{ mit } a_1 = a_2 = \cdots = a_k = 0 \tag{5}$$

ist. Wenn die Menge W nicht linear unabhängig ist, dann heißt sie *linear abhängig*.

Bemerkungen

- Prinzipiell müsste man für jede Linearkombination der $w \in W$ prüfen, ob sie Null ist.
- Die beiden folgenden Theoreme zeigen, dass es auch einfacher geht.

Theorem (Lineare Abhängigkeit von zwei Vektoren)

V sei ein Vektorraum. Zwei Vektoren $v_1, v_2 \in V$ sind linear abhängig, wenn einer der Vektoren ein skalares Vielfaches des anderen Vektors ist.

Beweis

v_1 sei ein skalares Vielfaches von v_2, also

$$v_1 = \lambda v_2 \text{ mit } \lambda \neq 0. \tag{6}$$

Dann gilt

$$v_1 - \lambda v_2 = 0. \tag{7}$$

Dies aber entspricht der Linearkombination

$$a_1 v_1 + a_2 v_2 = 0 \tag{8}$$

mit $a_1 = 1 \neq 0$ und $a_2 = -\lambda \neq 0$. Es gibt also eine Linearkombination des Nullelementes, die nicht die triviale Repräsentation ist, und damit sind v_1 und v_2 nicht linear unabhängig.

Theorem (Lineare Abhängigkeit einer Menge von Vektoren)

V sei ein Vektorraum und $w_1, ..., w_k \in V$ sei eine Menge von Vektoren in V. Wenn einer der Vektoren w_i mit $i = 1, ..., k$ eine Linearkombination der anderen Vektoren ist, dann ist die Menge der Vektoren linear abhängig.

Beweis

Die Vektoren $w_1, \ldots, w_k$ sind genau dann linear abhängig, wenn gilt, dass $\sum_{i=1}^{k} a_i w_i = 0$ mit mindestens einem $a_i \neq 0$. Es sei also zum Beispiel $a_j \neq 0$. Dann gilt

$$0 = \sum_{i=1}^{k} a_i w_i = \sum_{i=1, i\neq j}^{k} a_i w_i + a_j w_j \tag{9}$$

Also folgt

$$a_j w_j = - \sum_{i=1, i\neq j}^{k} a_i w_i \tag{10}$$

und damit

$$w_j = -a_j^{-1} \sum_{i=1, i\neq j}^{k} a_i w_i = - \sum_{i=1, i\neq j}^{k} (a_j^{-1} a_i) w_i \tag{11}$$

Also ist w_j eine Linearkombination der w_i für $i = 1, \ldots, k$ mit $i \neq j$. □

Theorem (Orthonormalität und lineare Unabhängigkeit in $\mathbb{R}^m$)

$q_1, ..., q_m$ seien orthonormale Vektoren in $\mathbb{R}^m$, es gelte also

$$q_i^T q_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \text{ für } 1 \leq i, j \leq m. \tag{12}$$

Dann sind die $q_1, ..., q_m$ linear unabhängig.

Beweis

Für $i = 1, \ldots, m$ gilt, dass

$$\begin{aligned} a_1 q_1 + a_2 q_2 + \cdots + a_m q_m &= 0_m \\ \Leftrightarrow (a_1 q_1 + a_2 q_2 + \cdots + a_m q_m)^T &= 0_m^T \\ \Leftrightarrow (a_1 q_1 + a_2 q_2 + \cdots + a_m q_m)^T q_i &= 0_m^T q_i \\ \Leftrightarrow \sum_{j=1}^{m} a_j q_j^T q_i &= 0 \\ \Leftrightarrow a_i &= 0. \end{aligned} \tag{13}$$

Die einzige Repräsentation des Nullelements 0_m durch eine Linearkombination der $q_1, \ldots, q_m$ ist also die triviale Repräsentation und die $q_1, \ldots, q_m$ sind linear unabhängig.

Definition (Basis)

V sei ein Vektorraum und es sei $B \subseteq V$. Dann heißt B eine *Basis von* V, wenn

- die Vektoren in B den Vektorraum V aufspannen und
- die Vektoren in B linear unabhängig sind.

Bemerkung

- Vektorräume haben in der Regel unendlich viele Basen.
- Wir zeigen unten, dass die Forderung der linearen Unabhängigkeit der Vektoren einer Basis impliziert, dass die Darstellung eines Vektors in V durch die Linearkombination der Vektoren in B eindeutig ist.

Theorem (Eigenschaften von Basen)

V sei ein Vektorraum. Dann gelten folgende Eigenschaften für Basen von V.

- Alle Basen von V haben die gleiche Kardinalität und diese wird die *Dimension* von V genannt.
- Jede Menge von m linear unabhängigen Vektoren ist Basis eines m-dimensionalen Vektorraums.

Bemerkung

- Wir verzichten auf einen Beweis des sehr tiefen Theorems.

Definition (Basisdarstellung und Koordinaten)

$B := \{b_1, ..., b_m\}$ sei eine Basis eines m-dimensionalen Vektorraumes V und es sei $v \in V$. Dann heißt die Linearkombination

$$v = \sum_{i=1}^{m} c_i b_i \tag{14}$$

die *Darstellung von v bezüglich der Basis B* und die Koeffizienten $c_1, ..., c_m$ heißen die *Koordinaten von v bezüglich der Basis B*.

Theorem (Eindeutigkeit der Basisdarstellung)

Die Basisdarstellung eines $v \in V$ bezüglich einer Basis B ist eindeutig.

Beweis

Ohne Beschränkung der Allgemeinheit nehmen wir an, dass der Vektorraum von Dimension m ist. Nehmen wir an, dass zwei Darstellungen von v bezüglich der Basis B existieren, also dass

$$\begin{aligned} v &= a_1 b_1 + \cdots + a_m b_m \\ v &= c_1 b_1 + \cdots + c_m b_m \end{aligned} \tag{15}$$

Subtraktion der unteren von der oberen Gleichung ergibt

$$0 = (a_1 - c_1)b_1 + \cdots + (a_m - c_m)b_m \tag{16}$$

Weil die $b_1, \ldots, b_m$ linear unabhängig sind, gilt aber, dass $(a_i - c_i) = 0$ für alle $i = 1, \ldots, m$ und somit sind die beiden Darstellungen von v bezüglich der Basis B identisch.

□

Bemerkung

- Die lineare Unabhängigkeit von Basisvektoren garantiert also, dass ein Vektor zu einer gegebenen Basis nur eine Darstellung, also insbesondere nur ein einziges Set an Koordinaten $c_1, \ldots, c_m$ hat und nicht etwa mehrere.

Definition (Orthonormalbasis von $\mathbb{R}^m$)

Eine Basis $B := \{q_1, ..., q_m\}$ von $\mathbb{R}^m$ heißt *Orthonormalbasis* von $\mathbb{R}^m$, wenn B eine Menge orthonormaler Vektoren ist, wenn also gilt, dass

$$q_i^T q_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \text{ für } 1 \leq i, j \leq m. \tag{17}$$

Bemerkungen

- Bei einer Orthonormalbasis von $\mathbb{R}^m$ sind die Basisvektoren orthonormal.
- Man beachte, dass eine Basis von $\mathbb{R}^m$ nicht aus orthonormalen Vektoren bestehen muss.

Beispiel (Orthonormalbasis von $\mathbb{R}^2$)

Es ist

$$B_1 := \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\} \tag{18}$$

eine Orthonormalbasis von $\mathbb{R}^2$, denn B_1 ist eine Basis und besteht aus orthonormalen Vektoren. Dabei erkennt man die Orthonormalität der Elemente von B_1 zunächst an

$$\begin{aligned} \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= 1 \cdot 1 + 0 \cdot 0 = 1 + 0 = 1 \\ \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= 0 \cdot 0 + 1 \cdot 1 = 0 + 1 = 1 \\ \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= 1 \cdot 0 + 0 \cdot 1 = 0 + 0 = 0 \end{aligned} \tag{19}$$

Da die Orthonormalität zweier Vektoren in $\mathbb{R}^2$ ihre lineare Unabhängkeit impliziert, sind die Elemente von B_1 auch linear unabhängig. Wir wollen schließlich noch nachweisen, dass die Elemente von B_1 den Vektorraum $\mathbb{R}^2$ aufspannen, dass also jedes $v \in \mathbb{R}^2$ durch eine Linearkombination der Elemente von B_1 geschrieben werden kann. Dazu sei $v = (v_1, v_2)^T \in \mathbb{R}^m$ beliebig. Dann aber gilt

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ für } c_1 := v_1 \text{ und } c_2 := v_2 \tag{20}$$

und damit ist B_1 eine Orthonormalbasis von $\mathbb{R}^2$.

Theorem (Kanonische Basis von $\mathbb{R}^m$)

Es sei

$$B := \left\{ e_1, ..., e_m | e_{i_j} = 1 \text{ für } i = j \text{ und } e_{i_j} = 0 \text{ für } i \neq j \right\} \subset \mathbb{R}^m. \tag{21}$$

Dann ist B eine Orthonormalbasis von $\mathbb{R}^m$ und wird die *kanonische Basis* von $\mathbb{R}^m$ genannt.

Beweis

Im Sinne der Definition einer Basis müssen wir zeigen, dass die Elemente von B linear unabhängig sind. Dies folgt aber direkt aus der Orthonormalität der Elemente von B. Weiterhin müssen wir zeigen, dass die Elemente von B den Vektorraum $\mathbb{R}^m$ aufspannen, dass also jeder Vektor $v \in \mathbb{R}^m$ als Linearkombination der Elemente von B geschrieben werden kann. Sei also $v \in \mathbb{R}^m$ beliebig. Dann kann $v := (v_1, ..., v_m)^T$ geschrieben werden als

$$v = \sum_{i=1}^{m} c_i e_i \text{ mit } c_i := v_i. \tag{22}$$

□

Bemerkungen

- Die kanonische Basis von $\mathbb{R}^3$ ist zum Beispiel $B := \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$.

Beispiel (Eine nichtkanonische Orthonormalbasis von $\mathbb{R}^2$)

Neben B_1 ist auch

$$B_2 := \left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\} \tag{23}$$

eine Orthonormalbasis von $\mathbb{R}^2$, denn B_2 ist ebenfalls eine Basis von $\mathbb{R}^2$ und besteht ebenfalls aus orthonormalen Vektoren. Dabei erkennt man die Orthonormalität der Elemente von B_2 zunächst an

$$\begin{aligned} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} &= \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} + \frac{1}{2} = 1 \\ \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} &= \left(-\frac{1}{\sqrt{2}} \right) \cdot \left(-\frac{1}{\sqrt{2}} \right) + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{1}{2} + \frac{1}{2} = 1 \\ \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} &= -\frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = -\frac{1}{2} + \frac{1}{2} = 0. \end{aligned} \tag{24}$$

Da die Orthonormalität zweier Vektoren in $\mathbb{R}^2$ ihre lineare Unabhängkeit impliziert, sind die Elemente von B_2 auch linear unabhängig. Es bleibt also nachzuweisen, dass die Elemente von B den Vektorraum $\mathbb{R}^2$ aufspannen, dass also jedes $v \in \mathbb{R}^2$ durch eine Linearkombination der Elemente von B_2 geschrieben werden kann.

Beispiel (Eine nichtkanonische Orthonormalbasis von $\mathbb{R}^2$ fortgeführt)

Sei also $v \in \mathbb{R}^m$ beliebig. Dann gilt, dass

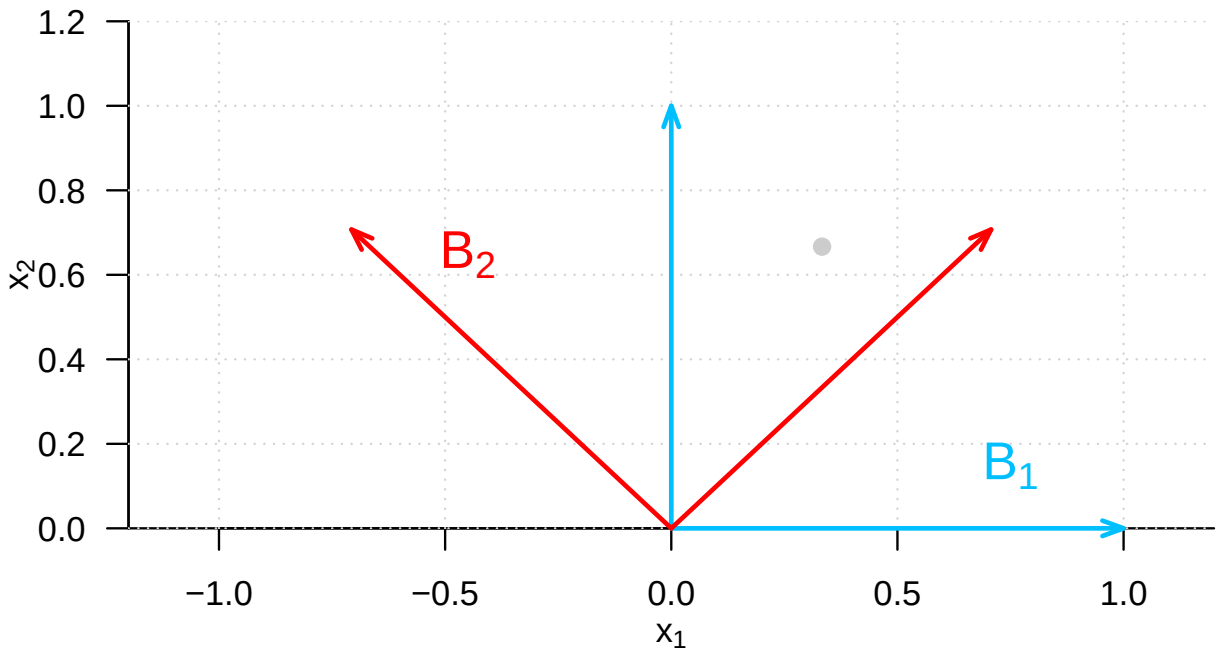
$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = c_1 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} + c_2 \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \text{ für } c_1 := \frac{\sqrt{2}}{2}(v_1 + v_2) \text{ und } c_2 := \frac{\sqrt{2}}{2}(v_2 - v_1), \tag{25}$$

denn dann ist

$$\begin{aligned} c_1 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} + c_2 \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} &= \frac{\sqrt{2}}{2}(v_1 + v_2) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} + \frac{\sqrt{2}}{2}(v_2 - v_1) \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \\ &= \frac{1}{2}(v_1 + v_2) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2}(v_2 - v_1) \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2}(v_1 + v_2) - \frac{1}{2}(v_2 - v_1) \\ \frac{1}{2}(v_1 + v_2) + \frac{1}{2}(v_2 - v_1) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2}v_1 + \frac{1}{2}v_2 - \frac{1}{2}v_2 + \frac{1}{2}v_1 \\ \frac{1}{2}v_1 + \frac{1}{2}v_2 + \frac{1}{2}v_2 - \frac{1}{2}v_1 \end{pmatrix} \\ &= \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \\ &= v. \end{aligned} \tag{26}$$

Damit ist B_2 eine Orthonormalbasis von $\mathbb{R}^2$.

Kanonische und nichtkanonische Orthonormalbasen B_1 und B_2 von $\mathbb{R}^2$



Im Rahmen der Hauptkomponentenanalyse sind wir daran interessiert, basierend auf den Koordinaten eines Vektors bezüglich einer Basis die Koordinaten desselben Vektors bezüglich einer anderen Basis zu berechnen. Dazu führen wir im Folgenden die Begriffe der *Orthogonalprojektion*, der *Vektorkoordinaten bezüglich einer Orthogonalbasis* und der *Vektorkoordinatentransformation* ein.

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

$$\tilde{x} = aq \text{ mit } a := \frac{q^T x}{q^T q}, \tag{27}$$

wobei der Skalar a *Projektionsfaktor* genannt wird.

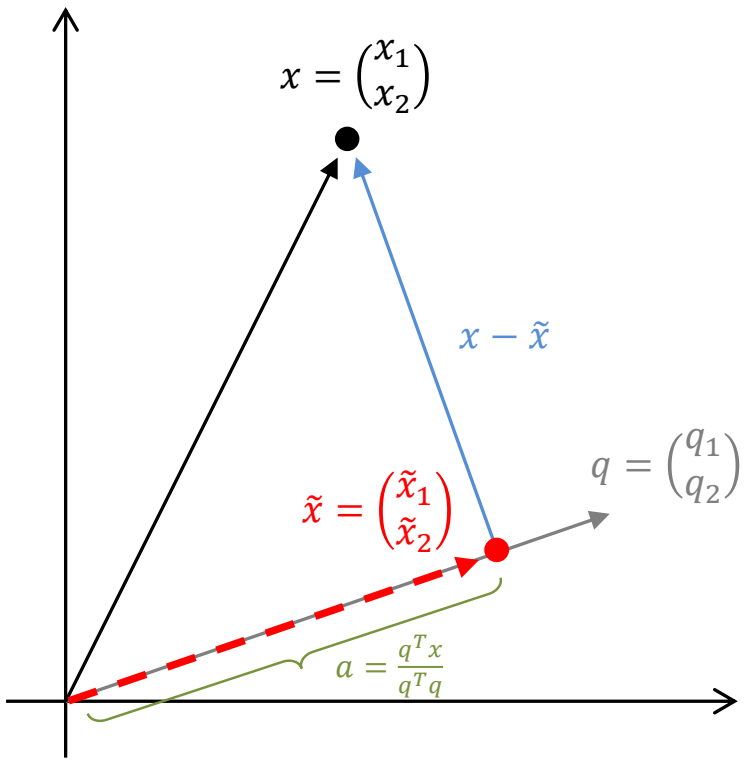
Bemerkungen

- 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

$$q^T(x - \tilde{x}) = 0 \Leftrightarrow q^T(x - aq) = 0 \Leftrightarrow q^T x - aq^T q = 0 \Leftrightarrow a = \frac{q^T x}{q^T q}. \tag{28}$$

- Wenn q die Länge $||q|| = \sqrt{q^T q} = 1$ hat, dann gilt $a = \frac{q^T x}{||q||^2} = q^T x$.
- Die Orthogonalprojektion von x auf q wird manchmal auch einfach als *Projektion von* x *auf* q bezeichnet

Orthogonalprojektion



Theorem (Vektorkoordinaten bezüglich einer Orthogonalbasis)

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

$$c_i = x^T q_i \tag{29}$$

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

$$x = \sum_{i=1}^{m} (x^T q_i) q_i . \tag{30}$$

Beweis

Für $i = 1, ..., m$ gilt

$$x = \sum_{j=1}^{m} c_j q_j \Leftrightarrow q_i^T x = q_i^T \sum_{j=1}^{m} c_j q_j \Leftrightarrow q_i^T x = \sum_{j=1}^{m} c_j q_i^T q_j \Leftrightarrow q_i^T x = c_i \Leftrightarrow c_i = x^T q_i . \tag{31}$$

Bemerkung

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

Theorem (Vektorkoordinatentransformation)

$B_v := \{v_1, ..., v_m\}$ und $B_w := \{w_1, ..., w_m\}$ 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 $x_i, i = 1, ..., m$ eines Vektors x bezüglich der Basis B_v in die Koordinaten $\tilde{x}_1, ..., \tilde{x}_m$ des Vektors bezüglich der Basis B_w durch

$$\tilde{x} = A^T x \tag{32}$$

transformiert werden. Analog können die Koordinaten $\tilde{x}_1, ..., \tilde{x}_m$ des Vektors hinsichtlich der Basis B_w in die Koordinaten $x_1, ..., x_m$ des Vektors hinsichtlich B_v durch

$$x = A\tilde{x}. \tag{33}$$

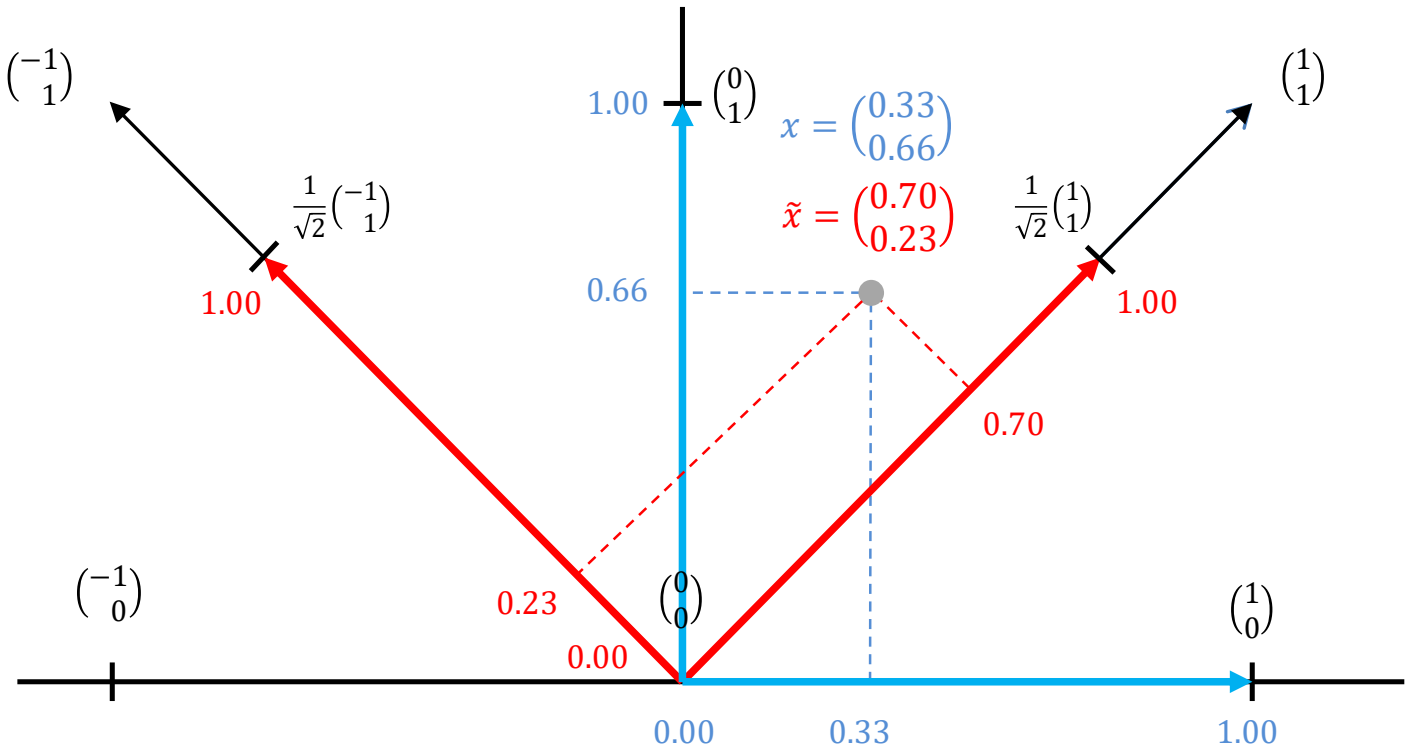
transformiert werden.

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

$$B_w := \left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\} \tag{34}$$

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

$$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = a_{11} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + a_{21} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ und } \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = a_{12} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + a_{22} \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{35}$$

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

$$a_{11} = \frac{1}{\sqrt{2}}, a_{21} = \frac{1}{\sqrt{2}}, a_{12} = -\frac{1}{\sqrt{2}}, a_{22} = \frac{1}{\sqrt{2}}. \tag{36}$$

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

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{37}$$

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

$$\tilde{x} = A^T x = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix} \approx \begin{pmatrix} 0.70 \\ 0.23 \end{pmatrix}. \tag{38}$$

Vektorraumbasen und Vektorkoordinatentransformationen

Definition und Eigenschaften

Datenkompression

Selbstkontrollfragen

Definition (Hauptkomponentenanalyse)

$\mathbb{C}(\xi)$ sei die Kovarianzmatrix eines m-dimensionalen Zufallsvektors ξ. Dann heißt die Orthonormalzerlegung

$$\mathbb{C}(\xi) = Q\Lambda Q^T, \tag{39}$$

wobei

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

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

$$\tilde{\xi} = Q^T\xi \tag{40}$$

heißt *Hauptkomponenten-transformierter oder Hauptkomponenten-projizierter Zufallsvektor*.

Bemerkungen

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

Theorem (Basiseigenschaften der Hauptkomponentenanalyse)

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

$$\mathbb{C}(\xi) = Q \Lambda Q^T, \tag{41}$$

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

(1) Die Spalten von Q, also die Hauptkomponenten von $\mathbb{C}(\xi)$, bilden eine Orthonormalbasis von $\mathbb{R}^m$.

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

Bemerkungen

- Werte von ξ werden üblicherweise zunächst hinsichtlich der kanonischen Basis von $\mathbb{R}^m$ verstanden.
- Die Hauptkomponentenanalyse resultiert in einer alternativen Orthonormalbasis für möglichen Werte von ξ.
- $\tilde{\xi}$ entspricht den Koordinaten von ξ bezüglich dieser alternativen Orthonormalbasis.

Beweis

(1) Mit dem Theorem zu den Eigenschaften von Basen gilt, dass jede Menge von m linear unabhängigen Vektoren Basis eines m-dimensionalen Vektorraums ist. Weiterhin gilt mit dem Theorem zu Orthonormalität und linearer Unabhängigkeit, dass m orthonormale Vektoren in $\mathbb{R}^m$ auch m linear unabhängige Vektoren in $\mathbb{R}^m$ sind. Damit sind die Spalten von Q m linear unabhängige Vektoren in $\mathbb{R}^m$ und folglich eine Basis von $\mathbb{R}^m$.

(2) Wir betrachten das Theorem zur Vektorkoordinatentransformation 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

$$q_i = \sum_{j=1}^{m} (q_i^T e_j) e_j = \sum_{j=1}^{m} q_{i_j} e_j = q_i. \tag{42}$$

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 dem Theorem zur Vektorkoordinatentransformation direkt, dass der Hauptkomponenten-projizierte Zufallsvektor

$$\tilde{\xi} = Q^T \xi \tag{43}$$

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

Theorem (Kovarianzeigenschaften der Hauptkomponentenanalyse)

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

$$\mathbb{C}(\xi) = Q \Lambda Q^T, \tag{44}$$

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

(1) Die Kovarianzmatrix des Hauptkomponenten-transformierten Zufallsvektors ist die Diagonalmatrix Λ, es gilt also insbesondere $\mathbb{V}(\tilde{\xi}_i) = \lambda_i$ für $i = 1, \ldots, m$ und $\mathbb{C}(\tilde{\xi}_i, \tilde{\xi}_j) = 0$ für $i \neq j, 1 \leq i, j \leq m$.

(2) Die Summen der Varianzen der Komponenten von ξ und der Komponenten von $\tilde{\xi}$ sind identisch,

$$\sum_{i=1}^{m} \mathbb{V}(\xi_i) = \sum_{i=1}^{m} \mathbb{V}(\tilde{\xi}_i). \tag{45}$$

Bemerkungen

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

  $$\mathbb{V}(\tilde{\xi}_1) > \mathbb{V}(\tilde{\xi}_2) > \cdots > \mathbb{V}(\tilde{\xi}_m) \Leftrightarrow \mathbb{V}(q_1^T \xi) > \mathbb{V}(q_2^T \xi) > \cdots > \mathbb{V}(q_m^T \xi). \tag{46}$$

  und $q_1^T \xi$ maximiert die Varianz der unkorrelierten Linearkombinationen der Komponenten von ξ mit orthonormalen Koeffizientenvektoren.
- Die paarweise nicht-identischen Kovarianzen der Komponenten von $\tilde{\xi}$ sind Null, die Komponenten von $\tilde{\xi}$ sind also unkorreliert und repräsentieren keine redundante Information.
- Die Gesamtvarianz der Komponenten von ξ und $\tilde{\xi}$ sind identisch, wobei Gesamtvarianz hier im Sinne der Summe der Varianzen der jeweiligen Komponenten zu verstehen ist.

Beweis

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

$$\mathbb{C}(\xi) = Q\Lambda Q^T \Leftrightarrow Q^T\mathbb{C}(\xi)Q = Q^TQ\Lambda Q^TQ \Leftrightarrow Q^T\mathbb{C}(\xi)Q = \Lambda. \tag{47}$$

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

$$\mathbb{C}(\xi) = \mathbb{E}\left((\xi - \mathbb{E}(\xi))\,(\xi - \mathbb{E}(\xi))^T\right) = \mathbb{E}\left(\xi\xi^T\right). \tag{48}$$

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

$$\begin{aligned}
\mathbb{C}(\tilde{\xi}) &= \mathbb{E}\left(\left(\tilde{\xi} - \mathbb{E}(\tilde{\xi})\right)\left(\tilde{\xi} - \mathbb{E}(\tilde{\xi})\right)^T\right) \\
&= \mathbb{E}\left(\left(Q^T\xi - \mathbb{E}(Q^T\xi)\right)\left(Q^T\xi - \mathbb{E}(Q^T\xi)\right)^T\right) \\
&= \mathbb{E}\left(\left(Q^T\xi - Q^T\mathbb{E}(\xi)\right)\left(Q^T\xi - Q^T\mathbb{E}(\xi)\right)^T\right) \\
&= \mathbb{E}\left((Q^T\xi)(Q^T\xi)^T\right) = Q^T\mathbb{E}\left(\xi\xi^T\right)Q = Q^T\mathbb{C}(\xi)Q = \Lambda.
\end{aligned} \tag{49}$$

(2) Wir erinnern daran, dass schon für jede symmetrische Matrix mit verschiedenen Eigenwerten, also auch für positiv-semidefinite Matrizen mit verschiedenen Eigenwerten die Spur gleich der Summe der Eigenwerte ist. Also gilt hier insbesondere für die Eigenwerte $\lambda_1, \ldots, \lambda_m$ von $\mathbb{C}(\xi)$

$$\sum_{i=1}^{m}\mathbb{V}(\xi_i) = \mathsf{tr}(\mathbb{C}(\xi)) = \sum_{i=1}^{m}\lambda_i = \sum_{i=1}^{m}\mathbb{V}(\tilde{\xi}_i) \tag{50}$$

Theorem (Projektionseigenschaften der ersten Hauptkomponente)

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

$$\mathbb{C}(\xi) = Q \Lambda Q^T, \tag{51}$$

die Hauptkomponentenanalyse von $\mathbb{C}(\xi)$. $q_1 \in \mathbb{R}^m$ sei die erste Spalte von Q, also der erste Eigenvektor von $\mathbb{C}(\xi)$, mit zugehörigen Diagonalelementen λ_1 von Λ, also dem größten Eigenwert von $\mathbb{C}(\xi)$. Dann gelten

(1) Unter allen Orthogonalprojektion von ξ auf Vektoren $v \in \mathbb{R}^m$ der Länge $\|v\| = 1$ hat der Projektionsfaktor

$$\alpha := q_1^T \xi \tag{52}$$

der Orthogonalprojektion von ξ auf q_1 die größte Varianz.

(2) Unter allen Orthogonalprojektionen $\tilde{\xi}_v := (v^T \xi) v$ von ξ auf Vektoren $v \in \mathbb{R}^m$ der Länge $\|v\| = 1$ mit *erwartetem quadratischem Fehler*

$$E(\xi_v) := \mathbb{E}\left(\|\xi - \tilde{\xi}_v\|^2\right) \tag{53}$$

hat die Orthogonalprojektion von ξ auf die erste Hauptkomponente q_1, $\tilde{\xi}_1 := (q_1^T \xi) q$ den geringsten erwarteten quadratischen Fehler.

Bemerkungen

- Man sagt auch oft etwas ungenau, dass die erste Hauptkomponente/der erste Eigenvektor “die Varianz maximiert und den (Rekonstruktions)fehler minimiert”.

Beweis

(1) Die erste Aussage entspricht der Lösung des restringierten Optimierungsproblems

$$\max_{v \in \mathbb{R}^m} \mathbb{V}\left(v^T \xi\right) \text{ unter der Nebenbedingung } v^T v = 1. \tag{54}$$

Zu seiner Lösung geben wir zunächst eine Repräsentation der Varianz des Projektionsfaktors $v^T \xi$ durch die Kovarianzmatrix von ξ an und lösen das Problem dann analytisch mithilfe seiner Lagrangefunktion.

Es gilt zunächst aufgrund von $\xi = 0_m$ und damit $\mathbb{C}(\xi) = \mathbb{E}\left(\xi\xi^T\right)$, dass

$$\begin{aligned}
\mathbb{V}(\alpha) &= \mathbb{V}\left(v^T \xi\right) \\
&= \mathbb{E}\left(\left(v^T \xi - \mathbb{E}(v^T \xi)\right)^2\right) \\
&= \mathbb{E}\left(\left(v^T \xi - v^T \mathbb{E}(\xi)\right)^2\right) \\
&= \mathbb{E}\left(\left(v^T \xi\right)^2\right) \\
&= \mathbb{E}\left((v^T \xi)(v^T \xi)^T\right) \\
&= \mathbb{E}\left(v \xi \xi^T v\right) \\
&= v^T \mathbb{E}\left(\xi\xi^T\right) v \\
&= v^T \mathbb{C}\left(\xi\right) v
\end{aligned} \tag{55}$$

Beweis (fortgeführt)

Das restringierte Optimierungsproblem hat also die Form

$$\max_{v \in \mathbb{R}^m} v^T \mathbb{C}(\xi)\, v \text{ unter der Nebenbedingung } v^T v = 1. \tag{56}$$

Die Lagrangefunktion dieses Problems hat entsprechend die Formel

$$L : \mathbb{R}^m \times \mathbb{R}, (v, l) \mapsto L(v, l) := v^T \mathbb{C}(\xi)\, v - l(v^T v - 1) \tag{57}$$

wobei wir mit $l \in \mathbb{R}$ den Lagrangemultiplikator für die Nebenbedingung $v^T v - 1 = 0$ bezeichnen. Pragmatisches Ableiten der Lagrangefunktion hinsichtlich v und Nullsetzen ergibt dann

$$\begin{aligned} \frac{\partial}{\partial v} L(v, l) &= 0 \\ \Leftrightarrow \frac{\partial}{\partial v} \left(v^T \mathbb{C}(\xi)\, v - l(v^T v - 1) \right) &= 0 \\ \Leftrightarrow 2\mathbb{C}(\xi)\, v - 2lv &= 0 \\ \Leftrightarrow \mathbb{C}(\xi)\, v &= lv. \end{aligned} \tag{58}$$

Also ist an der Stelle der notwendigen Bedingung für ein Maximum von L, dass v ein Eigenvektor von $\mathbb{C}(\xi)$ mit Eigenwert l ist. Weiterhin gilt an der Stelle der notwendigen Bedingung für ein Maximum aufgrund der Nebenbedingung, dass

$$\mathbb{C}(\xi)\, v = lv \Leftrightarrow v^T \mathbb{C}(\xi)\, v = v^T lv \Leftrightarrow \mathbb{V}\left(v^T \xi\right) = lv^T v \Leftrightarrow \mathbb{V}\left(v^T \xi\right) = l \tag{59}$$

Der größte Eigenwert von $\mathbb{C}(\xi)$ ist aber λ_1 und der entsprechende Eigenvektor, der das Optimierungsproblem löst damit q_1.

Beweis (fortgeführt)

(2) Die zweite Aussage entspricht der Lösung des restringierten Optimierungsproblems

$$\min_{v \in \mathbb{R}^m} E\left(\xi_v\right) \text{ unter der Nebenbedingung } v^T v = 1. \tag{60}$$

Zu seiner Lösung geben wir zunächst eine Repräsentation der erwarteten quadratischen Fehlers mithilfe der Kovarianzmatrix von ξ an führen die Lösung dann auf die Lösung des unter (1) betrachteten Problems zurück. Es gilt aufgrund von $\xi = 0_m$ und damit $\mathbb{C}(\xi) = \mathbb{E}\left(\xi\xi^T\right)$, dass

$$\begin{aligned}
E(\xi_v) &= \mathbb{E}\left(\|\xi - \tilde{\xi}_v\|^2\right) \\
&= \mathbb{E}\left((\xi - \tilde{\xi}_v)^T(\xi - \tilde{\xi}_v)\right) \\
&= \mathbb{E}\left(\xi^T\xi - 2\xi^T\tilde{\xi}_v + \tilde{\xi}_v^T\tilde{\xi}_v\right) \\
&= \mathbb{E}\left(\xi^T\xi - 2\xi^T(v^T\xi)v + ((v^T\xi)v)^T(v^T\xi)v\right) \\
&= \mathbb{E}\left(\xi^T\xi - 2(v^T\xi)(\xi^T v) + (v^T(\xi^T v))(v^T\xi)v\right) \\
&= \mathbb{E}\left(\xi^T\xi - 2(v^T\xi)(\xi^T v) + (v^T\xi)(\xi^T v)(v^T v)\right) \\
&= \mathbb{E}\left(\xi^T\xi - 2(v^T\xi)(\xi^T v) + (v^T\xi)(\xi^T v)\right) \\
&= \mathbb{E}\left(\xi^T\xi - (v^T\xi)(\xi^T v)\right) \\
&= \mathbb{E}\left(\xi^T\xi - v^T(\xi\xi^T)v\right) \\
&= \mathbb{E}\left(\xi^T\xi\right) - v^T\mathbb{E}\left(\xi\xi^T\right)v \\
&= \mathbb{E}\left(\xi^T\xi\right) - v^T\mathbb{C}(\xi)v
\end{aligned} \tag{61}$$

Beweis (fortgeführt)

Mit $\mathbb{E}\left(\xi^T\xi\right) \geq 0$ wird das Minimum von

$$E(\xi_v) = \mathbb{E}\left(\xi^T\xi\right) - v^T\mathbb{C}(\xi)v \text{ unter der Nebenbedingung } v^Tv = 1 \tag{62}$$

dann aber für das Maximum von

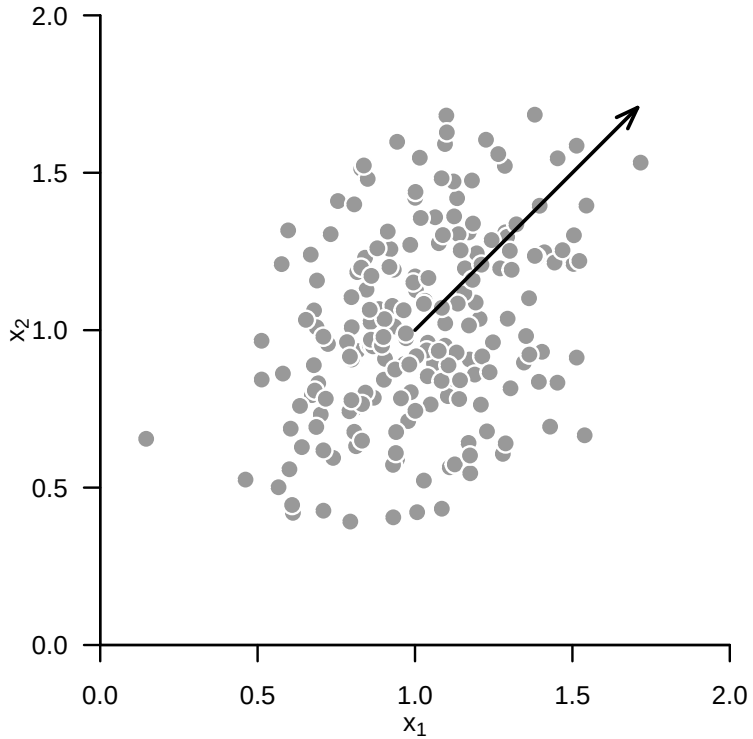
$$\mathbb{V}\left(v^T\xi\right) \text{ unter der Nebenbedingung } v^Tv = 1. \tag{63}$$

angenommen und ist durch Wahl von $v = q_1$ gegeben.

Definition und Eigenschaften

Visualisierung der Projektionseigenschaften der ersten Hauptkomponente für

$m = 2$, $\xi \sim N(\mu, \Sigma)$, • Realisierungen von ξ, $\rightarrow$ Erste Hauptkomponente q_1



Vektorraumbasen und Vektorkoordinatentransformationen

Definition und Eigenschaften

Datenkompression

Selbstkontrollfragen

Überblick

- Datenkompression entspricht der Reduktion der Dimensionalität m eines Datensatzes.
- Ziel der Datenkompression im Rahmen der Datenpräprozessierung bei prädiktiver Modellierung ist es, dem Undersampling hochdimensionaler Datenräume entgegenzuwirken.
- Inspiriert von den Eigenschaften der Hauptkomponentenanalyse möchte man mit dem unten skizzierten Verfahren den Fehler zwischen einem Datensatz und einem dimensionsreduziertem Datensatz bei gleichzeitiger Retention maximal variabler Datenaspekte minimieren.
- Prinzipiell gibt es viele weitere Möglichkeiten der Datenkompression.

Definition (Hauptkomponentenanalyse eines Datensatzes)

$C \in \mathbb{R}^{m \times m}$ sei die Stichprobenkovarianzmatrix eines Datensatzes $X \in \mathbb{R}^{m \times n}$ bestehend aus unabhängigen Realisierungen $x_j \in \mathbb{R}^m$ mit $j = 1, \ldots, n$ eines m-dimensionalen Zufallsvektors ξ. Dann heißt die Orthonormalzerlegung

$$C = Q \Lambda Q^T \tag{64}$$

wobei

- $Q \in \mathbb{R}^{m \times m}$ die spaltenweise Konkatenation der Eigenvektoren von C 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. Für $j = 1, \ldots, n$ heißt der m-dimensionale Vektor

$$\tilde{x}_j := Q^T x_j \tag{65}$$

Hauptkomponenten-transformierter oder *Hauptkomponenten-projizierter Datenvektor* und der $m \times n$-dimensionale Datensatz

$$\tilde{X} := Q^T X \tag{66}$$

heißt *Hauptkomponenten-transformierter* oder *Hauptkomponenten-projizierter Datensatz*.

Bemerkungen

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

Theorem (Basiseigenschaften der HKA eines Datensatzes)

$C \in \mathbb{R}^{m \times m}$ sei die Stichprobenkovarianzmatrix eines Datensatzes $X \in \mathbb{R}^{m \times n}$, es sei $\bar{X} = 0_m$ und es sei

$$C = Q \Lambda Q^T, \tag{67}$$

die Hauptkomponentenanalyse von C. Dann gelten:

(1) Die Spalten von Q, also die Hauptkomponenten von C, bilden eine Orthonormalbasis von $\mathbb{R}^m$.

(2) Multiplikation mit Q^T transformiert die Koordinaten der Datenvektoren $x_j, j = 1, \ldots, n$ bezüglich der kanonischen Basis von $\mathbb{R}^m$ in Koordinaten bezüglich der Hauptkomponenten von C.

Bemerkung

- Ein Beweis ergibt sich analog zum Beweis des Theorems zu den Basiseigenschaften der HKA.

Theorem (Stichprobenkovarianzeigenschaften der HKA)

$C \in \mathbb{R}^{m \times m}$ sei die Kovarianzmatrix eines Datensatzes $X \in \mathbb{R}^{m \times n}$, es sei $\bar{X} = 0_m$ und es sei

$$C = Q \Lambda Q^T, \tag{68}$$

die Hauptkomponentenanalyse von C. Dann gelten

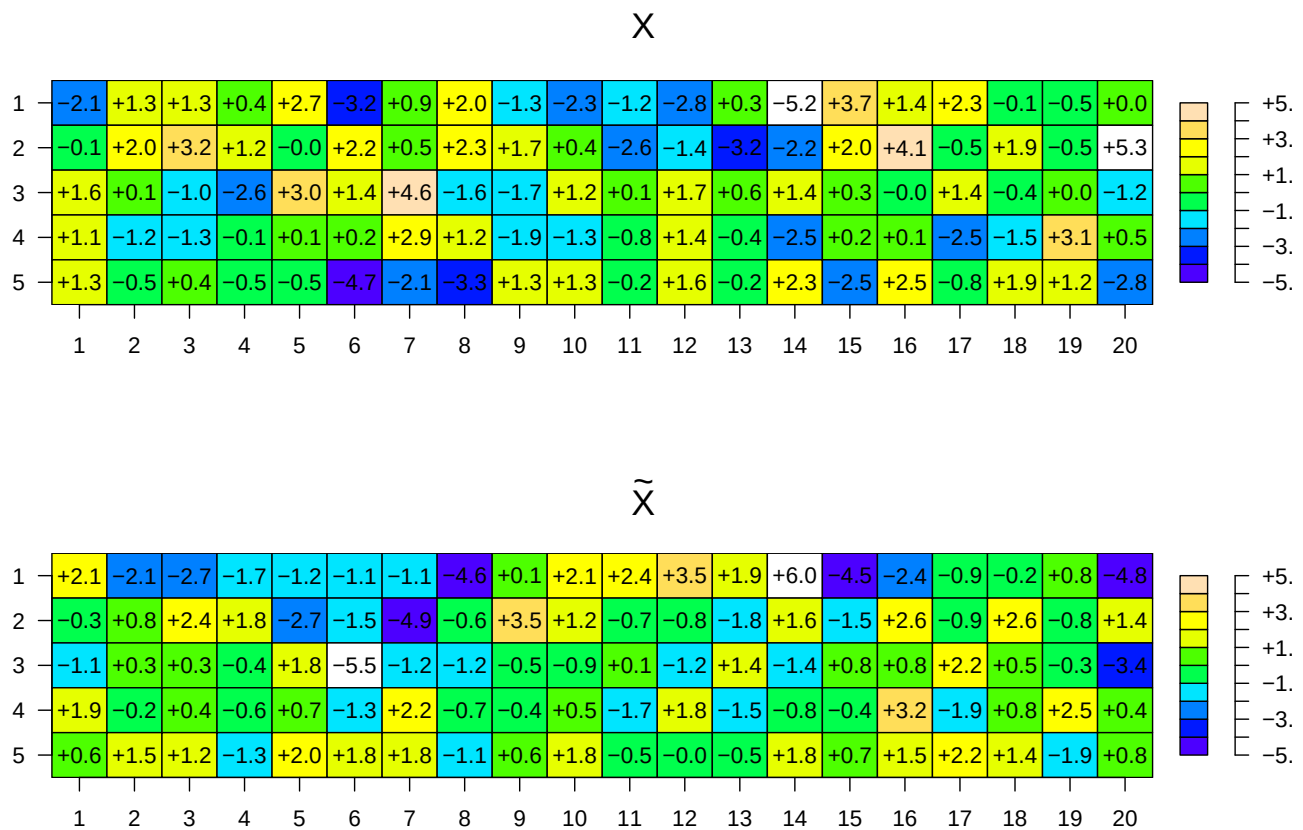
(1) Die Kovarianzmatrix des Hauptkomponenten-transformierten Datensatzes ist die Diagonalmatrix Λ. Es gilt also insbesondere, dass die Stichprobenvarianz der $i$ten Komponente der $\tilde{x}_j$ gegeben ist durch λ_i für $i = 1, \ldots m$ und dass die Stichprobenkovarianz der $i$ten und $k$ten Komponenten der $\tilde{x}_j$ gleich 0 ist.

(2) Die Summen der Stichprobenvarianzen der Komponenten der x_j und der Komponenten der $\tilde{x}_j$ sind identisch.

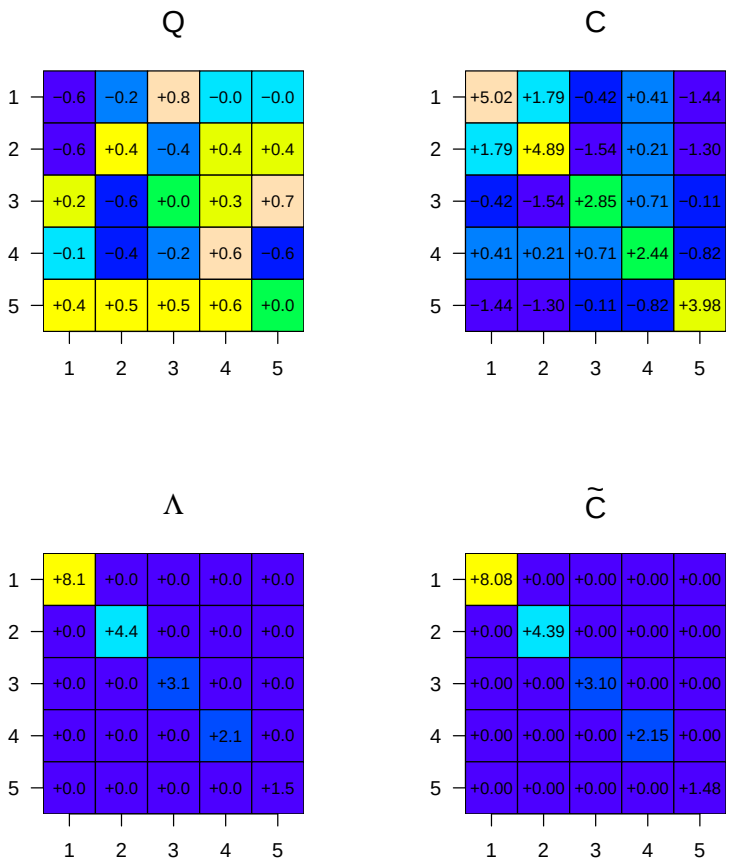
Bemerkung

- Ein Beweis ergibt sich analog zum Beweis des Theorems zu den Kovarianzeigenschaften der HKA.

Hauptkomponentenanalyse eines Datensatzes mit $m = 5$ und $n = 20$



Hauptkomponentenanalyse eines Datensatzes mit $m = 5$ und $n = 20$



Die Summe der Diagonalemente von C und $\tilde{C}$ ist hier 19.2.

Definition (Dimensionsreduzierter transformierte Datensatz)

$X \in \mathbb{R}^{m \times n}$ sei ein Datensatz, C sei seine Stichprobenkovarianzmatrix,

$$C = Q \Lambda Q^T \tag{69}$$

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 nennt man

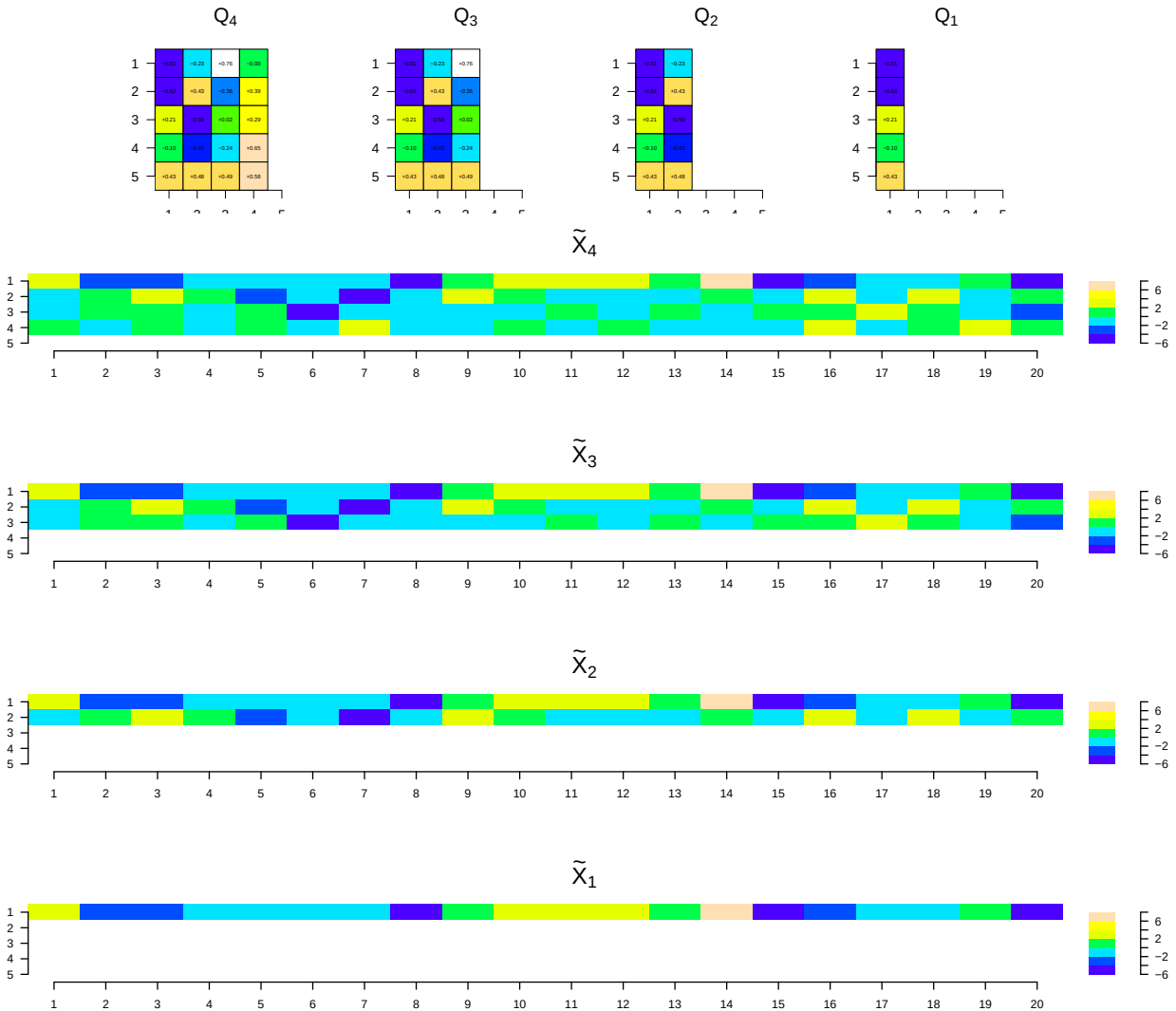
$$\tilde{X}_k = Q_k^T X \in \mathbb{R}^{k \times n} \tag{70}$$

einen *dimensionsreduzierten transformierten Datensatz*.

Bemerkung

- Die Definition ist durch das Theorem zu den Projektionseigenschaften der HKA motiviert.
- Man kann ein analoges Theorem für die HKA eines Datensatzes formulieren und beweisen.
- $\tilde{X}_k = Q_k^T X$ entspricht einer $(k \times n) = (k \times m) \cdot (m \times n)$ Matrixmultiplikation.
- $\tilde{X}_k$ ist der Datensatz, der aus $\tilde{X}$ durch Streichen der $(k+1)$-ten bis m-ten Zeile ensteht.

PCA-dimensionsreduzierte Datensätze



Definition (Rekonstruierter Datensatz, Datenrekonstruktionsfehler)

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

$$\tilde{X}_k = Q_k^T X \in \mathbb{R}^{k \times n} \tag{71}$$

ein dimensionsreduzierter Datensatz. Dann heißt

$$X_k = Q_k \tilde{X}_k \in \mathbb{R}^{m \times n} \tag{72}$$

rekonstruierter Datensatz und

$$e = \|\text{vec}(X - X_k)\|^2 \geq 0 \tag{73}$$

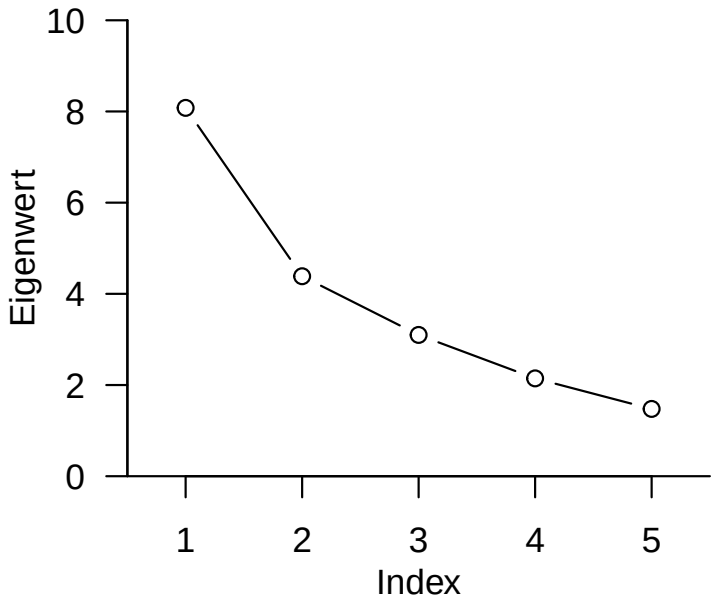
heißt *Datenrekonstruktionsfehler*.

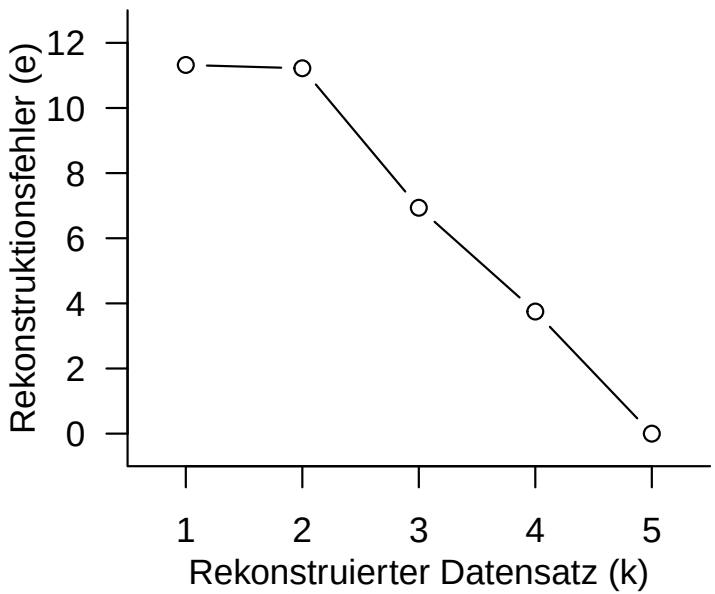
Bemerkungen

- $X_k = Q_k \tilde{X}_k$ 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 $Q\tilde{X}_k = QQ^T X = X$ und damit $e = 0$.
- Der Datenrekonstruktionsfehler ist das Stichprobenanalogon zum erwarteten quadratischen Fehler.

Datensatzrekonstruktion und Daterekonstruktionsfehler

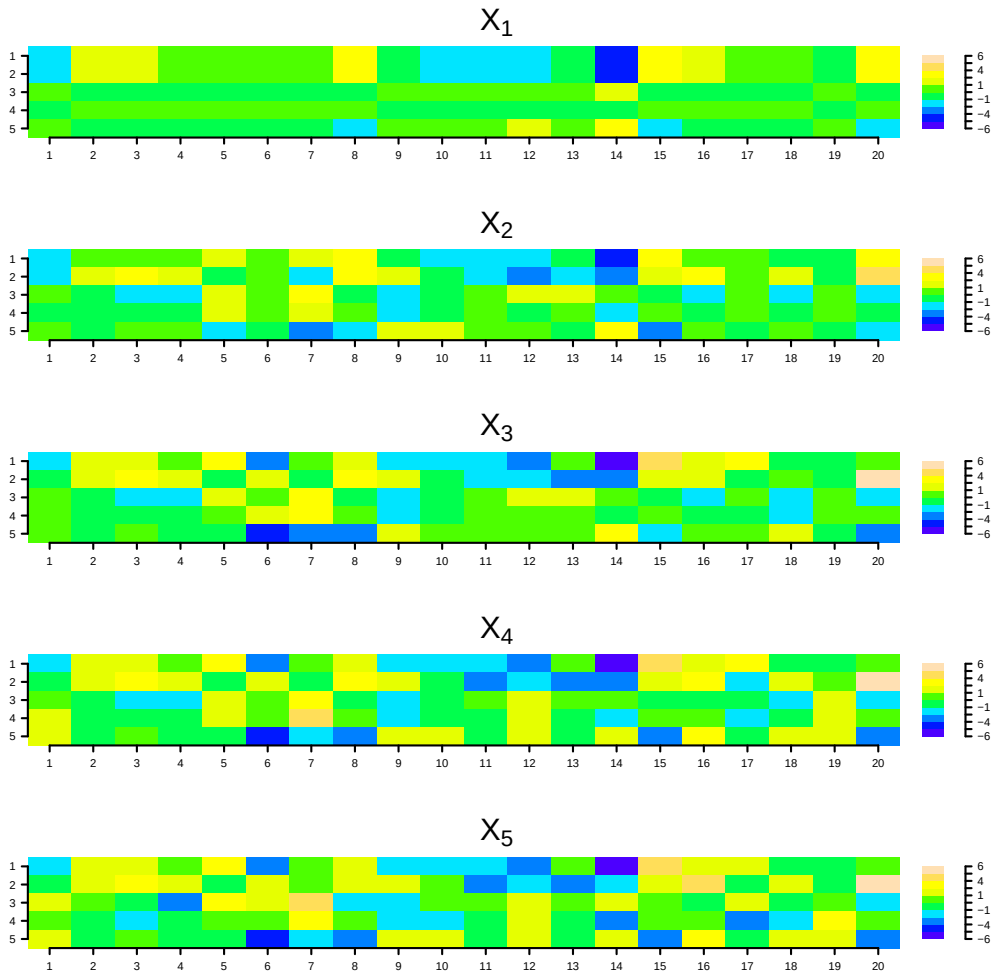
Eigenwerte (“Scree-Plot”) und Rekonstruktionsfehler





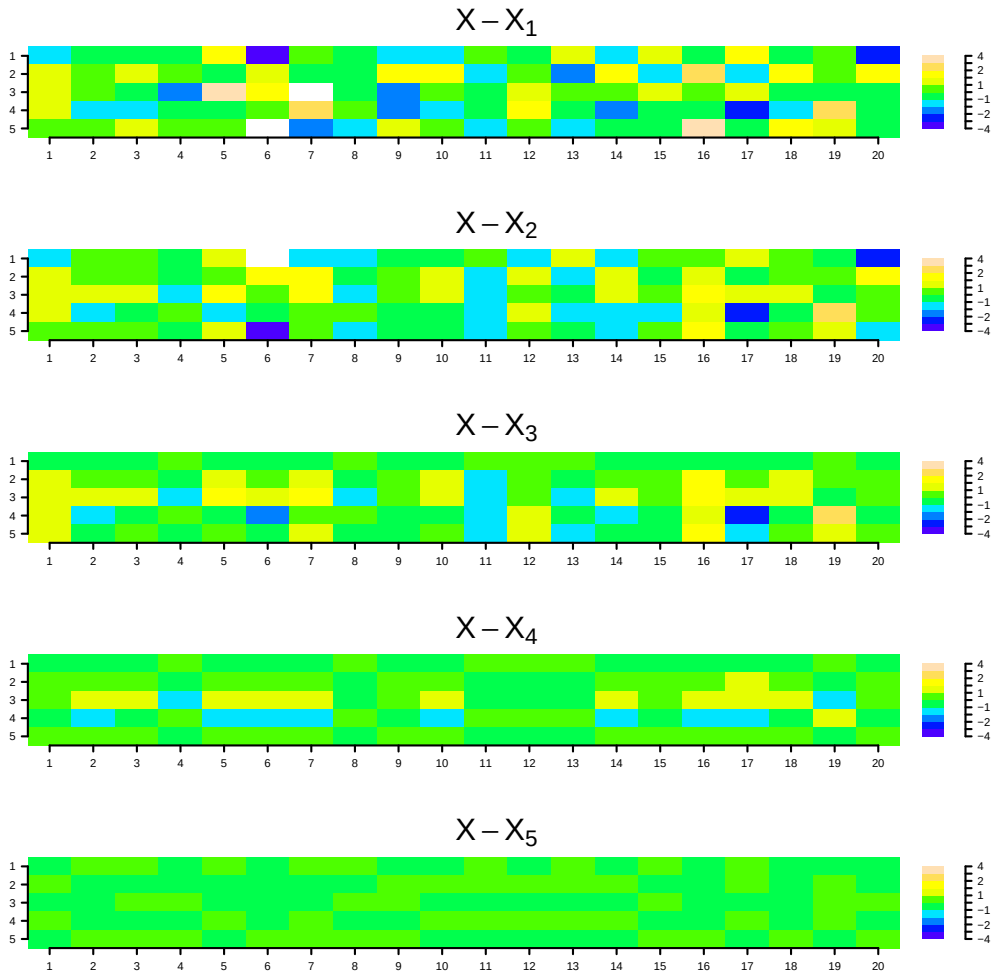
Datensatzrekonstruktion und Daterekonstruktionsfehler

Rekonstruierte Datensätze



Datensatzrekonstruktion und Daterekonstruktionsfehler

Rekonstruktionsfehler



Vektorraumbasen und Vektorkoordinatentransformationen

Definition und Eigenschaften

Datenkompression

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition der Linearen Unabhängigkeit von Vektoren wieder.
2. Geben Sie das Theorem zur Linearen Abhängigkeit von zwei Vektoren wieder.
3. Geben Sie die Definition der Basis eines Vektorraums wieder.
4. Geben Sie die Definition von Basisdarstellung und Koordinaten wieder.
5. Geben Sie das Theorem zur kanonischen Basis von $\mathbb{R}^m$ wieder.
6. Geben Sie die Definition des Begriffs der Orthogonalprojektion wieder.
7. Geben Sie das Theorem zu Vektorkoordinaten bezüglich einer Orthogonalbasis wieder.
8. Geben Sie das Vektorkoordinatentransformationstheorem wieder.
9. Erläutern Sie das Vektorkoordinatentransformationstheorem.
10. Geben Sie die Definition der Hauptkomponentenanalyse wieder.
11. Geben Sie das Theorem zu den Basiseigenschaften der Hauptkomponentenanalyse wieder.
12. Geben Sie das Theorem zu den Kovarianzeigenschaften der Hauptkomponentenanalyse wieder.
13. Geben Sie das Theorem zu den Projektionseigenschaften der ersten Hauptkomponente wieder.
14. Erläutern Sie das Prinzip der Datenkompression durch Hauptkomponentenanalyse.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(10) Lineare Diskriminanzanalyse

Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie Non-Response-Rate wird auf etwa 20 - 30% geschätzt

Vorhersage von Behandlungserfolg basierend auf klinischen Markern wäre hilfreich

- Therapieauswahloptimierung
- Lebensqualitätverbesserung
- Ressourcensensitivität

Digitale Datenbank von Psychotherapieverläufen als Trainingsdatensatz

Prädiktive Modellierung zur Etablierung eines prädiktiven klinischen Markerprofils

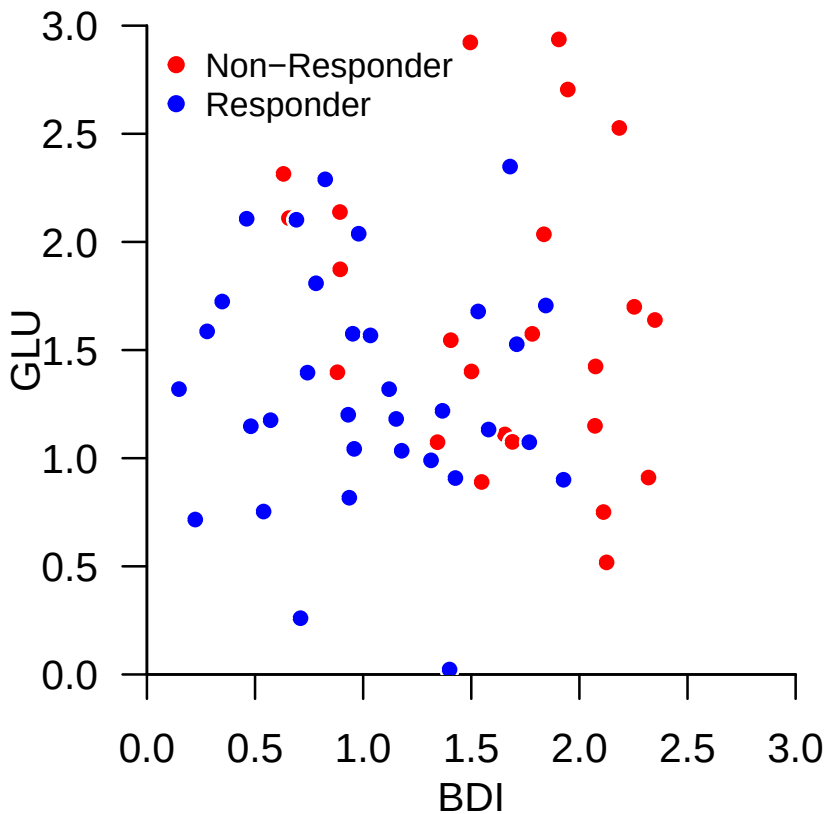
Treatmentsuccessvorhersage für neue Patient:innen

Anwendungsbeispiele

- BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg
- Lineare Diskriminanzanalyse, Logistische Regression, Neuronale Netze

Beispieldatensatz

- BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg



Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Definition (Bernoulli-Zufallsvariable)

Es sei ξ eine Zufallsvariable mit Ergebnisraum $\mathcal{X} := \{0, 1\}$ und WMF

$$p : \mathcal{X} \to [0, 1], x \mapsto p(x) := \mu^x (1 - \mu)^{1-x} \text{ mit } \mu \in [0, 1]. \tag{1}$$

Dann sagen wir, dass ξ einer *Bernoulli-Verteilung mit Parameter* $\mu \in [0, 1]$ unterliegt und nennen ξ eine *Bernoulli-Zufallsvariable*. Wir kürzen dies mit $\xi \sim \text{Bern}(\mu)$ ab. Die WMF einer Bernoulli-Zufallsvariable bezeichnen wir mit

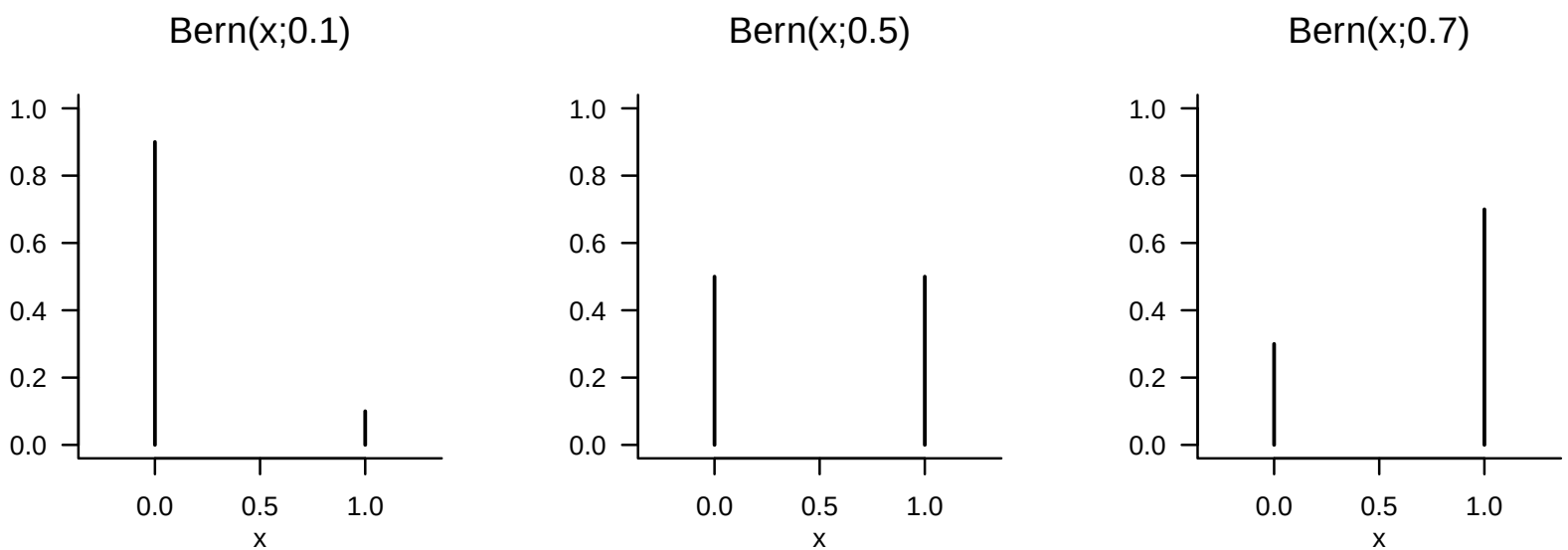
$$\text{Bern}(x; \mu) := \mu^x (1 - \mu)^{1-x}. \tag{2}$$

Bemerkungen

- Eine Bernoulli-Zufallsvariable kann als Modell eines Münzwurfs dienen.
- μ ist die Wahrscheinlichkeit dafür, dass ξ den Wert 1 annimmt,

$$\mathbb{P}(\xi = 1) = \mu^1 (1 - \mu)^{1-1} = \mu. \tag{3}$$

Bernoulli-Zufallsvariable



Definition (Modell der Linearen Diskriminanzanalyse)

ξ sei ein m-dimensionaler Zufallsvektor mit Ergebnisraum $\mathbb{R}^m$ und υ sei eine Zufallsvariable mit Ergebnisraum $\{0, 1\}$. Dann ist das *Modell der Linearen Diskriminanzanalyse* die gemeinsame Verteilung

$$\mathbb{P}(\xi, \upsilon) = \mathbb{P}(\upsilon)\mathbb{P}(\xi|\upsilon), \tag{4}$$

wobei die diskrete marginale Verteilung $\mathbb{P}(\upsilon)$ durch die WMF

$$p(y) = \mathsf{Bern}(y; \mu) \tag{5}$$

mit $\mu \in]0, 1[$ und die kontinuierliche bedingte Verteilung $\mathbb{P}(\xi|\upsilon)$ durch die WDF

$$p(x|y) = N(x; \mu_0, \Sigma)^{1-y} N(x; \mu_1, \Sigma)^y \tag{6}$$

mit $\mu_0, \mu_1 \in \mathbb{R}^m$ und $\Sigma \in \mathbb{R}^{m \times m}$ pd definiert ist. Wir bezeichnen die gemischte WMF und WDF (WMDF) des Modells der Linearen Diskriminanzanalyse mit

$$p(x, y) := p(y)p(x|y) = \mathsf{Bern}(y; \mu) N(x; \mu_0, \Sigma)^{1-y} N(x; \mu_1, \Sigma)^y \tag{7}$$

Modellformulierung

Bemerkung

Aus generativer Sicht wird ein Trainingsdatensatz

$$\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n} \text{ mit } x^{(i)} \in \mathbb{R}^m \text{ und } y^{(i)} \in \{0,1\} \tag{8}$$

eines Modells zur Linearen Diskriminanzanalyse wie folgt erzeugt:

(1) $y^{(i)}$ wird zunächst durch Ziehen aus einer Bernoulliverteilung mit Parameter μ erzeugt.

(2) In Abhängigkeit vom Wert von $y^{(i)}$ wird $x^{(i)}$ dann durch Ziehen aus einer multivariaten Normalverteilung mit Kovarianzmatrixparameter Σ und Erwartungswertparameter μ_0 für $y^{(i)} = 0$ oder μ_1 für $y^{(i)} = 1$ erzeugt.

Datengeneration

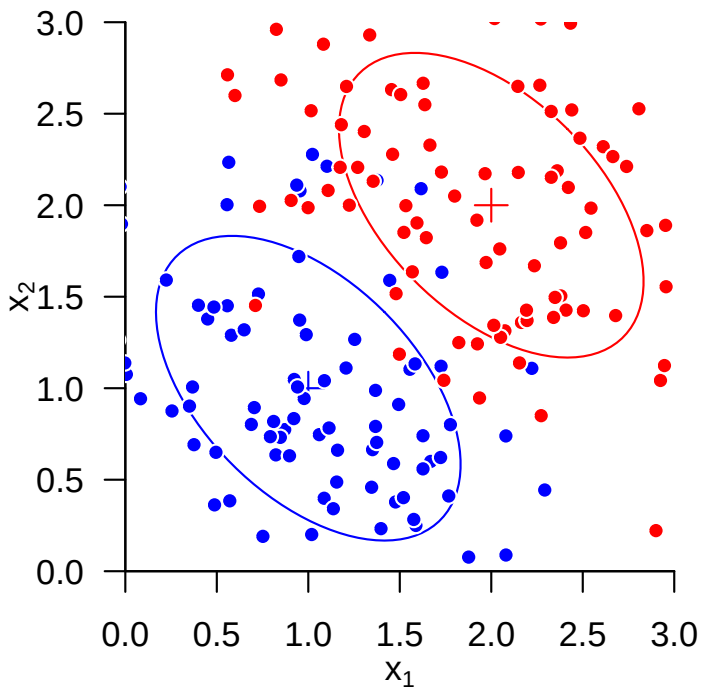
```
# Modellformulierung
library(mvtnorm)                                  # Multivariate Normalverteilung
set.seed(0)                                       # Zufallszahlengenerator
m      = 2                                        # Featurevektordimension
n      = 2e2                                      # Anzahl Trainingsdatenpunkte
mu     = 0.5                                      # wahrer, aber unbekannter, Bernoulliparameter \mu
mu_0   = c(1,1)                                   # wahrer, aber unbekannter, Normalverteilungsparameter \mu_0
mu_1   = c(2,2)                                   # wahrer, aber unbekannter, Normalverteilungsparameter \mu_1
Sigma  = matrix(c( 0.50, -0.25,                   # Kovarianzmatrixparameter
                  -0.25,  0.50),
                byrow = TRUE,
                nrow = m)

# Modellsampling
y      = matrix(rep(NaN,n)  , nrow = 1)           # Labeldatenarray
x      = matrix(rep(NaN,n*m), nrow = m)           # Featurevektorarray
for(i in 1:n){                                    # Iteration über Datenpunkte
  y[i]  = rbinom(1,1,mu)                          # y^{(i)} \sim Bern(\mu)
  x[,i] = ((rmvnorm(1, mu_0, Sigma)**(1-y[i]))    # x^{i} \sim N(\mu_0,\Sigma)^{1-y}N(\mu_1,\Sigma)^y
          *(rmvnorm(1, mu_1, Sigma)**(y[i])))
}

# Datensatzkonkatenation
D = rbind(x,y)
```

Modellformulierung

Datengeneration

$$m = 2, n = 200, \mu = 0.5, \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.50 & -0.10 \\ -0.10 & 0.50 \end{pmatrix}$$

$+$ μ_0 • $x^{(i)}$ mit $y^{(i)} = 0$, $+$ μ_1 • $x^{(i)}$ mit $y^{(i)} = 1, i = 1, \ldots, n$



Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Theorem (Inferenz bei Linearer Diskriminanzanalyse)

$p(x, y)$ sei die WMDF des Modells einer Linearen Diskriminanzanalyse. Dann gilt

$$p(y = 1|x) = \frac{1}{1 + \exp(-\tilde{x}^T \beta)} \text{ und } p(y = 0|x) = 1 - p(y = 1|x), \tag{9}$$

wobei

$$\tilde{x} := \begin{pmatrix} 1 \\ x \end{pmatrix} \in \mathbb{R}^{m+1} \tag{10}$$

der *erweiterten Featurevektor* und

$$\beta := \begin{pmatrix} \frac{1}{2} \left(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1 \right) + \ln \left(\frac{\mu}{1-\mu} \right) \\ -\Sigma^{-1} (\mu_0 - \mu_1) \end{pmatrix} \in \mathbb{R}^{m+1}. \tag{11}$$

der *Inferenzparametervektor* sind.

Bemerkungen

- $p(y|x)$ kann zur Prädiktion der Klasse eines $x \in \mathbb{R}^m$ genutzt werden.
- Diese Prädiktion hängt von den Parameter $\mu, \mu_0, \mu_1, \Sigma$ des Modells der Linearen Diskriminanzanalyse ab.

Beweis

Wir halten zunächst fest, dass

$$\begin{aligned} p(y=1|x) &= \frac{p(x,y=1)}{p(x,y=0)+p(x,y=1)} \\ &= \frac{\frac{p(x,y=1)}{p(x,y=1)}}{\frac{p(x,y=0)}{p(x,y=1)}+\frac{p(x,y=1)}{p(x,y=1)}} \\ &= \frac{1}{1+\frac{p(x,y=0)}{p(x,y=1)}} \\ &= \frac{1}{1+\exp\left(\ln\left(\frac{p(x,y=0)}{p(x,y=1)}\right)\right)} \\ &= \frac{1}{1+\exp\left(-\ln\left(\frac{p(x,y=1)}{p(x,y=0)}\right)\right)} \end{aligned} \tag{12}$$

Mit der Definition des Modells der Linearen Diskriminanzanalyse gilt dann

$$p(x,y=1) = p(x|y=1)p(y=1) = N(x;\mu_1,\Sigma)\mu \tag{13}$$

und

$$p(x,y=0) = p(x|y=0)p(y=0) = N(x;\mu_0,\Sigma)(1-\mu) \tag{14}$$

Beweis (fortgeführt)

Wir erhalten also

$$
\begin{aligned}
&= \ln\left(\frac{p(x, y=1)}{p(x, y=0)}\right) \\
&= \ln\left(\frac{N(x; \mu_1, \Sigma)\mu}{N(x; \mu_0, \Sigma)(1-\mu)}\right) \\
&= \ln\left(N(x; \mu_1, \Sigma)\mu\right) - \ln\left(N(x; \mu_0, \Sigma)(1-\mu)\right) \\
&= \ln(\mu) + \ln N(x; \mu_1, \Sigma) - \ln(1-\mu) - \ln N(x; \mu_0, \Sigma) \\
&= \ln\mu - \ln(1-\mu) - \frac{m}{2}\ln 2\pi - \ln|\Sigma| - \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \\
&\qquad\qquad + \frac{m}{2}\ln 2\pi + \ln|\Sigma| + \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) \\
&= -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + \ln\mu - \ln(1-\mu) \\
&= -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \ln\left(\frac{\mu}{1-\mu}\right) \\
&= \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - x^T\Sigma^{-1}(\mu_0-\mu_1) + \ln\left(\frac{\mu}{1-\mu}\right) \\
&= \begin{pmatrix}1 & x^T\end{pmatrix}\begin{pmatrix}\frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right) + \ln\left(\frac{\mu}{1-\mu}\right) \\ -\Sigma^{-1}(\mu_0-\mu_1)\end{pmatrix} \\
&=: \tilde{x}^T\beta
\end{aligned}
$$

□

Definition (Klassifikationsregel der linearen Diskriminanzanalyse)

$p(x, y)$ sei die WMDF des Modells der Linearen Diskriminanzanalyse. Dann ist die *Klassifikationregel* definiert als

$$\delta : \mathbb{R}^m \to \{0, 1\}, x \mapsto \delta(x) := \begin{cases} 0 & \text{für } p(y = 0|x) \geq p(y = 1|x) \\ 1 & \text{für } p(y = 0|x) < p(y = 1|x) \end{cases} \tag{15}$$

Bemerkung

- Es gilt

$$\delta(x) = 1 \Leftrightarrow p(y = 1|x) > p(y = 0|x) \Leftrightarrow p(y = 1|x) > 0.5. \tag{16}$$

Inferenz und Klassifikation bei bekannten Modellparametern

$$\mu = 0.5, \mu_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.50 & -0.10 \\ -0.10 & 0.50 \end{pmatrix}$$

```r
# Inferenz und Klassifikation für die ersten k Datenpunkte
k           = 10                                          # Anzahl Datenpunkte
x_tilde     = rbind(rep(1,k), x[,1:k])                    # erweiterte Featurevektoren
beta        = matrix(                                     # Inferenzparametervektor
              c((1/2)*(  t(mu_0) %*% solve(Sigma) %*% mu_0
                       - t(mu_1) %*% solve(Sigma) %*% mu_1)
                       + log(mu/(1-mu)),
                -solve(Sigma) %*% (mu_0-mu_1)), nrow = 3)
p_y_giv_x   = 1/(1+exp(-t(x_tilde) %*% beta))             # p(y = 1|x)
delta       = as.numeric(p_y_giv_x >= 0.5)                # Klassifikationsregel
```

	1	2	3	4	5	6	7	8	9	10
x_1	1.14	1.80	2.38	0.87	2.48	2.15	0.08	0.82	2.61	1.47
x_2	3.24	2.05	1.50	0.77	2.37	2.65	0.94	0.64	2.32	0.59
p(y = 1|x)	1.00	0.97	0.97	0.00	1.00	1.00	0.00	0.00	1.00	0.02
delta(x)	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00
y	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	1.00	0.00

Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Theorem (ML-Schätzer der Linearen Diskriminanzanalyse)

$p(x, y)$ sei die WMDF des Modells einer Linearen Diskriminanzanalyse mit Parametern $\{\mu, \mu_0, \mu_1, \Sigma\}$, $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ sei ein LDA Trainingsdatensatz, und $1_{\{S\}}$ sei die Indikatorfunktion der Aussage A, d.h. $1_{\{A\}} = 1$, wenn A WAHR ist uns $1_{\{A\}} = 0$, wenn A FALSCH ist. Dann sind die Maximum-Likelihood-Schätzer für μ, μ_0, μ_1 und Σ gegeben durch

$$
\begin{aligned}
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{y^{(i)}=1\}}, \\
\hat{\mu}_0 &= \frac{1}{\sum_{i=1}^{n} 1_{\{y^{(i)}=0\}}} \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} x^{(i)}, \\
\hat{\mu}_1 &= \frac{1}{\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}}} \sum_{i=1}^{n} 1_{\{y^{(i)}=1\}} x^{(i)}, \\
\hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)} - \hat{\mu}_{y^{(i)}} \right) \left(x^{(i)} - \hat{\mu}_{y^{(i)}} \right)^T .
\end{aligned}
\tag{17}
$$

Bemerkungen

- μ wird als die relative Häufigkeit der 1en im Trainingsdatensatz geschätzt.
- μ_0 und μ_1 werden als Stichprobenmittel aller $x^{(i)}$ mit $y^{(i)} = 0$ bzw. $y^{(i)} = 1$ geschätzt.
- Σ wird durch die empirische Kovarianzmatrix aller $x^{(i)}, i = 1, \ldots, n$ geschätzt.
- Substitution ergibt den Schätzer $\hat{\beta}$

Beweis

(1) Formulierung der Log Likelihood Funktion

$$
\begin{aligned}
\ell(\mu, \mu_0, \mu_1, \Sigma) &:= \ln \prod_{i=1}^{n} p\left(x^{(i)}, y^{(i)}\right) \\
&= \sum_{i=1}^{n} \ln p\left(x^{(i)}, y^{(i)}\right) \\
&= \sum_{i=1}^{n} \ln p\left(x^{(i)} | y^{(i)}\right) p\left(y^{(i)}\right) \\
&= \sum_{i=1}^{n} \ln p\left(x^{(i)} | y^{(i)}\right) + \ln p\left(y^{(i)}\right) \\
&= \sum_{i=1}^{n} \ln \left(N(x^{(i)}; \mu_0, \Sigma)\right)^{1-y^{(i)}} \left(N(x^{(i)}; \mu_1, \Sigma)\right)^{y^{(i)}} + \ln \left(\mu^{y^{(i)}} (1-\mu)^{1-\mu^{y^{(i)}}}\right) \\
&= \sum_{i=1}^{n} \left(1-y^{(i)}\right) \ln N\left(x^{(i)}; \mu_0, \Sigma\right) + y^{(i)} \ln N\left(x^{(i)}; \mu_1, \Sigma\right) + y^{(i)} \ln \mu + \left(1-y^{(i)}\right) \ln(1-\mu) \\
&= \sum_{i=1}^{n} \left(1-y^{(i)}\right) \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_0\right)\right) \\
&\quad + \sum_{i=1}^{n} y^{(i)} \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x^{(i)} - \mu_1\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_1\right)\right) \\
&\quad + \sum_{i=1}^{n} y^{(i)} \ln \mu + \sum_{i=1}^{n} \left(1-y^{(i)}\right) \ln(1-\mu).
\end{aligned}
$$

Beweis (fortgeführt)

(2) Gradient der Log Likelihood Funktion

Der Gradient der Log Likelihood Funktion des Modells der Linearen Diskriminanzanalyse besteht aus den partiellen Ableitungen von ℓ hinsichtlich von μ, μ_0, μ_1 und Σ. Wie unten gezeigt ergibt er sich als

$$\nabla \ell\left(\mu, \mu_0, \mu_1, \Sigma\right) = \begin{pmatrix} \frac{\partial}{\partial \mu} \ell\left(\mu, \mu_0, \mu_1, \Sigma\right) \\ \frac{\partial}{\partial \mu_0} \ell\left(\mu, \mu_0, \mu_1, \Sigma\right) \\ \frac{\partial}{\partial \mu_1} \ell\left(\mu, \mu_0, \mu_1, \Sigma\right) \\ \frac{\partial}{\partial \Sigma} \ell\left(\mu, \mu_0, \mu_1, \Sigma\right) \end{pmatrix} = \begin{pmatrix} \frac{1}{\mu} \sum_{i=1}^{n} 1_{\{y^{(i)}=1\}} - \frac{1}{1-\mu} \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} \\ -\frac{1}{2} \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} \left(\left(x^{(i)} - \mu_0\right)^T \Sigma^{-1}\right) \\ -\frac{1}{2} \sum_{i=1}^{n} 1_{\{y^{(i)}=1\}} \left(\left(x^{(i)} - \mu_1\right)^T \Sigma^{-1}\right) \\ \frac{n}{2}\Sigma - \frac{1}{2} \sum_{i=1}^{n} \left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right) \left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)^T \end{pmatrix}.$$

Beweis (fortgeführt)

Für die partielle Ableitung hinsichtlich μ_0 und ähnlich für μ_1 ergibt sich

$$\begin{aligned}
\frac{\partial}{\partial \mu_0} \ell(\mu, \mu_0, \mu_1, \Sigma) &= \frac{\partial}{\partial \mu_0} \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_0\right)\right) \\
&= \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \frac{\partial}{\partial \mu_0} \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_0\right)\right) \\
&= \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(-\frac{1}{2} \frac{\partial}{\partial \mu_0} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_0\right)\right) \\
&= \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(-\frac{1}{2} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1}\right) \\
&= -\frac{1}{2} \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(\left(x^{(i)} - \mu_0\right)^T \Sigma^{-1}\right). \\
&= -\frac{1}{2} \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} \left(\left(x^{(i)} - \mu_0\right)^T \Sigma^{-1}\right).
\end{aligned}$$

Beweis (fortgeführt)

Für die partielle Ableitung hinsichtlich Σ ergibt sich

$$
\begin{aligned}
\frac{\partial}{\partial \Sigma} \ell(\mu, \mu_1, \mu_0, \Sigma) = \quad & \frac{\partial}{\partial \Sigma} \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_0\right)\right) \\
& + \frac{\partial}{\partial \Sigma} \sum_{i=1}^{n} y^{(i)} \left(-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left(x^{(i)} - \mu_1\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_1\right)\right) \\
= \quad & \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(-\frac{1}{2} \frac{\partial}{\partial \Sigma} \ln |\Sigma| - \frac{1}{2} \frac{\partial}{\partial \Sigma} \left(x^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_0\right)\right) \\
& + \sum_{i=1}^{n} y^{(i)} \left(-\frac{1}{2} \frac{\partial}{\partial \Sigma} \ln |\Sigma| - \frac{1}{2} \frac{\partial}{\partial \Sigma} \left(x^{(i)} - \mu_1\right)^T \Sigma^{-1} \left(x^{(i)} - \mu_1\right)\right) \qquad (18) \\
= \quad & \sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(-\frac{1}{2} \Sigma - \frac{1}{2} \left(x^{(i)} - \mu_0\right) \left(x^{(i)} - \mu_0\right)^T\right) \\
& + \sum_{i=1}^{n} y^{(i)} \left(-\frac{1}{2} \Sigma - \frac{1}{2} \left(x^{(i)} - \mu_1\right) \left(x^{(i)} - \mu_1\right)^T\right) \\
= \quad & \frac{n}{2} \Sigma - \frac{1}{2} \sum_{i=1}^{n} \left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right) \left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)^T .
\end{aligned}
$$

Beweis (fortgeführt)

Für die partielle Ableitung hinsichtlich μ ergibt sich

$$\begin{aligned}\frac{\partial}{\partial\mu}\ell(\mu,\mu_1,\mu_0,\Sigma) &= \frac{\partial}{\partial\mu}\left(\sum_{i=1}^{n} y^{(i)}\ln\mu + \sum_{i=1}^{n}\left(1-y^{(i)}\right)\ln(1-\mu)\right)\\ &= \sum_{i=1}^{n} y^{(i)}\frac{\partial}{\partial\mu}\ln\mu + \sum_{i=1}^{n}\left(1-y^{(i)}\right)\frac{\partial}{\partial\mu}\ln(1-\mu)\\ &= \frac{1}{\mu}\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}} - \frac{1}{1-\mu}\sum_{i=1}^{n} 1_{\{y^{(i)}=0\}}.\end{aligned}$$

(4) Auflösen der Maximum Likelihood Gleichungen

Nullsetzen der partiellen Ableitungen des Gradienten der Log Likelihood Funktion und Auflösen der resultierenden Log Likelihood Gleichungen ergibt dann die Maximum-Likelihood-Schätzer des Modells der Linearen Diskriminanzanalyse.

Beweis (fortgeführt)

Nullsetzen der ersten Gradientenkomponente ergibt

$$\begin{aligned}
&\frac{1}{\hat{\mu}} \sum_{i=1}^{n} 1_{\{y^{(i)}=1\}} - \frac{1}{1-\hat{\mu}} \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} = 0 \\
\Leftrightarrow\ &\frac{1}{\hat{\mu}} \sum_{i=1}^{n} y^{(i)} - \frac{1}{1-\hat{\mu}} \sum_{i=1}^{n} \left(1 - y^{(i)}\right) = 0 \\
\Leftrightarrow\ &\frac{1-\hat{\mu}}{\hat{\mu}} \sum_{i=1}^{n} y^{(i)} - \sum_{i=1}^{n} \left(1 - y^{(i)}\right) = 0 \\
\Leftrightarrow\ &\frac{1-\hat{\mu}}{\hat{\mu}} \sum_{i=1}^{n} y^{(i)} - n + \sum_{i=1}^{n} y^{(i)} = 0 \\
\Leftrightarrow\ &(1-\hat{\mu}) \sum_{i=1}^{n} y^{(i)} - \hat{\mu} n + \hat{\mu} \sum_{i=1}^{n} y^{(i)} = 0 \\
\Leftrightarrow\ &(1-\hat{\mu}) \sum_{i=1}^{n} y^{(i)} - \hat{\mu} n + \hat{\mu} \sum_{i=1}^{n} y^{(i)} = 0 \\
\Leftrightarrow\ &(1-\hat{\mu}+\hat{\mu}) \sum_{i=1}^{n} y^{(i)} = \hat{\mu} n \\
\Leftrightarrow\ &\hat{\mu} n = \sum_{i=1}^{n} y^{(i)} \\
\Leftrightarrow\ &\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{y^{(i)}=1\}}.
\end{aligned}$$

Beweis (fortgeführt)

Nullsetzen der zweiten Gradientenkomponente ergibt

$$\begin{aligned}
&\sum_{i=1}^{n} \left(1 - y^{(i)}\right) \left(\left(x^{(i)} - \hat{\mu}_0\right)^T \Sigma^{-1}\right) = 0 \\
&\quad \Leftrightarrow \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} \left(x^{(i)} - \hat{\mu}_0\right)^T = 0 \\
\Leftrightarrow &\sum_{i=1}^{n} 1_{\{y^{(i)}\}} x^{(i)} - \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} \hat{\mu}_0 = 0 \\
&\quad \Leftrightarrow \sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} \hat{\mu}_0 = \sum_{i=1}^{n} 1_{\{y^{(i)}\}} x^{(i)} \\
&\quad \Leftrightarrow \hat{\mu}_0 = \frac{1}{\sum_{i=1}^{n} 1_{\{y^{(i)}=0\}}} \sum_{i=1}^{n} 1_{\{y^{(i)}\}} x^{(i)}.
\end{aligned}$$

Nullsetzen der dritten Gradientenkomponente ergibt dann in ähnlicher Weise den Maximum-Likelihood-Schätzer $\hat{\mu}_1$.

Beweis (fortgeführt)

Nullsetzen der vierten Gradientenkomponente ergibt dann schließlich

$$
\begin{aligned}
0 &= \frac{n}{2}\hat{\Sigma} - \frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)\left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)^T \\
\Leftrightarrow n\hat{\Sigma} &= \sum_{i=1}^{n}\left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)\left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)^T \\
\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)\left(x^{(i)} - \mu_{1_{\{y^{(i)}=1\}}}\right)^T
\end{aligned}
\quad .
$$

Anwendung

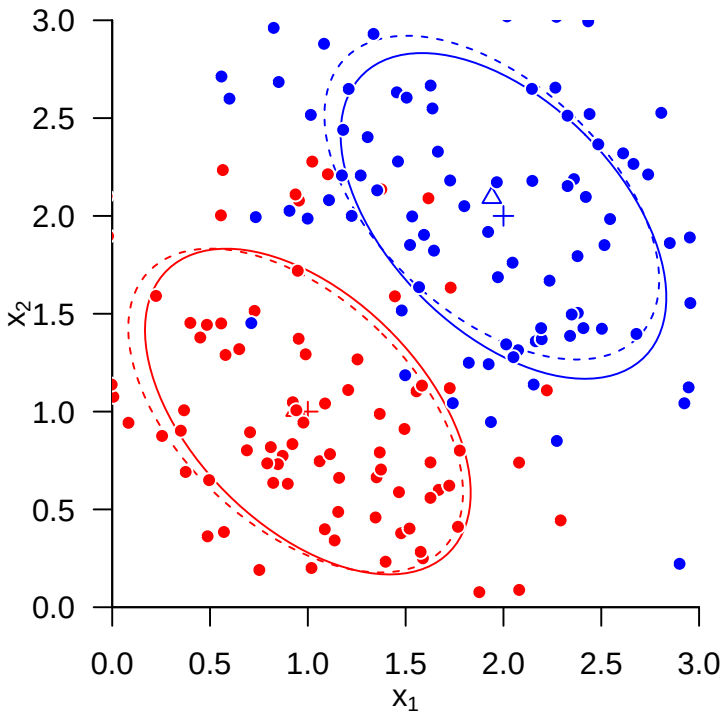
$$\hat{\mu} = 0.52, \hat{\mu}_0 = \begin{pmatrix} 0.94 \\ 1.01 \end{pmatrix}, \hat{\mu}_1 = \begin{pmatrix} 1.94 \\ 2.09 \end{pmatrix}, \hat{\Sigma} = \begin{pmatrix} 0.53 & -0.26 \\ -0.26 & 0.49 \end{pmatrix}$$

```
# Parameterlernen
n            = ncol(x)                                                    # n
m            = nrow(x)                                                    # m
mu_hat       = mean(y)                                                    # \hat{\mu}
mu_0_hat     = rowMeans(x[,y == 0])                                       # \hat{\mu}_0
mu_1_hat     = rowMeans(x[,y == 1])                                       # \hat{\mu}_1
Sigma_hat    = matrix(rep(0,m^2), nrow = m)                               # \hat{\Sigma}
for(i in 1:n){
  Sigma_hat  = (Sigma_hat + (1/n)*
               ((y[i] == 0)*(x[,i]-mu_0_hat) %*% t((x[,i]-mu_0_hat))
               +(y[i] == 1)*(x[,i]-mu_1_hat) %*% t((x[,i]-mu_1_hat))))
}
beta_hat     = matrix(c((1/2)*( t(mu_0_hat) %*% solve(Sigma_hat) %*% mu_0_hat   # \hat{beta}
                              - t(mu_1_hat) %*% solve(Sigma_hat) %*% mu_1_hat)
                              + log(mu_hat/(1-mu_hat)),
                        -solve(Sigma_hat) %*% (mu_0_hat-mu_1_hat)),
                      nrow = m+1)
```

Anwendung

$$m = 2, n = 200, \hat{\mu} = 0.52, \hat{\mu}_0 = \begin{pmatrix} 0.94 \\ 1.01 \end{pmatrix}, \hat{\mu}_1 = \begin{pmatrix} 1.94 \\ 2.09 \end{pmatrix}, \hat{\Sigma} = \begin{pmatrix} 0.53 & -0.26 \\ -0.26 & 0.49 \end{pmatrix}$$

$+\ \mu_0$, $\bullet\ x^{(i)}$ mit $y^{(i)} = 0$, $\triangle \hat{\mu}_0$, $+\ \mu_1$, $\bullet\ x^{(i)}$ mit $y^{(i)} = 1$, $\triangle\ \hat{\mu}_1$



Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

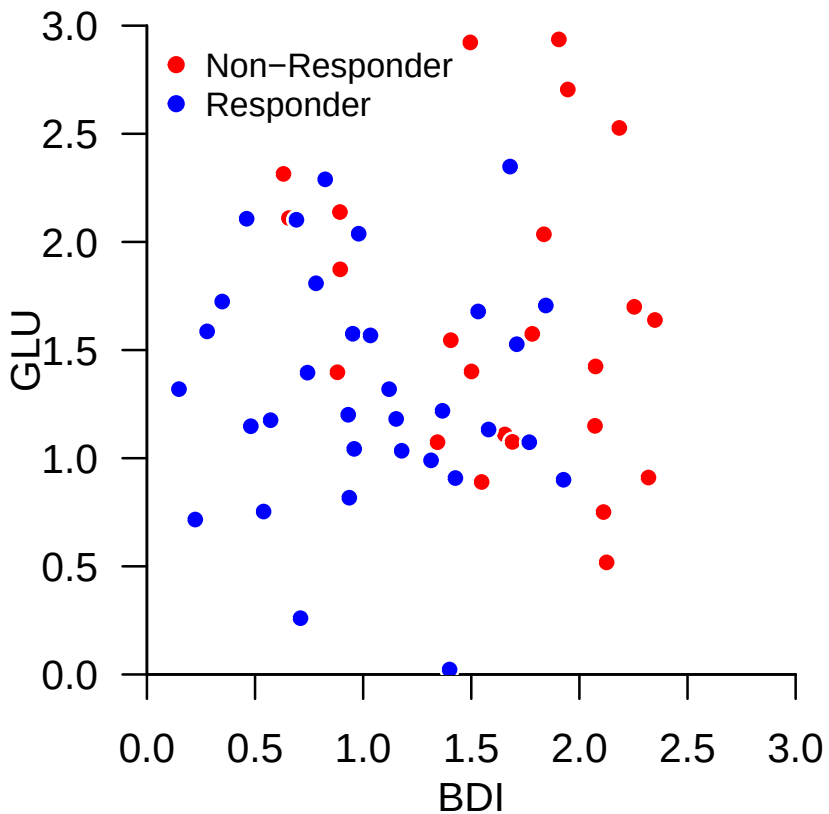
Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Beispieldatensatz

BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg



Beispieldatensatz

`BDI` und `GLU` Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg `RES`

BDI	GLU	RES
0.74	1.40	1
0.22	0.72	1
0.82	2.29	1
2.07	1.15	0
1.71	1.53	1
1.77	1.07	1
1.95	2.70	0
2.18	2.53	0
0.93	1.20	1
1.34	1.07	0
2.35	1.64	0
1.43	0.91	1
1.66	1.11	0
0.28	1.59	1
2.13	0.52	0
1.37	1.22	1
0.89	2.14	0
0.88	1.40	0
0.98	2.04	1
1.93	0.90	1

LOOCV zur Bestimmung der Featureprädiktivität

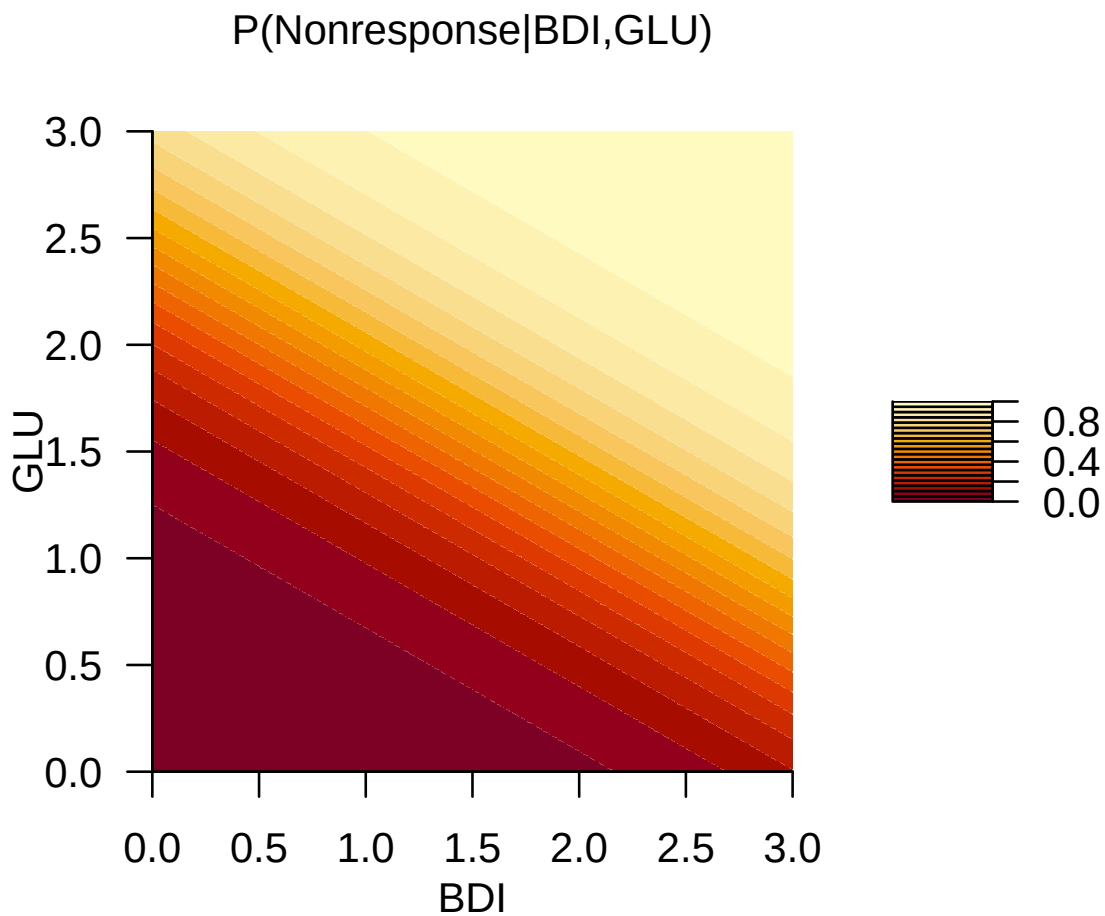
```
D       = read.csv("./10_Daten/10_Lineare_Diskriminanzanalyse.csv")             # Datensatz
K       = nrow(D)                                                               # Anzahl Cross Folds
p_y     = matrix(rep(NaN, K)  , nrow = 1)                                       # p(y = 1|x)
y_pred  = matrix(rep(NaN, K*2), nrow = K)                                       # Prädiktionsperformancearray
for(k in 1:K){                                                                  # K-fold LOOCV
  x_train     = t(D[-k,1:2])                                                    # Trainingsdatensatzfeatures
  y_train     = t(D[-k,3  ])                                                    # Trainingsdatensatzlabels
  x_test      = t(D[ k,1:2])                                                    # Testdatensatzfeaturevektor
  y_pred[k,1] = t(D[ k,3])                                                      # Testdatensatzfeaturevektorlabel
  n           = ncol(x_train)                                                   # n
  m           = nrow(x_train)                                                   # m
  mu_hat      = mean(y_train)                                                   # \hat{\mu}
  mu_0_hat    = rowMeans(x_train[,y_train == 0])                                # \hat{\mu}_0
  mu_1_hat    = rowMeans(x_train[,y_train == 1])                                # \hat{\mu}_1
  Sigma_hat   = matrix(rep(0,m^2), nrow = m)                                    # \hat{\Sigma}
  for(i in 1:n){
    Sigma_hat  = (Sigma_hat + (1/n)*
                 ((y_train[i] == 0)*(x_train[,i]-mu_0_hat) %*% t((x_train[,i]-mu_0_hat))
                 +(y_train[i] == 1)*(x_train[,i]-mu_1_hat) %*% t((x_train[,i]-mu_1_hat))))}
  beta_hat    = matrix(c((1/2)*( t(mu_0_hat) %*% solve(Sigma_hat) %*% mu_0_hat # \hat{beta}
                               - t(mu_1_hat) %*% solve(Sigma_hat) %*% mu_1_hat)
                               + log(mu_hat/(1-mu_hat)),
                               -solve(Sigma_hat) %*% (mu_0_hat-mu_1_hat)), nrow = m+1)
  x_test_tilde = rbind(1, x_test)                                               # \tilde{x}
  p_y[k]       = 1/(1+exp(-t(x_test_tilde) %*% beta_hat))                       # p(y=1|x)
  y_pred[k,2]  = as.numeric(p_y[k] >= 0.5)}                                     # Prädiktion
rp      = sum(y_pred[y_pred[,1] == 1,2] == 1)                                   # |(1,1)|
rn      = sum(y_pred[y_pred[,1] == 0,2] == 0)                                   # |(0,0)|
fp      = sum(y_pred[y_pred[,1] == 0,2] == 1)                                   # |(0,1)|
fn      = sum(y_pred[y_pred[,1] == 1,2] == 0)                                   # |(1,0)|
ACC     = (rp+rn)/(rp+fp+rn+fn)                                                 # Accuracy
SEN     = rp/(rp+fn)                                                            # Sensitivity
SPE     = rn/(rn+fp)                                                            # Specificity
cat("Accuracy   : "  , ACC, ", Sensitivity: " , SEN, ", Specificity: " , SPE)   # Ergebnisausgabe
```

```
Accuracy   :  0.7166667 , Sensitivity:  0.8235294 , Specificity:  0.5769231
```

Evaluation der Non-Response Wahrscheinlichkeit

```
D           = read.csv("./10_Daten/10_Lineare_Diskriminanzanalyse.csv")          # Datensatz
x           = t(D[,1:2])                                                        # Featurevektoren
y           = t(D[,3])                                                          # Label
n           = ncol(x)                                                           # n
m           = nrow(x)                                                           # m
mu_hat      = mean(y)                                                           # \hat{\mu}
mu_0_hat    = rowMeans(x[,y == 0])                                              # \hat{\mu}_0
mu_1_hat    = rowMeans(x[,y == 1])                                              # \hat{\mu}_1
Sigma_hat   = matrix(rep(0,m^2), nrow = m)                                      # \hat{\Sigma}
for(i in 1:n){
  Sigma_hat  = (Sigma_hat + (1/n)*
               ((y[i] == 0)*(x[,i]-mu_0_hat) %*% t((x[,i]-mu_0_hat))
               +(y[i] == 1)*(x[,i]-mu_1_hat) %*% t((x[,i]-mu_1_hat))))}
beta_hat     = matrix(c((1/2)*( t(mu_0_hat) %*% solve(Sigma_hat) %*% mu_0_hat   # \hat{beta}
                              - t(mu_1_hat) %*% solve(Sigma_hat) %*% mu_1_hat)
                              + log(mu_hat/(1-mu_hat)),
                        -solve(Sigma_hat) %*% (mu_0_hat-mu_1_hat)),
                      nrow = m+1)
x_min     = 0                                                                   # GLU/BDI Minimum
x_max     = 3                                                                   # GLU/BDI Maximum
x_res     = 5e2                                                                 # GLU/BDI Auflösung
bdi       = seq(x_min, x_max, length.out = x_res)                               # BDI
glu       = seq(x_min, x_max, length.out = x_res)                               # GLU
p_y       = matrix(rep(NaN, x_res*x_res), nrow = x_res)                         # p(y=1|(BDI, GLU))
for(i in 1:x_res){                                                              # BDI Iterationen
    for(j in 1:x_res){                                                          # GLU Iterationen
        x_tilde   = rbind(1, bdi[i], glu[j])                                    # \tilde{x}
        p_y[i,j]  = 1/(1+exp(-t(x_tilde) %*% beta_hat))}}                       # p(y=1|(BDI, GLU))
```

Evaluation der Non-Response Wahrscheinlichkeit



Anwendungsszenario

Modellformulierung

Inferenz und Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition einer Bernoullizufallsvariable wieder.
2. Geben Sie die Definition des Modells der Linearen Diskriminanzanalyse wieder.
3. Erläutern Sie die Generation von Daten unter dem Modell der Linearen Diskriminanzanalyse.
4. Geben Sie das Theorem zur Inferenz der Linearen Diskriminanzanalyse wieder.
5. Geben Sie die Definition der Klassifikationsregel der Linearen Diskriminanzanalyse wieder.
6. Geben Sie das Theorem zur Maximum-Likelihood-Schätzung der Linearen Diskriminanzanalyse wieder.
7. Erläutern Sie wie, mithilfe einer Linearen Diskriminanzanalyse die psychotherapeutische Nonresponse-wahrscheinlichkeit geschätzt werden kann.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(11) Nichtlineare Optimierung

Motivation

- Viele Verfahren der Prädiktiven Modellierung oder, wie heutzutage auch oft gesagt wird, der "Künstlichen Intelligenz", z.B. Neuronale Netze als generalisierte logistische Regression oder Support-Vektor-Maschinen, basieren auf mathematisch recht überschaubaren Modellen.
- Ihre momentane breite Verwendung verdanken diese Verfahren im Wesentlichen den verbesserten Computer-Hardware-Komponenten der letzten 15 Jahre, die zum Lernen ihrer Parameter ("Trainieren") genutzt werden, weniger wesentlichen neuen theoretischen Einsichten. Für einen aktuellen Überblick, siehe zum Beispiel Prince (2023) und Murphy (2023).
- Das Lernen von Parametern von Modellen der Prädiktiven Modellierung enstpricht der Optimierung von Funktionen, wie wir in (12) Logistische Regression und (13) Neuronale Netze sehen werden, insbesondere der Minimierung sogenannter *Loss Functions*.
- Zur Optimierung von Funktionen werden in diesem Bereich insbesondere und im einfachsten Fall sogenannten *Gradientenverfahren* eingesetzt. In diesem Abschnitt wollen wir deshalb zunächst ein Grundverständnis von Gradientenverfahren erarbeiten.

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Definition (Multivariate reellwertige Funktion)

Eine Funktion der Form

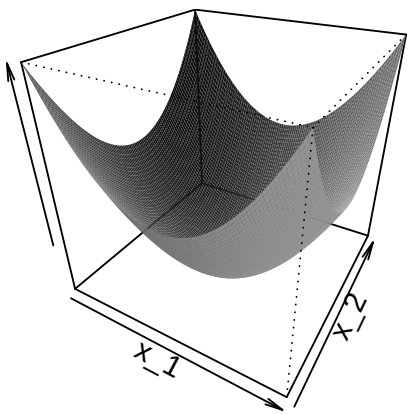
$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) = f(x_1, ..., x_n) \tag{1}$$

heißt *multivariate reellwertige* Funktion.

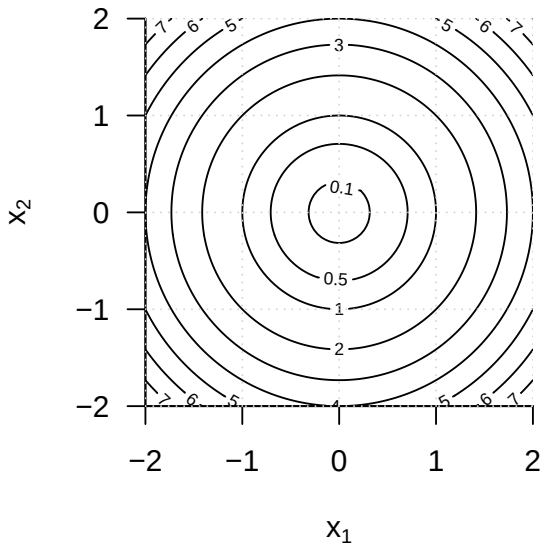
Beispiel für $n := 2$

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2 \tag{2}$$

f(x): = $x_1^2 + x_2^2$



f(x): = $x_1^2 + x_2^2$



Definition (Partielle Ableitung)

Es sei $D \subseteq \mathbb{R}^n$ eine Menge und

$$f : D \to \mathbb{R}, x \mapsto f(x) \tag{3}$$

eine multivariate reellwertige Funktion. f heißt in $x \in D$ *nach* x_i *partiell differenzierbar*, wenn der Grenzwert

$$\frac{\partial}{\partial x_i} f(x) := \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \tag{4}$$

existiert. $\frac{\partial}{\partial x_i} f(x)$ heißt dann die *partielle Ableitung von* f *nach* x_i *an der Stelle* x. Wenn f für alle $x \in D$, nach x_i partiell differenzierbar ist, dann heißt f *nach* x_i *partiell differenzierbar* und die Funktion

$$\frac{\partial}{\partial x_i} f : D \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_i} f(x) \tag{5}$$

heißt *partielle Ableitung von* f *nach* x_i.

f heißt *partiell differenzierbar in* $x \in D$, wenn f für alle $i = 1, \ldots, n$ in $x \in D$ nach x_i partiell differenzierbar ist, und f heißt partiell differenzierbar, wenn f für alle $i = 1, \ldots, n$ in allen $x \in D$ nach x_i partiell differenzierbar ist.

Bemerkungen

- $e_i \in \mathbb{R}^n$ bezeichnet den $i$ten Einheitsvektor.
- $\frac{f(x+he_i)-f(x)}{h}$ misst die Änderung $f(x + he_i) - f(x)$ von f pro Strecke h in Richtung e_i.
- Für $h \to 0$ misst der Differenzquotient die Änderungsrate von f in x in Richtung e_i.
- $\frac{\partial}{\partial x_i} f(x)$ ist eine Zahl, $\frac{\partial}{\partial x_i} f$ ist eine Funktion.
- Praktisch berechnet man $\frac{\partial}{\partial x_i} f$ als die (einfache) Ableitung

$$\frac{d}{dx_i} \tilde{f}_{x_1,...x_{i-1},x_{i+1},...,x_n}(x_i) \tag{6}$$

der univariaten reellwertigen Funktion

$$\tilde{f} : \mathbb{R} \to \mathbb{R}, x_i \mapsto \tilde{f}_{x_1,...x_{i-1},x_{i+1},...,x_n}(x_i) := f(x_1, ..., x_i, ..., x_n). \tag{7}$$

- Man betrachtet alle x_j mit $j \neq i$ also als Konstanten.

Beispiel (1)

Wir betrachten die Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{8}$$

Weil die Definitionsmenge dieser Funktion zweidimensional ist, kann man zwei partielle Ableitungen berechnen

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) \text{ und } \frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x). \tag{9}$$

Um die erste dieser partiellen Ableitungen zu berechnen, betrachtet man die Funktion

$$f_{x_2} : \mathbb{R} \to \mathbb{R}, x_1 \mapsto f_{x_2}(x_1) := x_1^2 + x_2^2, \tag{10}$$

wobei x_2 hier die Rolle einer Konstanten einnimmt. Um explizit zu machen, dass x_2 kein Argument der Funktion ist, die Funktion aber weiterhin von x_2 abhängt haben wir die Subskriptnotation $f_{x_2}(x_1)$ verwendet. Um nun die partielle Ableitung zu berechnen, berechnen wir die (einfache) Ableitung von f_{x_2},

$$f'_{x_2}(x) = 2x_1\,. \tag{11}$$

Es ergibt sich also

$$\frac{\partial}{\partial x_1} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_1} f(x) = \frac{\partial}{\partial x_1}(x_1^2 + x_2^2) = f'_{x_2}(x) = 2x_1\,. \tag{12}$$

Analog gilt mit der entsprechenden Formulierung von f_{x_1}, dass

$$\frac{\partial}{\partial x_2} f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto \frac{\partial}{\partial x_2} f(x) = \frac{\partial}{\partial x_2}(x_1^2 + x_2^2) = f'_{x_1}(x) = 2x_2\,. \tag{13}$$

Definition (Zweite partielle Ableitungen)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion und $\frac{\partial}{\partial x_i} f$ sei die partielle Ableitung von f nach x_i. Dann ist die zweite partielle Ableitung von f nach x_i und x_j definiert als

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) := \frac{\partial}{\partial x_j} \left(\frac{\partial}{\partial x_i} f \right) \tag{14}$$

Bemerkungen

- Wie die zweite Ableitung ist auch die zweite partielle Ableitung rekursiv definiert.
- Zu jeder partiellen Ableitung $\frac{\partial}{\partial x_i} f$ gibt es n zweite partiellen Ableitungen $\frac{\partial^2}{\partial x_j \partial x_i} f, j = 1, \ldots, n$.

Theorem (Satz von Schwarz)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine partiell differenzierbare multivariate reellwertige Funktion. Dann gilt

$$\frac{\partial^2}{\partial x_j \partial x_i} f(x) = \frac{\partial^2}{\partial x_i \partial x_j} f(x) \text{ für alle } 1 \leq i, j \leq n. \tag{15}$$

Bemerkungen

- Wir verzichten auf einen Beweis.
- Das Theorem von Schwarz besagt, dass die Reihenfolge des partiellen Ableitens irrelevant ist
- Das Theorem erleichtert die Berechnung von zweiten partiellen Ableitungen.
- Das Theorem hilft, Fehler bei der Berechnung zweiter partieller Ableitungen aufzudecken.

Beispiel (1) (fortgeführt)

Wir wollen die partiellen Ableitungen zweiter Ordnung der Funktion

$$f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2. \tag{16}$$

berechnen. Mit den Ergebnissen für die partiellen Ableitungen erster Ordnung dieser Funktion ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_1} (2x_1) = 2 \\
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_1} (2x_2) = 0 \\
\frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_1} f(x) \right) = \frac{\partial}{\partial x_2} (2x_1) = 0 \\
\frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2} \left(\frac{\partial}{\partial x_2} f(x) \right) = \frac{\partial}{\partial x_2} (2x_2) = 2
\end{aligned} \tag{17}$$

Offenbar gilt

$$\frac{\partial^2}{\partial x_1 \partial x_2} f(x) = \frac{\partial^2}{\partial x_2 \partial x_1} f(x). \tag{18}$$

Beispiel (2)

Wir wollen die partiellen Ableitungen erster und zweiter Ordnung der Funktion

$$f : \mathbb{R}^3 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_1 x_2 + x_2\sqrt{x_3}. \tag{19}$$

berechnen.

Mit den Rechenregeln für Ableitungen ergibt sich für die partiellen Ableitungen erster Ordnung

$$\begin{aligned}
\frac{\partial}{\partial x_1} f(x) &= \frac{\partial}{\partial x_1}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = 2x_1 + x_2, \\
\frac{\partial}{\partial x_2} f(x) &= \frac{\partial}{\partial x_2}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = x_1 + \sqrt{x_3}, \\
\frac{\partial}{\partial x_3} f(x) &= \frac{\partial}{\partial x_3}\left(x_1^2 + x_1 x_2 + x_2\sqrt{x_3}\right) = \frac{x_2}{2\sqrt{x_3}}.
\end{aligned} \tag{20}$$

Beispiel (2) (fortgeführt)

Für die zweiten partiellen Ableitungen hinsichtlich x_1 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_1} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_1}(2x_1 + x_2) = 2, \\
\frac{\partial^2}{\partial x_2 \partial x_1} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_2}(2x_1 + x_2) = 1, \\
\frac{\partial^2}{\partial x_3 \partial x_1} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_1} f(x)\right) = \frac{\partial}{\partial x_3}(2x_1 + x_2) = 0.
\end{aligned} \tag{21}$$

Für die zweiten partiellen Ableitungen hinsichtlich x_2 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_1}(x_1 + \sqrt{x_3}) = 1, \\
\frac{\partial^2}{\partial x_2 \partial x_2} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_2}(x_1 + \sqrt{x_3}) = 0, \\
\frac{\partial^2}{\partial x_3 \partial x_2} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_2} f(x)\right) = \frac{\partial}{\partial x_3}(x_1 + \sqrt{x_3}) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{22}$$

Beispiel (2) (fortgeführt)

Für die zweiten partiellen Ableitungen hinsichtlich x_3 ergibt sich

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial}{\partial x_1}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_1}\left(\frac{x_2}{2}\sqrt{x_3}\right) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial}{\partial x_2}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_2}\left(\frac{x_2}{2\sqrt{x_3}}\right) = \frac{1}{2\sqrt{x_3}}, \\
\frac{\partial^2}{\partial x_3 \partial x_3} f(x) &= \frac{\partial}{\partial x_3}\left(\frac{\partial}{\partial x_3} f(x)\right) = \frac{\partial}{\partial x_3}\left(x_2 \frac{1}{2} x_3^{-\frac{1}{2}}\right) = -\frac{1}{4} x_2 x_3^{-\frac{3}{2}}.
\end{aligned} \tag{23}$$

Weiterhin erkennt man, dass die Reihenfolge der partiellen Ableitungen irrelevant ist, denn es gilt

$$\begin{aligned}
\frac{\partial^2}{\partial x_1 \partial x_2} f(x) &= \frac{\partial^2}{\partial x_2 \partial x_1} f(x) = 1, \\
\frac{\partial^2}{\partial x_1 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_1} f(x) = 0, \\
\frac{\partial^2}{\partial x_2 \partial x_3} f(x) &= \frac{\partial^2}{\partial x_3 \partial x_2} f(x) = \frac{1}{2\sqrt{x_3}}.
\end{aligned} \tag{24}$$

Definition (Gradient)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist der *Gradient* $\nabla f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^n. \tag{25}$$

Bemerkung

- $\nabla f(x)$ fasst die partiellen Ableitungen von f an der Stelle $x \in \mathbb{R}^n$ in einem Vektor zusammen.
- Gradienten sind multivariate vektorwertige Abbildungen der Form $\nabla f : \mathbb{R}^n \to \mathbb{R}^n, x \mapsto \nabla f(x)$.
- Wir zeigen später, dass $-\nabla f(x)$ die Richtung des steilsten Abstiegs von f in $\mathbb{R}^n$ anzeigt.
- Für $n = 1$ gilt $\nabla f(x) = f'(x)$.

Beispiele

Für die in Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \in \mathbb{R}^2. \tag{26}$$

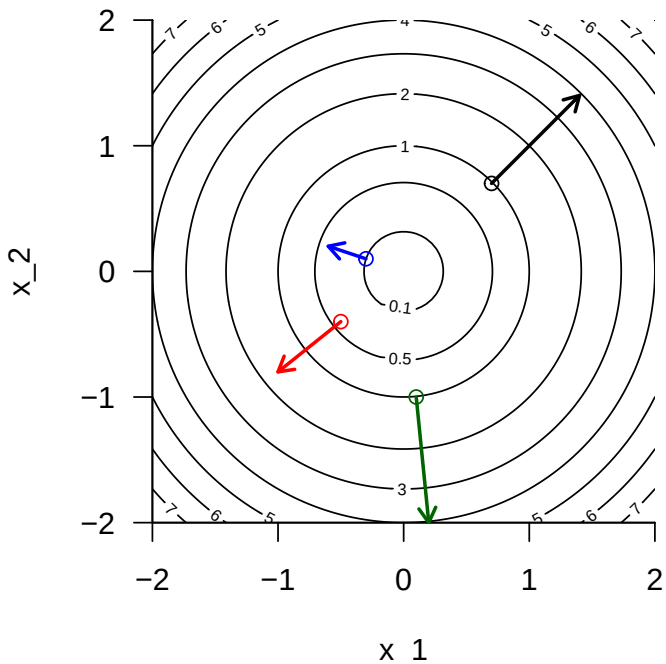
Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\nabla f(x) := \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \frac{\partial}{\partial x_3} f(x) \end{pmatrix} = \begin{pmatrix} 2x_1 + x_2 \\ x_1 + \sqrt{x_3} \\ \frac{x_2}{2\sqrt{x_3}} \end{pmatrix} \in \mathbb{R}^3. \tag{27}$$

Beispiel (1) (fortgeführt)

Gradienten von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$ bei

$$x = \begin{pmatrix} 0.7 \\ 0.7 \end{pmatrix} \qquad x = \begin{pmatrix} -0.3 \\ 0.1 \end{pmatrix} \qquad x = \begin{pmatrix} -0.5 \\ -0.4 \end{pmatrix} \qquad x = \begin{pmatrix} 0.1 \\ -1.0 \end{pmatrix}$$



Definition (Hesse-Matrix)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion. Dann ist die *Hesse-Matrix* $\nabla^2 f(x)$ von f an der Stelle $x \in \mathbb{R}^n$ definiert als

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_n} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(x) & \frac{\partial^2}{\partial x_n \partial x_2} f(x) & \cdots & \frac{\partial^2}{\partial x_n \partial x_n} f(x) \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{28}$$

Bemerkung

- $\nabla^2 f(x)$ fasst die partiellen Ableitungen zweiter Ordnung von f in einer Matrix zusammen.
- Hesse-Matrizen sind multivariate matrixwertige Abbildungen der Form $\nabla^2 f : \mathbb{R}^n \to \mathbb{R}^{n \times n}, x \mapsto \nabla^2 f(x)$.
- Für $n = 1$ gilt $\nabla^2 f(x) = f''(x)$.
- Mit $\frac{\partial^2}{\partial x_i \partial x_j} f(x) = \frac{\partial^2}{\partial x_j \partial x_i} f(x)$ für $1 \leq i, j \leq n$ folgt, dass $\left(\nabla^2 f(x)\right)^T = \nabla^2 f(x)$.

Beispiel

Für die in Beispiel (1) betrachtete Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

Für die in Beispiel (2) betrachtete Funktion $f : \mathbb{R}^3 \to \mathbb{R}$ gilt

$$\nabla^2 f(x) := \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \frac{\partial^2}{\partial x_1 \partial x_3} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2 \partial x_2} f(x) & \frac{\partial^2}{\partial x_2 \partial x_3} f(x) \\ \frac{\partial^2}{\partial x_3 \partial x_1} f(x) & \frac{\partial^2}{\partial x_3 \partial x_2} f(x) & \frac{\partial^2}{\partial x_3 \partial x_3} f(x) \end{pmatrix}$$

$$:= \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & \frac{1}{2\sqrt{3}} \\ 0 & \frac{1}{2\sqrt{3}} & -\frac{1}{4} x_2 x_3^{-3/2} \end{pmatrix}$$

Definition (Glatte multivariate reellwertige Funktion)

Eine multivariate reellwertige Funktion

$$f : \mathbb{R}^n \to \mathbb{R}, x \mapsto f(x) \tag{29}$$

heißt *glatt*, wenn ihr Gradient und ihre Hesse-Matrix existieren und für alle $x \in \mathbb{R}^n$ stetig sind.

Bemerkungen

- Der Gradient und die Hesse-Matrix einer glatten Funktion könnten überall in $\mathbb{R}^n$ berechnet werden.

Theorem (Multivariater Mittelwertsatz erster Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $p \in \mathbb{R}^n$. Dann gibt es ein $t \in]0, 1[$, so dass gilt

$$f(x + p) = f(x) + \nabla f(x + tp)^T p. \tag{30}$$

Theorem (Multivariater Mittelwertsatz zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $p \in \mathbb{R}^n$. Dann gibt es ein $t \in]0, 1[$, so dass gilt

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x + tp) p. \tag{31}$$

Bemerkung

- Wir verzichten auf Beweise.
- Nocedal and Wright (2006) bezeichnen die Theoreme als “Taylor’s Theorem”, das ist ein wenig misleading.
- ∇f und $\nabla^2 f$ werden an einer Stelle zwischen x und $x + p$ evaluiert.

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Definition (Optimierungsproblem)

Ein *Optimierungsproblem* hat die allgemeine Form

$$\min_x f(x), \tag{32}$$

wobei $x \in \mathbb{R}^n$ und $f : \mathbb{R}^n \to \mathbb{R}$ eine glatte multivariate reellwertige Funktion ist. Die Lösung x^* eines Optimierungsproblems wird bezeichnet mit

$$x^* = \arg\min_x f(x). \tag{33}$$

Bemerkungen

- Weil gilt, dass $\max_x f(x) = \min_x -f(x)$ genügt es, sich mit Minimierungsproblemen zu befassen.
- Im Allgemeinen ist die Lösung x^* eines Optimierungsproblems eine Menge.
- Man denkt bei $\arg\min_x f(x)$ alllerdings auch einfach an Elemente dieser Menge.

Definition (Globale und lokale Minimalstellen/Minima)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine multivariate reellwertige Funktion.

- $x^* \in \mathbb{R}^n$ heißt *globale Minimalstelle* von f, wenn $f(x^*) \leq f(x)$ für alle $x \in \mathbb{R}^n$ gilt. $f(x^*) \in \mathbb{R}$ heißt dann das *globale Minimum* von f.

- $x^* \in \mathbb{R}^n$ heißt *lokale Minimalstelle* von f, wenn es eine Umgebung N von x^* gibt, so dass $f(x^*) \leq f(x)$ für alle $x \in N \subset \mathbb{R}^n$. $f(x^*) \in \mathbb{R}$ wird dann ein *lokales Minimum* von f genannt.

- $x^* \in \mathbb{R}^n$ heißt *strikte lokale Minimalstelle* von f, wenn es eine Umgebung N von x^* gibt, so dass $f(x^*) < f(x)$ für alle $x \in N \subset \mathbb{R}^n$. $f(x^*) \in \mathbb{R}$ wird dann ein *striktes lokales Minimum* von f genannt.

Bemerkung

- Eine Umgebung von $x \in \mathbb{R}^n$ ist eine offene Menge, die x enthält.

Theorem (Notwendige Bedingung erster Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion. Wenn x^* eine lokale Minimalstelle von f ist, dann gilt

$$\nabla f(x^*) = 0_n. \tag{34}$$

Beweis

Wir beweisen das Theorem mithilfe eines indirekten Beweises (Beweis durch Widerspruch). Dazu nehmen wir an, dass x^* zwar eine lokale Minimalstelle von f ist, aber $\nabla f(x^*) \neq 0_n$ ist. Dazu definieren wir zunächst $p := -\nabla f(x^*)$. Dann gilt, dass

$$p^T \nabla f(x^*) = -\nabla f(x^*)^T \nabla f(x^*) = -||\nabla f(x^*)|| < 0. \tag{35}$$

Weil ∇f in einer Umgebung von x^* stetig ist, existiert ein Skalar $T > 0$, so dass auch

$$p^T \nabla f(x^* + tp) < 0 \text{ für alle } t \in [0, T]. \tag{36}$$

gilt. Nun gilt für $\tilde{t} \in]0, T]$ aber mit dem Mittelwertssatz erster Ordnung, dass

$$f(x^* + \tilde{t}p) = f(x^*) + \nabla f(x^* + tp)^T \tilde{t}p = f(x^*) + \tilde{t}p^T \nabla f(x^* + tp) \text{ für ein } t \in]0, \tilde{t}[. \tag{37}$$

Also folgt $f(x^* + \tilde{t}p) < f(x^*)$ für alle $\tilde{t} \in]0, T]$. Wir haben also eine Richtung von x^* weg gefunden, in der f abnimmt. Also kann x^* keine Minimalstelle sein, wenn $\nabla f(x^*) \neq 0_n$ gilt. Dies ist aber ein Widerspruch, zur Annahme, dass es möglich ist, dass x^* eine lokale Minimalstelle von f ist und $\nabla f(x^*) \neq 0_n$ gilt. Also muss $\nabla f(x^*) = 0_n$ gelten, wenn x^* eine lokale Minimalstelle ist.

Theorem (Notwendige Bedingung zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion. Wenn x^* eine lokale Minimalstelle von f ist, dann ist $\nabla f(x^*) = 0_n$ und $\nabla^2 f(x^*)$ ist positiv semidefinit.

Beweis

Wir beweisen das Theorem mithilfe eines indirekten Beweises (Beweis durch Widerspruch). Wir haben schon gesehen, dass $\nabla f(x^*) = 0_n$ ist, wenn x^* eine lokale Minimalstelle von f. ist. Für einen Widerspruchsbeweis nehmen wir nun an, dass x^* zwar eine lokale Minimalstelle von f ist, aber dass $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Dann ist es möglichen einen Vektor p zu finden, so dass gilt

$$p^T \nabla^2 f(x^*) p < 0. \tag{38}$$

Weil $\nabla^2 f(x^*)$ in einer Umgebung von x^* stetig ist, existiert ein Skalar $T > 0$, so dass

$$p^T \nabla^2 f(x^* + tp) p < 0 \text{ für alle } t \in [0, T]. \tag{39}$$

gilt. Mithilfe des Mittelwertsatzes zweiter Ordnung gilt dann für alle $\bar{t} \in]0, T]$ und ein $t \in]0, \bar{t}[$, dass

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t} p^T \nabla f(x^*) + \frac{1}{2} \bar{t}^2 p^T \nabla^2 f(x^* + tp) p < f(x^*). \tag{40}$$

Wir haben also wieder eine Richtung von x^* weg gefunden, in der f abnimmt. Also kann x^* keine Minimalstelle sein, wenn $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Dies ist aber ein Widerspruch, zur Annahme, dass es möglich ist, dass x^* eine lokale Minimalstelle von f ist und $\nabla^2 f(x^*)$ nicht positiv semidefinit ist. Also muss $\nabla^2 f(x^*)$ positiv semidefinit sein, wenn x^* eine lokale Minimalstelle ist.

Theorem (Hinreichende Bedingungen zweiter Ordnung)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es seien $\nabla f(x^*) = 0_n$ und $\nabla^2 f(x^*)$ positiv definit. Dann ist x^* eine strikte Minimalstelle von f.

Beweis

Wir halten zunächst fest, dass weil die Hesse-Matrix stetig und positiv definit in x^* ist, wir ein $r > 0$ wählen können, so dass $\nabla^2 f(x)$ positiv definit für alle x in

$$D = \{x | ||x - x^*|| < r\} \tag{41}$$

ist. Für einen Vektor p mit $||p|| > 0$ und $||p|| < r$ gilt $x^* + p \in D$. Für ein $t \in]0, 1[$ gilt dann mit dem Mittelwertssatz zweiter Ordnung, dass

$$\begin{aligned} f(x^* + p) &= f(x^*) + \nabla f(x^*) p^T + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p \\ &= f(x^*) + \frac{1}{2} p^T \nabla^2 f(x^* + tp) p. \end{aligned} \tag{42}$$

Weil aber $x^* + tp \in D$ ist, gilt, dass $p^T \nabla^2 f(x^* + tp) p > 0$ ist und somit $f(x^* + p) > f(x^*)$. In jeder Richtung p von x^* weg erhöht sich also der Wert von f und damit ist x^* eine strikte Minimalstelle.

Zusammenfassung

Optimierungsproblem

$$\min_x f(x) = \max_x -f(x) \text{ für } f : \mathbb{R}^n \to \mathbb{R}$$

Lokale Minimalstelle

$$x^* = \arg\min_x f(x), x^* \in \mathbb{R}^n \Leftrightarrow f(x^*) \leq f(x) \text{ für alle } x \in N \subset \mathbb{R}^n$$

Notwendige Bedingung erster Ordnung

$$x^* = \arg\min_x f(x) \Rightarrow \nabla f(x^*) = 0_n$$

Notwendige Bedingung zweiter Ordnung

$$x^* = \arg\min_x f(x) \Rightarrow \nabla f(x^*) = 0_n \text{ und } \nabla^2 f(x^*) \text{ positiv semidefinit}$$

Hinreichende Bedingung zweiter Ordnung

$$\nabla f(x^*) = 0_n \text{ und } \nabla^2 f(x^*) \text{ positiv definit } \Rightarrow x^* = \arg\min_x f(x)$$

Multivariate Differentialrechnung

Grundlagen der nichtlinearen Optimierung

Gradientenverfahren

Selbstkontrollfragen

Allgemeine Form von Optimierungsalgorithmen

Initialisierung

0. Wahl eines Startpunktes $x_0 \in \mathbb{R}^n$.

Iterationen

Für $k = 0, 1, 2, ...$

1. Berechnung von x_{k+1} basierend auf Information über f an der Stelle x_k.
2. STOP, wenn Minimalstelle gefunden ist oder kein Fortschritt mehr erzielt wird.

Definition (Gradientenverfahren)

Es sei $f : \mathbb{R}^n \to \mathbb{R}$ eine multivariate reellwertige Funktion. Dann hat das *Gradientenverfahren* zur Minimierung von f folgende allgemeine Form:

Initialisierung

0. Wähle einen Startpunkt $x_0 \in \mathbb{R}^n$, eine Lernrate $\alpha > 0$ und ein Konvergenzkriterium $\delta > 0$.

Iterationen

Für $k = 0, 1, 2, ...$

1. Setze $x_{k+1} := x_k - \alpha \nabla f(x_k)$.
2. STOP, wenn $||\nabla f(x_{k+1})|| < \delta$, ansonsten gehe zu 1.

Bemerkungen

- Die Lernrate α bestimmt, wie weit ein Schritt in Richtung des Gradienten erfolgt.
- Das Konvergenzkriterium bestimmt, wie klein der Gradient sein muss, damit das Verfahren endet.

Theorem (Gradientenverfahren)

$f : \mathbb{R}^n \to \mathbb{R}$ sei eine glatte Funktion und es sei $x_k \in \mathbb{R}^n$. Dann ist die Gradientenrichtung

$$p_k^G := -\nabla f(x_k) \tag{43}$$

die Richtung des steilsten Abstiegs von f in x_k.

Bemerkungen

- Es gibt unendliche viele mögliche Richtungen p in x_k.
- $\nabla f(x) \in \mathbb{R}^n$ ist eine Richtung in der Definitionsmenge von f (Parameterraum).
- Die Gradientenrichtung ist davon die Richtung, in der die Zielfunktion f am schnellsten abnimmt.
- Zum Vergleich von Richtungen genügt es, Richtungen der Länge $||p|| = 1$ zu vergleichen.

Beweis

Mit dem Mittelwertsatz zweiter Ordnung gilt für jede Richtung p und Schrittlängenparameter α, dass

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f(x_k) + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k + tp)p \text{ für ein } t \in]0, \alpha[. \tag{44}$$

Die Änderungsrate von f in Richtung p in x_k ist also der Koeffizient von α, also $p^T \nabla f(x_k)$ (man denke an $x = tv$ für einen Ort x, eine Geschwindigkeit v und eine Zeit t). Also gilt, dass die Richtung des steilsten Abstiegs p in x_k mit Länge 1 die Lösung des Optimierungsproblems

$$\min_p p^T \nabla f(x_k) \text{ mit der Nebenbedingung } ||p|| = 1. \tag{45}$$

ist. Wir erinnern nun zunächst daran, dass für $x, y \in \mathbb{R}^n$ gilt der Kosinus des Winkel zwischen x und y durch

$$\cos\alpha = \frac{\langle x, y\rangle}{||x||||y||} = \frac{x^T y}{||x||||y||} \tag{46}$$

gegeben ist. Damit aber gilt, dass

$$p^T \nabla f(x_k) = ||p|| \cdot ||\nabla f(x_k)|| \cos\theta = 1 \cdot ||\nabla f(x_k)|| \cos\theta = ||\nabla f(x_k)|| \cos\theta \tag{47}$$

und somit liegt hier bei $\cos\theta = -1$ eine Minimalstelle vor. Dies bedeutet aber, dass die minimierende Länge p exakt antiparallel zu $\nabla f(x_k)$ und von Länge 1 sein muss. Also ist die Minimalstelle des Optimierungsproblems

$$p = \frac{-\nabla f(x_k)}{||\nabla f(x_k)||}. \tag{48}$$

Damit ist $p_k^G := -\nabla f(x_k)$ aber der Richtungsvektor beliebiger Länge in der die Abnahme von f maximal ist.

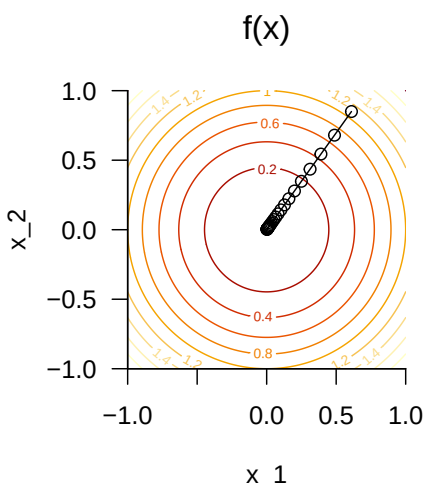
Beispiel

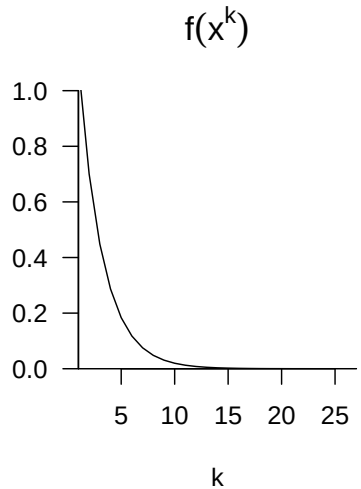
Minimierung von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$

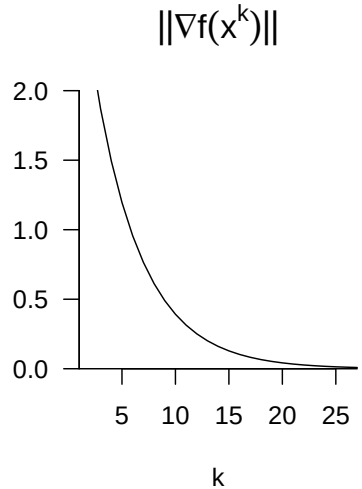
```
# Funktionsdefinitionen
# ------------------------------------------------------------------------------
# Zielfunktion
f = function(x) {
  return(x[1]^2 + x[2]^2)                # f(x) := x_1^2 + x_2^2
}
# Gradient der Zielfunktion
nabla_f = function(x) {
  return(matrix(c(2*x[1],2*x[2]),        # \nabla f(x) := (2x_1, 2x_2)^T
                nrow = 2))
}
# Gradientenverfahren
# ------------------------------------------------------------------------------
# Parameter
n      = 2                             # Dimension
alpha  = 1e-1                          # Lernrate
delta  = 1e-2                          # Konvergenzkriterium

# Initialisierung
x_k    = matrix(c(.61,.85), nrow = 2)  # Zufälliger Startpunkt in [0,1]^2
x      = x_k                           # Initialisierung Iteranden
fx     = f(x_k)                        # Initialisierung Funktionswerte
crt    = norm(nabla_f(x_k))            # Initialisierung Kriterium

# Iterationen
while(norm(nabla_f(x_k)) > delta){

  # Argumentupdate
  x_k    = x_k - alpha*nabla_f(x_k)

  # Dokumentation
  x     = cbind(x,x_k)
  fx    = c(fx,f(x_k))
  crt   = c(crt, norm(nabla_f(x_k)))
}
```

Beispiel

Minimierung von $f : \mathbb{R}^2 \to \mathbb{R}, x \mapsto f(x) := x_1^2 + x_2^2$







Liniensuchverfahren als generalisierte Gradientenverfahren

Initialisierung

0. Wahl eines Startpunktes $x_0 \in \mathbb{R}^n$.

Iterationen

Für $k = 0, 1, 2, ...$

1. Wahl einer Abstiegsrichtung p_k
2. Wahl eines Lernparameters $\alpha_k \approx \min_\alpha f(x_k + \alpha p_k)$.
3. Setze $x_{k+1} := x_k + \alpha_k p_k$.
4. Konvergenztest.

$\Rightarrow$ Die Wahl sinnvoller Lernraten α_k ist für eine gute Performanz entscheidend!

(vgl. Ostwald and Starke (2016))

Selbstkontrollfragen

1. Geben Sie die Definition einer multivariaten reellwertigen Funktion wieder.
2. Geben Sie die Definition der partiellen Ableitung wieder.
3. Geben Sie die Definition der zweiten partiellen Ableitung wieder.
4. Geben Sie den Satz von Schwarz wieder.
5. Geben Sie die Definition eines Gradienten einer multivariaten reellwertigen Funktion wieder.
6. Geben Sie die Definition der Hesse-Matrix einer multivariaten reellwertigen Funktion wieder.
7. Geben Sie die allgemeine Form eines Optimierungsproblems wieder.
8. Geben Sie die notwendige Bedingung erster Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
9. Geben Sie die notwendige Bedingung zweiter Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
10. Geben Sie die hinreichende Bedingung zweiter Ordnung für ein Minimum von $f : \mathbb{R}^n \to \mathbb{R}$ an.
11. Geben Sie die Definition des Gradientverfahrens wieder.
12. Erläutern Sie die Bedeutung der Lernrate und des Konvergenzkriteriums im Gradientenverfahren.
13. Geben Sie das Theorem zum Gradientenverfahren wieder.

Referenzen

Murphy, Kevin P. 2023. *Probabilistic Machine Learning: Advanced Topics*. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press.

Nocedal, Jorge, and Stephen J. Wright. 2006. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research. New York: Springer.

Ostwald, Dirk, and Ludger Starke. 2016. "Probabilistic Delay Differential Equation Modeling of Event-Related Potentials." *NeuroImage* 136 (August): 227–57. https://doi.org/10.1016/j.neuroimage.2016.04.025.

Prince, Simon J. D. 2023. *Understanding Deep Learning*. Cambridge, Massachusetts: The MIT Press.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(12) Logistische Regression

Anwendungsszenario

Modellformulierung

Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

Modellformulierung

Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie Non-Response-Rate wird auf etwa 20 - 30% geschätzt

Vorhersage von Behandlungserfolg basierend auf klinischen Markern wäre hilfreich

- Therapieauswahloptimierung
- Lebensqualitätverbesserung
- Ressourcensensitivität

Digitale Datenbank von Psychotherapieverläufen als Trainingsdatensatz

Prädiktive Modellierung zur Etablierung eines prädiktiven klinischen Markerprofils

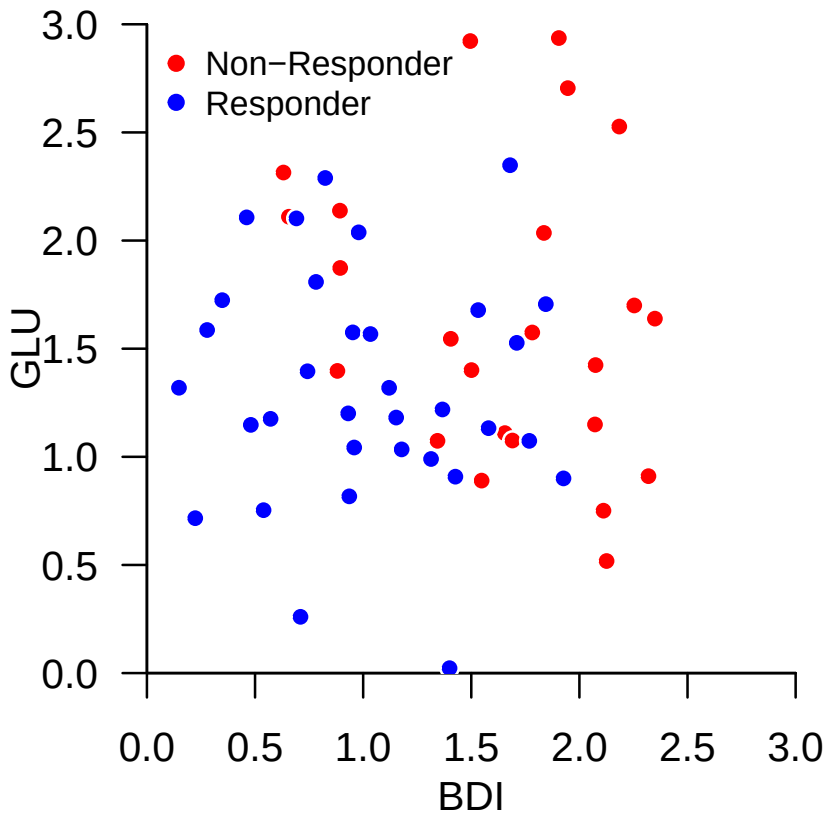
Treatmentsuccessvorhersage für neue Patient:innen

Anwendungsbeispiele

- BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg
- Lineare Diskriminanzanalyse, Logistische Regression, Neuronale Netze

Anwendungsszenario

BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg



Anwendungsszenario

Modellformulierung

Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Definition (Generalisiertes Lineares Modell)

$x \in \mathbb{R}^{m+1}$ sei ein erweiterter Featurevektor und υ das assoziierte Label. Weiterhin sei für einen *Parametervektor* $\beta \in \mathbb{R}^{m+1}$

$$\eta := x^T \beta \tag{1}$$

ein *linearer Prädiktor*. Dann ist ein generalisierte lineares Modell definiert mithilfe einer zweimal differenzierbaren und invertierbaren *Link-Funktion*

$$g : \mathbb{R} \to \mathbb{R}, \mathbb{E}(\upsilon) \mapsto g(\mathbb{E}(\upsilon)) =: \eta. \tag{2}$$

definiert. Die Inverse der Link-Funktion,

$$g^{-1} : \mathbb{R} \to \mathbb{R}, \eta \mapsto g^{-1}(\eta) = \mathbb{E}(\upsilon) \tag{3}$$

heißt *Mean-Funktion* und wird mit f bezeichnet, so dass

$$f : \mathbb{R} \to \mathbb{R}, \eta \mapsto f(\eta) = \mathbb{E}(\upsilon). \tag{4}$$

Definition (ALM als Generalisiertes Lineares Modell)

Das Allgemeine Lineare Modell mit u.i.v. Störvariablen ist das Generalisierte Lineare Model, bei dem

1. die Labelvariable eine univariat normalverteilte Zufallsvariable

$$\upsilon \sim N(\mu, \sigma^2), \tag{5}$$

   ist und

2. die Link-Funktion durch die Identität

$$g : \mathbb{R} \to \mathbb{R}, \mu \mapsto g(\mu) := \mu =: \eta. \tag{6}$$

   gegeben ist.

Weil die Inverse der Identität wiederrum die Identität ist, folgt, dass die Mean-Funktion des ALM durch

$$f : \mathbb{R} \to \mathbb{R}, \eta \mapsto f(\eta) = \eta = \mu. \tag{7}$$

gegeben ist. Die Parameter des Allgemeinen Linearen Modells sind die Komponenten des Vektors $\beta \in \mathbb{R}^m$ des linearen Prädiktors $\eta = x^T\beta$ und der Parameter $\sigma^2 > 0$.

Definition (Logistische Regression als Generalisiertes Lineares Modell)

Das Modell der Logistischen Regression (LR) ist das Generalisierte Lineare Modell, bei dem

1. die Labelvariable eine Bernoulli-Zufallsvariable

$$\upsilon \sim B(\mu) \tag{8}$$

   ist und

2. die Link-Funktion durch die *standard logit function*

$$g : [0,1] \to \mathbb{R}, \mu \mapsto g(\mu) := \ln\left(\frac{\mu}{1-\mu}\right) =: \eta \tag{9}$$

   gegeben ist.

Die Parameter des Logistischen Regressionsmodels sind die Komponenten des Vektors $\beta \in \mathbb{R}^m$ des linearen Prädiktors $\eta = x^T\beta$.

Theorem (Mean-Funktion der Logistischen Regression)

Die Inverse der Link-Funktion des Modells der Logistischen Regression und somit seine Mean-Funktion ist die *Logistische Standardfunktion*

$$f : \mathbb{R} \to [0,1], \eta \mapsto f(\eta) = \frac{1}{1+\exp(-\eta)}. \tag{10}$$

Beweis

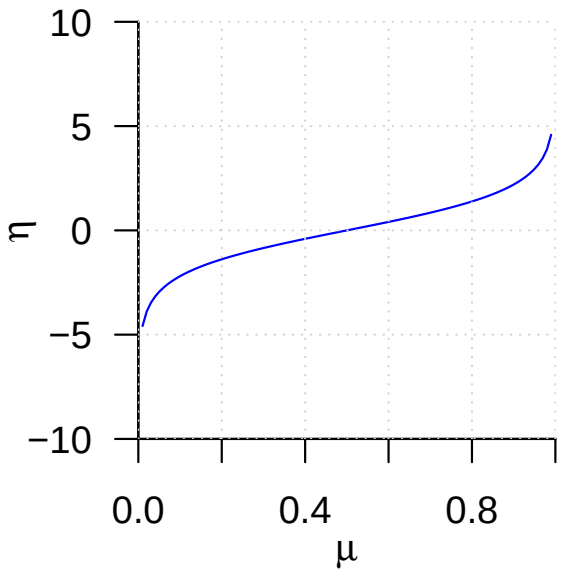
Umformen der logit function ergibt

$$\begin{aligned}
\eta &= \ln(\mu/(1-\mu)) \\
\Leftrightarrow -\eta &= -\ln(\mu/(1-\mu)) \\
\Leftrightarrow -\eta &= \ln((1-\mu)/\mu) \\
\Leftrightarrow \exp(-\eta) &= (1-\mu)/\mu \\
\Leftrightarrow \mu\exp(-\eta) &= 1-\mu \\
\Leftrightarrow \exp(-\eta) &= \mu^{-1} - 1 \\
\mu &= 1/(\exp(-\eta)+1)
\end{aligned}$$

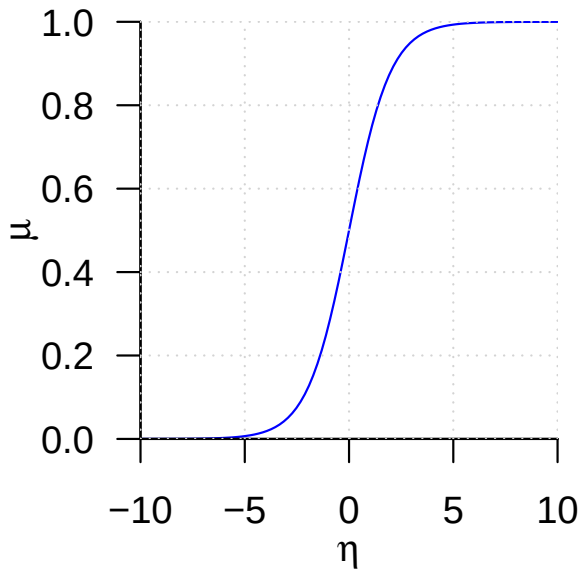
□

Link und Mean Funktionen

$$g(\mu) := \ln(\mu / 1 - \mu)$$



$$f(\eta) := 1/(1 + \exp(-\eta))$$



Definition (Modell der Logistischen Regression)

υ sei eine Zufallsvariable mit Ergebnisraum $\{0, 1\}$. Dann ist das *Modell der Logistischen Regressions* definiert als die WMF von υ

$$p(y) = B\left(y; \frac{1}{1 + \exp(-x^T \beta)}\right), \tag{11}$$

wobei $x \in \mathbb{R}^{m+1}$ einen erweiterten Featurevektor und $\beta \in \mathbb{R}^{m+1}$ den *Parametervektor* bezeichnen

Bemerkung

- Aus generativer Sicht wird ein Trainingsdatensatz

$$\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n} \text{ mit } x^{(i)} \in \mathbb{R}^{m+1} \text{ und } y^{(i)} \in \{0, 1\} \tag{12}$$

  eines LR Modells wie folgt erzeugt:

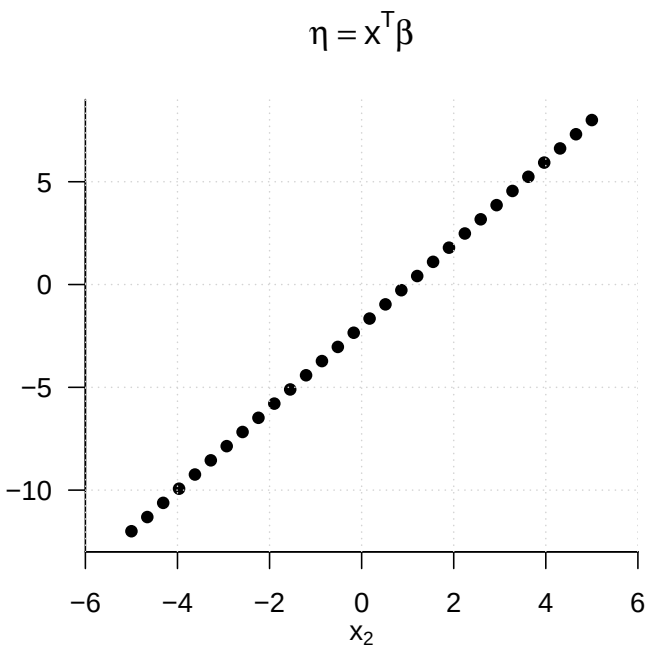
(1) Definition von $x^{(i)}$,
(2) Ziehen von $y^{(i)}$ aus $p(y) = B(y; \mu)$ mit Erwartungswertparameter $\mu = \frac{1}{1+\exp(-x^{(i)T} \beta)}$.

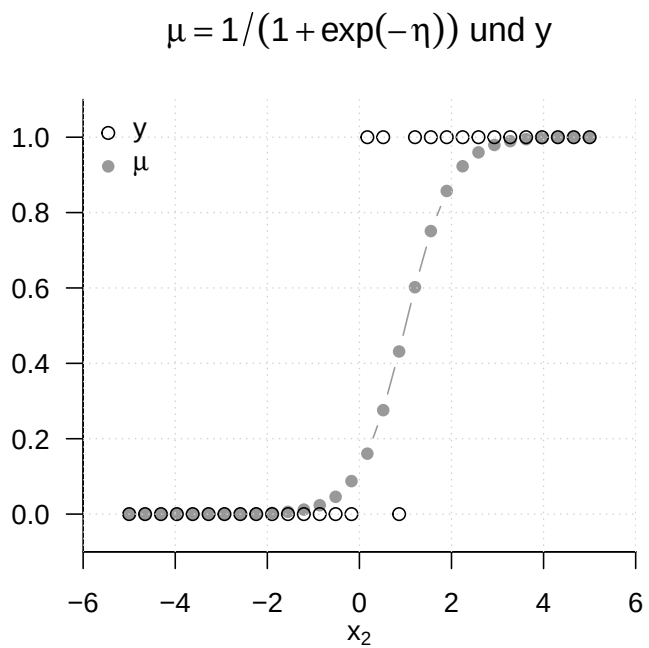
Modellformulierung

Datengeneration bei einfacher Logistischer Regression für $m = 1, \beta = (-2, 2)^T, n = 30$.

```
# Modellparameter
m    = 1                              # Featurevektoredimensionalität
n    = 30                             # Anzahl Datenpunkte
x    = matrix(c(rep(1,n),              # Definition des erweiterten Featurevektors
                seq(-5,5, len = n)),
              nrow  = 2,
              byrow = TRUE)
beta = matrix(c(-2,2), nrow = 2)      # wahrer, aber unbekannter, Parametervektor
eta  = t(x) %*% beta                  # wahrer, aber unbekannter linearer Prädiktor
mu   = 1/(1+exp(-eta))                # wahrer, aber unbekannter, Bernoulliparametervektor

# Dategeneration
set.seed(2)                           # Zufallsgeneratorzustand
y    = rep(NaN,2)                     # Datenarray
for(i in 1:n){
  y[i] = rbinom(1,1,mu[i])            # Bernoullivariablenrealisierung
}
```

Datengeneration bei einfacher Logistischer Regression für $m = 1, \beta = (-2, 2)^T, n = 30$.





Anwendungsszenario

Modellformulierung

Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Definition (Klassifikationsregel der Logistischen Regression)

$p(y)$ sei die WMF eines Logistischen Regressionsmodells. Dann ist die *Klassifikationregel* definiert als

$$\delta : \mathbb{R}^m \to \{0,1\}, x \mapsto \delta(x) := \begin{cases} 0 & \text{für } p(y=0) \geq p(y=1) \\ 1 & \text{für } p(y=0) < p(y=1) \end{cases} \tag{13}$$

Bemerkung

- Es gilt

$$\delta(x) = 1 \Leftrightarrow p(y=1) > p(y=0) \Leftrightarrow p(y=1) > 0.5. \tag{14}$$

Anwendungsszenario

Modellformulierung

Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Theorem (Log-Likelihood-Funktion der Logistischen Regression)

$\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ sei ein Trainingsdatensatz aus erweiterten Featurevektoren und assoziierten Labelvariablenrealisierungen und f sei die Logistische Standardfunktion. Dann hat die Log-Likelihood-Funktion der Logistischen Regression die Form

$$\ell : \mathbb{R}^m \to \mathbb{R}, \beta \mapsto \ell(\beta) := \sum_{i=1}^{n} y^{(i)} \ln(f({x^{(i)}}^T \beta)) + (1 - y^{(i)}) \ln(1 - f({x^{(i)}}^T \beta)).$$

Beweis

Wir halten zunächst fest, dass für u.i.v. Labelvariablen gilt, dass

$$\ell(\beta) := \ln p(y^{(1)}, ..., y^{(n)}) = \ln \prod_{i=1}^{n} p(y^{(i)}) = \sum_{i=1}^{n} \ln p(y^{(i)}) \tag{15}$$

Mit der WMF der Bernoulliverteilung folgt dann

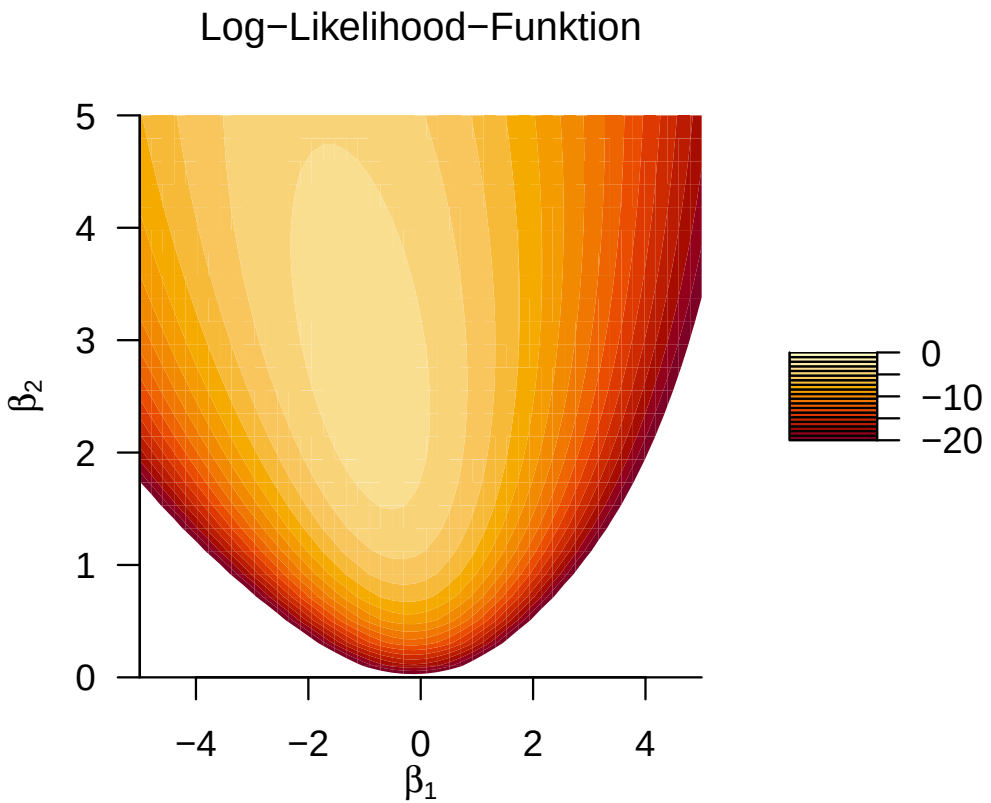
$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{n} \ln \left(f\left({x^{(i)}}^T \beta\right)^{y^{(i)}} \left(1 - f\left({x^{(i)}}^T \beta\right)\right)^{1-y^{(i)}} \right) \\ &= \sum_{i=1}^{n} y^{(i)} \ln \left(f\left({x^{(i)}}^T \beta\right) \right) + \left(1 - y^{(i)}\right) \ln \left(1 - f\left({x^{(i)}}^T \beta\right)\right) \end{aligned} \tag{16}$$

Lernen

Log-Likelihood-Funktion des generierten Datensatzes für $m = 1, \beta = (-2, 2)^T, n = 30$.

```
# Funktionsdefinitionen
# ------------------------------------------------------------------------------------------
# Standard Logistic Function
f = function(eta){
    return(1/(1 + exp(-eta)))
}

# Log Likelihood Function
llh = function(x,y,beta){
    n   = ncol(x)
    ell = 0
    for(i in 1:n){
        ell = ell + y[i]*log(f(t(x[,i]) %*% beta)) + (1-y[i])*log(1-f(t(x[,i]) %*% beta))
    }
    return(ell)
}

# Log-Likelihood-Funktion Auswertung
# ------------------------------------------------------------------------------------------
beta_min  = -5                                                    # beta Minimum
beta_max  =  5                                                    # beta Maximum
beta_res  = 5e1                                                   # beta Auflösung
beta_1    = seq(beta_min, beta_max, length.out = beta_res)        # beta_1 Raum
beta_2    = seq(beta_min, beta_max, length.out = beta_res)        # beta_2 Raum
ell       = matrix(rep(NaN, beta_res*beta_res), nrow = beta_res) # Log-Likelihood-Funktion Array
for(i in 1:beta_res){
    for(j in 1:beta_res){
        beta12   = matrix(c(beta_1[i], beta_2[j]), nrow = 2)
        ell[i,j] = llh(x,y,beta12)
    }
}
```

Log-Likelihood-Funktion des generierten Datensatzes für $m = 1, \beta = (-2, 2)^T, n = 30$.



Theorem (Gradientverfahren der Logistischen Regression)

Gegeben sei das Modell einer Logistischen Regression und $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ sei ein entsprechender Trainingsdatensatz. Dann kann ein Maximum-Likelihood-Schätzer $\hat{\beta}$ für den Parametervektor β des Logistischen Regressionsmodell durch folgendes Gradientenverfahren gewonnen werden:

(0) Wähle $\beta^0 \in \mathbb{R}, \alpha > 0, \delta > 0$

(1) Für $k = 0, 1, 2, \ldots$ bis zur Konvergenz setze

$$\beta^{(k+1)} := \beta^{(k)} + \alpha \nabla \ell(\beta^{(k)}). \tag{17}$$

wobei $\nabla \ell \left(\beta^k\right)$ den Gradienten der Log-Likelihood-Funktion der Logistischen Regression bezeichnet und die Form

$$\nabla \ell(\beta) = \begin{pmatrix} \frac{\partial}{\partial \beta_1} \ell(\beta) \\ \frac{\partial}{\partial \beta_2} \ell(\beta) \\ \vdots \\ \frac{\partial}{\partial \beta_m} \ell(\beta) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} (y^{(i)} - f({x^{(i)}}^T \beta)) x_1^{(i)} \\ \sum_{i=1}^{n} (y^{(i)} - f({x^{(i)}}^T \beta)) x_2^{(i)} \\ \vdots \\ \sum_{i=1}^{n} (y^{(i)} - f({x^{(i)}}^T \beta)) x_m^{(i)} \end{pmatrix} \tag{18}$$

Bemerkungen

- Das reine Gradientenverfahren zum Lernen der Parameter eines LR Modells ist recht instabil.
- Iteratively Weighted Least Squares Verfahren werden zur ML Schätzung in GLMs bevorzugt (Green (1984)).
- IWLS Verfahren nutzen Gradienten und Hesse-Matrix ähnlich wie Gauss-Newton Verfahren.
- R implementiert in der `glm()` ein IWLS Verfahren.

Beweis

Um die $j$te partielle Ableitung der Log-Likelihood-Funktion zu bestimmen, halten wir zunächst fest, dass sich die Ableitung der logistic function f hinsichtlich η zu

$$\frac{d}{d\eta} f : \mathbb{R} \to \mathbb{R}, \eta \mapsto \frac{d}{d\eta} f(\eta) = f(\eta)(1 - f(\eta)) \tag{19}$$

ergibt. Dies kann wie folgt eingesehen werden:

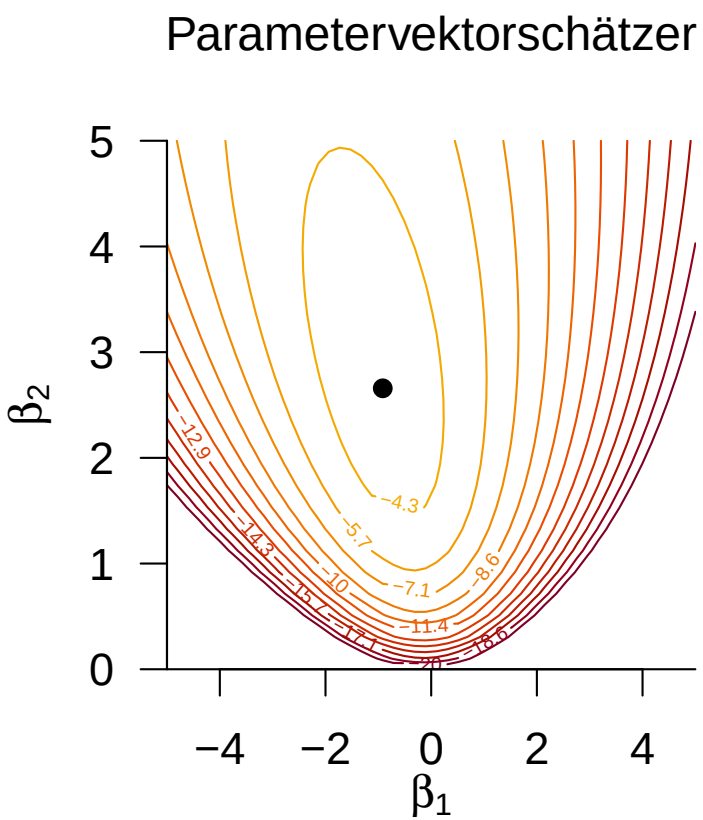
$$\begin{aligned}
\frac{d}{d\eta} f(\eta) &= \frac{d}{d\eta} \left(1 + \exp(-\eta)\right)^{-1} \\
&= -\left(1 + \exp(-\eta)\right)^{-2} \cdot \exp(-\eta) \cdot (-1) \\
&= \frac{\exp(-\eta)}{(1 + \exp(-\eta))^2} \\
&= \frac{1 + \exp(-\eta) - 1}{(1 + \exp(-\eta))^2} \\
&= \frac{1 + \exp(-\eta)}{(1 + \exp(-\eta))^2} - \frac{1}{(1 + \exp(-\eta))^2} \\
&= \frac{1}{1 + \exp(-\eta)} - \frac{1}{(1 + \exp(-\eta))^2} \\
&= \frac{1}{1 + \exp(-\eta)} \left(1 - \frac{1}{1 + \exp(-\eta)}\right) \\
&= f(\eta)(1 - f(\eta))
\end{aligned}$$

Damit ergibt sich dann für $\frac{\partial}{\partial \beta_j} \ell, j = 1, \ldots, m$:

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j} \ell(\beta) &= \frac{\partial}{\partial \beta_j} \left(\sum_{i=1}^{n} y^{(i)} \ln\left(f\left({x^{(i)}}^T \beta \right) \right) + \left(1 - y^{(i)}\right) \ln\left(1 - f\left({x^{(i)}}^T \beta \right) \right) \right) \\
&= \sum_{i=1}^{n} y^{(i)} \frac{\partial}{\partial \beta_j} \left(\ln\left(f\left({x^{(i)}}^T \beta \right) \right) \right) + \left(1 - y^{(i)}\right) \frac{\partial}{\partial \beta_j} \left(\ln\left(1 - f\left({x^{(i)}}^T \beta \right) \right) \right) \\
&= \sum_{i=1}^{n} y^{(i)} \frac{1}{f\left({x^{(i)}}^T \beta \right)} \left(\frac{\partial}{\partial \beta_j} \left(f\left({x^{(i)}}^T \beta \right) \right) \right) + \left(1 - y^{(i)}\right) \frac{1}{1 - f\left({x^{(i)}}^T \beta \right)} \frac{\partial}{\partial \beta_j} \left(1 - f\left({x^{(i)}}^T \beta \right) \right) \\
&= \sum_{i=1}^{n} \left(y^{(i)} \frac{1}{f\left({x^{(i)}}^T \beta \right)} - \left(1 - y^{(i)}\right) \frac{1}{1 - f\left({x^{(i)}}^T \beta \right)} \right) \frac{\partial}{\partial \beta_j} \left(f\left({x^{(i)}}^T \beta \right) \right) \\
&= \sum_{i=1}^{n} \left(y^{(i)} \frac{1}{f\left({x^{(i)}}^T \beta \right)} - \left(1 - y^{(i)}\right) \frac{1}{1 - f\left({x^{(i)}}^T \beta \right)} \right) f\left({x^{(i)}}^T \beta \right) \left(1 - f\left({x^{(i)}}^T \beta \right) \right) \frac{\partial}{\partial \beta_j} \left({x^{(i)}}^T \beta \right) \\
&= \sum_{i=1}^{n} \left(y^{(i)} \frac{1}{f\left({x^{(i)}}^T \beta \right)} - \left(1 - y^{(i)}\right) \frac{1}{1 - f\left({x^{(i)}}^T \beta \right)} \right) f\left({x^{(i)}}^T \beta \right) \left(1 - f\left({x^{(i)}}^T \beta \right) \right) x_j^{(i)} \\
&= \sum_{i=1}^{n} \left(y^{(i)} \frac{f\left({x^{(i)}}^T \beta \right) \left(1 - f\left({x^{(i)}}^T \beta \right) \right)}{f\left({x^{(i)}}^T \beta \right)} - \left(1 - y^{(i)}\right) \frac{f\left({x^{(i)}}^T \beta \right) \left(1 - f\left({x^{(i)}}^T \beta \right) \right)}{1 - f\left({x^{(i)}}^T \beta \right)} \right) x_j^{(i)} \\
&= \sum_{i=1}^{n} \left(y^{(i)} \left(1 - f\left({x^{(i)}}^T \beta \right) \right) - \left(1 - y^{(i)}\right) f\left({x^{(i)}}^T \beta \right) \right) x_j^{(i)} \\
&= \sum_{i=1}^{n} \left(y^{(i)} - y^{(i)} f\left({x^{(i)}}^T \beta \right) - f\left({x^{(i)}}^T \beta \right) + y^{(i)} f\left({x^{(i)}}^T \beta \right) \right) x_j^{(i)} \\
&= \sum_{i=1}^{n} \left(y^{(i)} - f\left({x^{(i)}}^T \beta \right) \right) x_j^{(i)}.
\end{aligned}
$$

Praktisches Parameterschätzung mit dem IWLS Verfahren in **R**

```
lr        = glm(y ~ x[2,], family = 'binomial')          # generalized linear model fit
beta_hat  = lr$coefficients                              # Parametervektorschätzer
```



Anwendungsszenario

Modellformulierung

Klassifikation

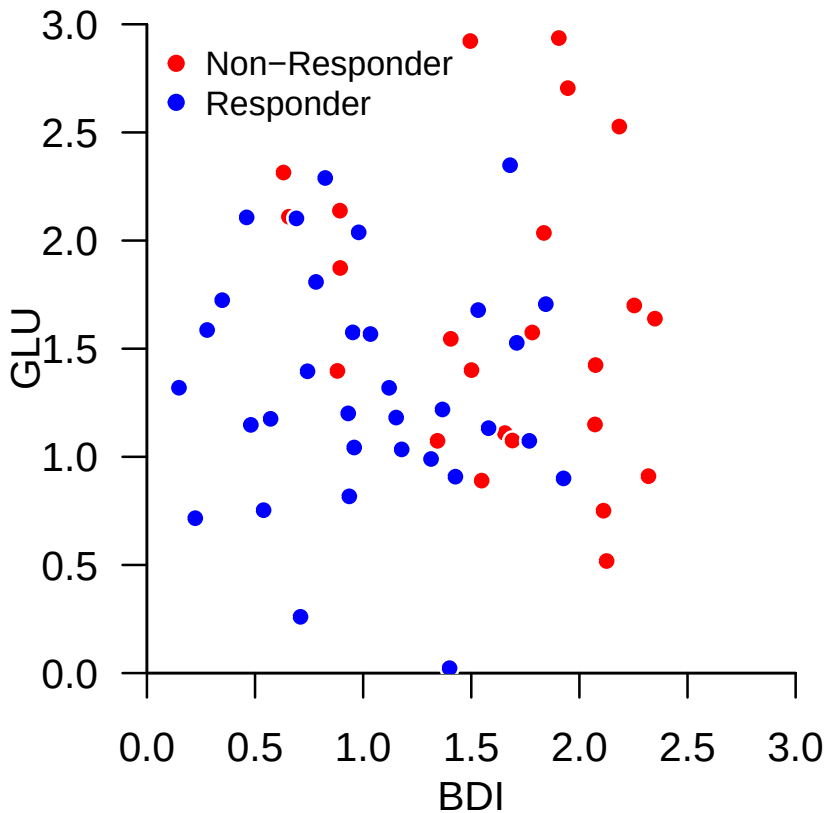
Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsbeispiel

BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg



Anwendungsbeispiel

`BDI` und `GLU` Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg `RES`

BDI	GLU	RES
0.74	1.40	1
0.22	0.72	1
0.82	2.29	1
2.07	1.15	0
1.71	1.53	1
1.77	1.07	1
1.95	2.70	0
2.18	2.53	0
0.93	1.20	1
1.34	1.07	0
2.35	1.64	0
1.43	0.91	1
1.66	1.11	0
0.28	1.59	1
2.13	0.52	0
1.37	1.22	1
0.89	2.14	0
0.88	1.40	0
0.98	2.04	1
1.93	0.90	1

LOOCV zur Bestimmung der Featureprädiktivität

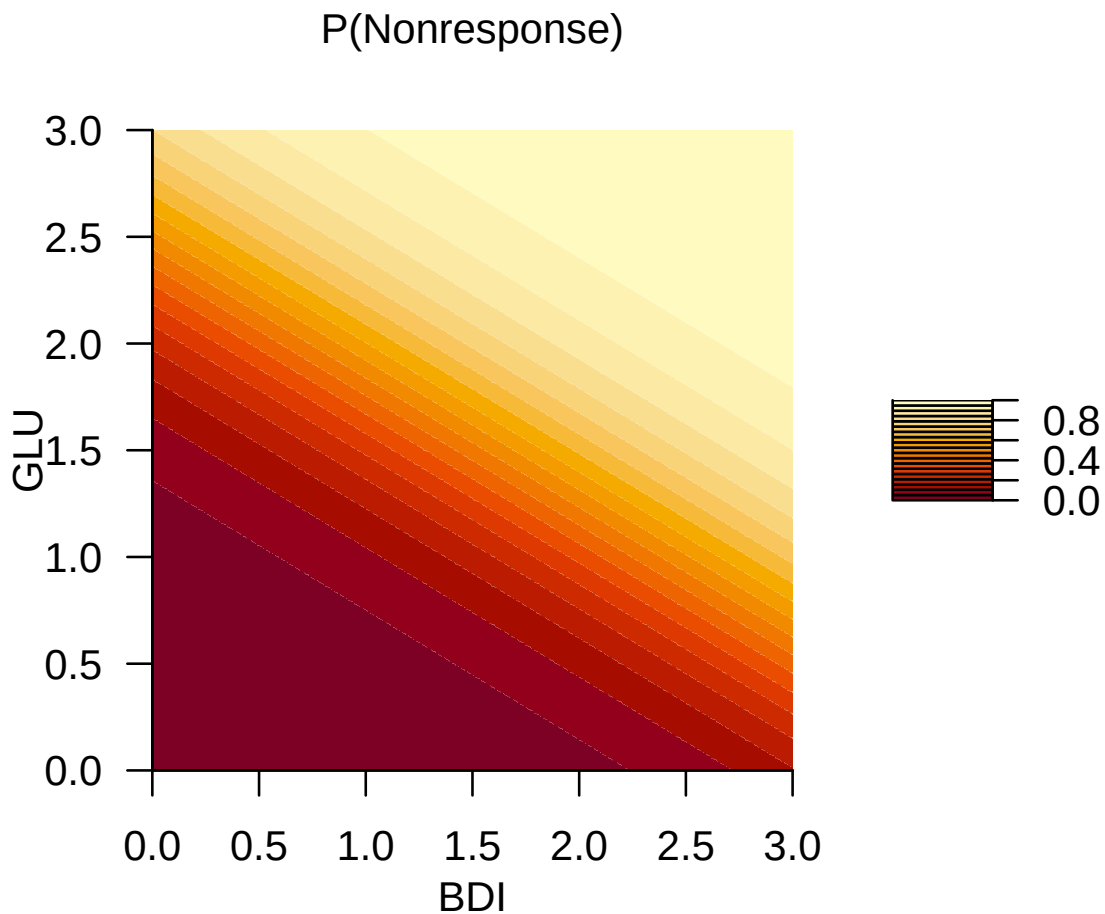
```
D       = read.csv("./12_Daten/12_Logistische_Regression.csv")                  # Datensatz
K       = nrow(D)                                                               # Anzahl Cross Folds
p_y     = matrix(rep(NaN, K)  , nrow = 1)                                       # p(y = 1)
y_pred  = matrix(rep(NaN, K*2), nrow = K)                                       # Prädiktionsperformancearray
for(k in 1:K){                                                                  # K-fold LOOCV
    x_train      = t(D[-k,1:2])                                                 # Trainingsdatensatzfeatures
    y_train      = t(D[-k,3  ])                                                 # Trainingsdatensatzlabels
    x_test       = t(D[ k,1:2])                                                 # Testdatensatzfeaturevektor
    y_pred[k,1]  = t(D[ k,3])                                                   # Testdatensatzfeaturevektorlabel
    n            = ncol(x_train)                                                # n
    m            = nrow(x_train)                                                # m
    lr           = glm(t(y_train) ~ t(x_train), family = 'binomial')            # IWLS Parameterlernen
    beta_hat     = as.matrix(lr$coefficients, nrow = m + 1)                     # Parameterschätzer
    x_test_tilde = rbind(1, x_test)                                             # erweiteter Featurevektor
    p_y[k]       = 1/(1+exp(-t(x_test_tilde) %*% beta_hat))                     # p(y)
    y_pred[k,2]  = as.numeric(p_y[k] >= 0.5)                                    # Klassifikationsregel \delta
}
rp      = sum(y_pred[y_pred[,1] == 1,2] == 1)                                   # |(1,1)|
rn      = sum(y_pred[y_pred[,1] == 0,2] == 0)                                   # |(0,0)|
fp      = sum(y_pred[y_pred[,1] == 0,2] == 1)                                   # |(0,1)|
fn      = sum(y_pred[y_pred[,1] == 1,2] == 0)                                   # |(1,0)|
ACC     = (rp+rn)/(rp+fp+rn+fn)                                                 # Accuracy
SEN     = rp/(rp+fn)                                                            # Sensitivity
SPE     = rn/(rn+fp)                                                            # Specificity
cat("Accuracy   : "  , ACC, ", Sensitivity: " , SEN, ", Specificity: " , SPE)  # Ergebnisausgabe
```

```
Accuracy   :  0.7166667 , Sensitivity:  0.8235294 , Specificity:  0.5769231
```

Evaluation der Non-Response Wahrscheinlichkeit

```
D         = read.csv("./12_Daten/12_Logistische_Regression.csv")          # Datensatz
x         = t(D[,1:2])                                                    # Featurevektoren
y         = t(D[,3])                                                      # Label
n         = ncol(x)                                                       # n
m         = nrow(x)                                                       # m
lr        = glm(t(y_train) ~ t(x_train), family = 'binomial')             # IWLS Parameterlernen
beta_hat  = as.matrix(lr$coefficients, nrow = m + 1)                      # Parameterschätzer
x_min     = 0                                                             # GLU/BDI Minimum
x_max     = 3                                                             # GLU/BDI Maximum
x_res     = 5e2                                                           # GLU/BDI Auflösung
bdi       = seq(x_min, x_max, length.out = x_res)                         # BDI
glu       = seq(x_min, x_max, length.out = x_res)                         # GLU
p_y       = matrix(rep(NaN, x_res*x_res), nrow = x_res)                   # p_{(BDI, GLU}(y=1)
for(i in 1:x_res){                                                        # BDI Iterationen
    for(j in 1:x_res){                                                    # GLU Iterationen
        x_tilde   = rbind(1, bdi[i], glu[j])                              # \tilde{x}
        p_y[i,j]  = 1/(1+exp(-t(x_tilde) %*% beta_hat))}}                 # p_{(BDI, GLU}(y=1)
```

Evaluation der Non-Response Wahrscheinlichkeit



Anwendungsszenario

Modellformulierung

Klassifikation

Lernen

Anwendungsbeispiel

Selbstkontrollfragen

Selbstkontrollfragen

1. Geben Sie die Definition des Generalisierten Linearen Modells wieder.
2. Geben Sie die Definition der Logistischen Regression als Generalisiertes Lineares Modells wieder.
3. Geben Sie das Theorem zur Mean-Funktion der Logistischen Regression wieder.
4. Geben Sie die Definition des Modells der Logistischen Regression wieder.
5. Erläutern Sie die Generation von Daten unter dem Modell der Logistischen Regression.
6. Geben Sie die Definition der Klassifikationsregel der Logistischen Regression wieder.
7. Erläutern Sie wie mithilfe einer Logistischen Regression die psychotherapeutische Nonresponsewahrscheinlichkeit geschätzt werden kann.

Referenzen

Green, P. J. 1984. “Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives.” *Journal of the Royal Statistical Society: Series B (Methodological)* 46 (2): 149–70. https://doi.org/10.1111/j.2517-6161.1984.tb01288.x.

Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(13) Neuronale Netze

Anwendungsszenario

Funktionale Architektur

Lernen

Backpropagation

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

Funktionale Architektur

Lernen

Backpropagation

Anwendungsbeispiel

Selbstkontrollfragen

Anwendungsszenario

Psychotherapie Non-Response-Rate wird auf etwa 20 - 30% geschätzt

Vorhersage von Behandlungserfolg basierend auf klinischen Markern wäre hilfreich

- Therapieauswahloptimierung
- Lebensqualitätverbesserung
- Ressourcensensitivität

Digitale Datenbank von Psychotherapieverläufen als Trainingsdatensatz

Prädiktive Modellierung zur Etablierung eines prädiktiven klinischen Markerprofils

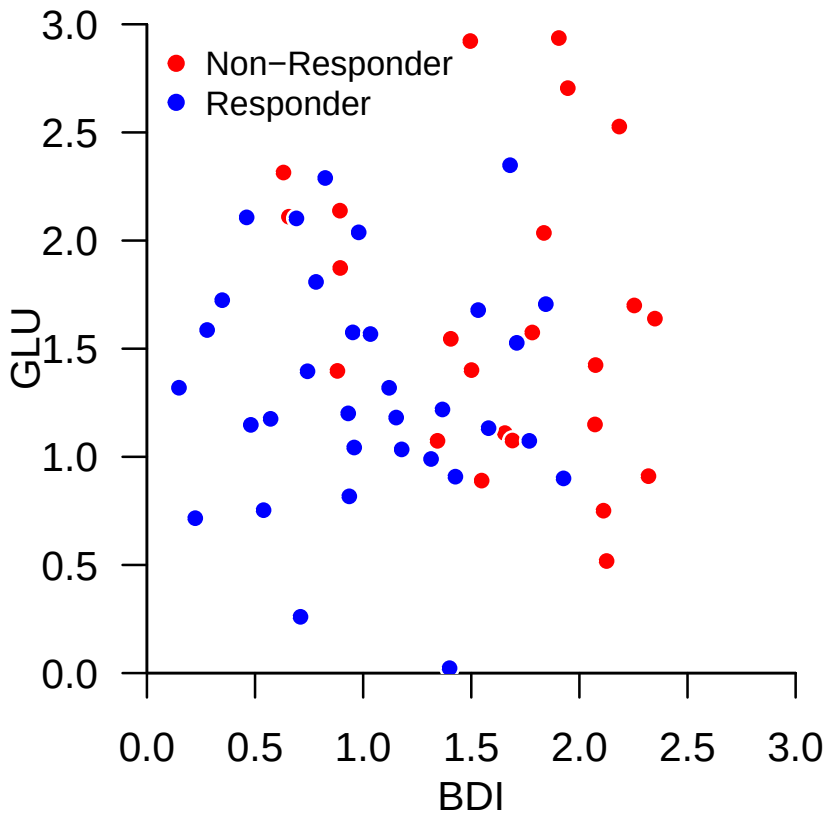
Treatmentsuccessvorhersage für neue Patient:innen

Anwendungsbeispiel

- BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg
- Lineare Diskriminanzanalyse, Logistische Regression, Neuronale Netze

Anwendungsszenario

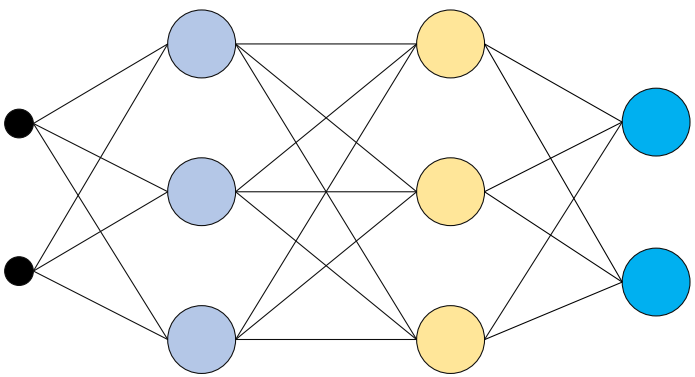
BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg

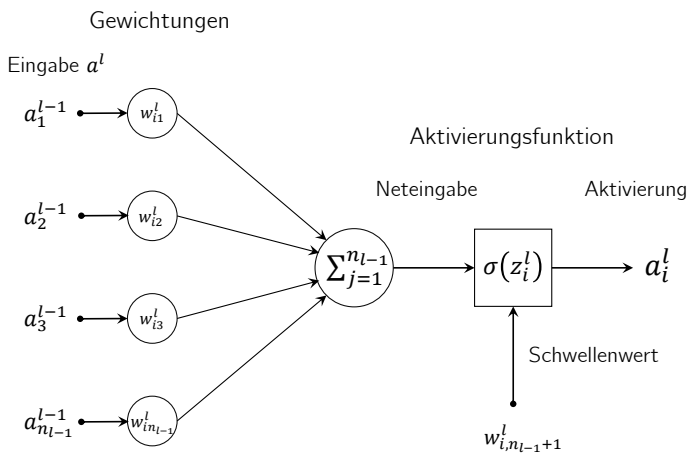


Neuronale Netze (Neural Networks)

- AKA *Künstliche Neuronale Netze (Artificial Neural Networks)*.
- Keine Modelle für biologische neuronale Netze.
- Mathematische Modelle zur Approximation multivariater vektorwertiger Funktionen.

Typische Visualisierungen





Zur Geschichte neuronaler Netze

Anfänge

- McCulloch and Pitts (1943) | Analyse der mit biologischen Neuronen möglichen logischen Operationen.
- Rosenblatt (1958) | Implementation eines Mustererkennungsalgorithmus in einem frühen Computer.
- Minsky and Papert (1969) | Mathematische Analyse der logischen Stärken und Schwächen eines Perzeptrons.

$\Rightarrow$ Erster Winter Neuronaler Netze

Erste Renaissance

- Hopfield (1982) | Mehrschichtige neuronale Netze beleben das Interesse an neuronalen Netzen erneut.
- Rumelhart, Hinton, and Williams (1986) | Popularisierung des Backpropagation Algorithmus.
- Hauptinteresse in den 1990er und 2000er Jahren im Machine Learning gilt aber SVMs und Bayesian Inference.

$\Rightarrow$ Zweiter Winter Neuronaler Netze

Zweite Renaissance

- 2009 - 2012 | Schmidhuber (2015) gewinnen Klassifikationswettbewerbe mit neuronalen Netzen.
- LeCun, Bengio, and Hinton (2015) | Neuronale Netze unter dem Label “Deep Learning” wieder sehr in Mode.
- 2015 - 2025 | Viele Menschen verwechseln die Begriffe “Künstliche Intelligenz” und “Neuronales Netz”.
- Ostwald and Usée (2021) | Beweis der Validität des Backpropagation Algorithmus in Matrixform.

Neuronale Netze und Prädiktive Modellierung

Explanatorische Modellierung ⇔ Grundlagenforschung

Bestimmung von $\hat{\phi} \coloneqq \operatorname{argmin} \left\| \hat{\phi} - \phi \right\|$

Bestimmung von $\hat{f} \coloneqq \operatorname{argmin}_{f \in F} \| \upsilon - f(\xi) \|, F$ beliebig

Prädiktive Modellierung ⇔ Anwendungsorientierte Forschung

⇒ Neuronale Netze zur Approximation multivariater vektorwertiger Funktionen im prädiktiven Sinn.

Universelle Approximationstheoreme

Topologische Aussagen über die Dichten von Funktionenräumen (vgl. Friedman (1970)).

Neuronale Netze können eine Vielzahl von Funktionen sehr genau approximieren, wenn

- die Anzahl der Neurone gegen Unendlich geht (*arbitrary width case*) bzw.
- die Anzahl der Neuronenschichten gegen Unendlich geht (*arbitrary depth case*).

Arbitrary width case ⇒ Cybenko (1989), Hornik (1991), Leshno et al. (1993), Pinkus (1999)

Arbitrary depth case ⇒ Lu et al. (2017), Hanin and Sellke (2018), Kidger and Lyons (2020)

Universelle Approximationstheoreme sind Existenzaussagen, keine Konstruktionsaussagen.

⇒ Parameter neuronaler Netze müssen durch Gradientenverfahren gelernt werden.

Beispiel eines Universellen Approximationstheorems

Theorem (Universelles Approximationstheorem nach Kidger (2020))

$\mathcal{X}$ sei eine kompakte Teilmenge von $\mathbb{R}^m$, $\sigma : \mathbb{R} \to \mathbb{R}$ sei eine nicht-affine stetige und zumindest in einem Punkt stetig differenzierbare Funktion mit einer von Null verschiedenen Ableitung in diesem Punkt. F sei die Menge der neuronalen Netze f mit Inputdimension m, Outputdimension n_k und einer beliebigen Anzahl verdeckter Schichten mit jeweils $m + n_k + 2$ Neuronen und Aktivierungsfunktion σ, sowie der Identitätsabbildung als Aktivierungsfunktion der Outputschicht. Dann existiert zu jeder stetigen multivariaten vektorwertigen Funktion

$$g : \mathcal{X} \to \mathbb{R}^{n_k}, x \mapsto g(x) \tag{1}$$

ein neuronales Netz $f \in F$, so dass für ein beliebig kleines $\epsilon > 0$ gilt, dass

$$\sup_{x \in \mathcal{X}} \|f(x) - g(x)\| < \epsilon. \tag{2}$$

Bemerkungen

- Das Supremum $\sup$ kann intuitiv als Maximum verstanden werden.
- $\|\cdot\|$ bezeichnet eine *Metrik* (Abstandsfunktion) auf $\mathbb{R}^{n_k}$.
- Für jedes $x \in \mathcal{X}$ wird der Abstand zwischen dem Wert von g und dem Wert von f also beliebig klein.
- Man sagt dazu auch, dass F im Raum der stetigen multivariaten vektorwertigen Funktionen *dicht* ist.
- Man kann das Theorem sicherlich noch präziser formulieren und sollte es beweisen.
- Wir verzichten hier darauf und führen das Theorem nur als “intuitives Beispiel” auf.

Anwendungsszenario

Funktionale Architektur

Lernen

Backpropagation

Anwendungsbeispiel

Selbstkontrollfragen

Definition (Potentialfunktionen)

$W \in \mathbb{R}^{m\times(n+1)}$ sei eine Matrix, die wir *Wichtungsmatrix* nennen und $a \in \mathbb{R}^n$ sei ein Vektor, den wir *Aktivierungsvektor* nennen. Dann nennen wir eine Funktion der Form

$$\Phi : \mathbb{R}^{m\times(n+1)} \times \mathbb{R}^n \to \mathbb{R}^m, (W, a) \mapsto \Phi(W, a) := W \cdot \begin{pmatrix} a \\ 1 \end{pmatrix} \tag{3}$$

eine *bivariate Potentialfunktion*. Für ein festes $a \in \mathbb{R}^n$ nennen wir eine Funktion der Form

$$\Phi_a : \mathbb{R}^{m\times(n+1)} \to \mathbb{R}^m,\ W \mapsto \Phi_a(W) := \Phi(W, a) \tag{4}$$

eine *Wichtungsmatrix-variate Potentialfunktion*. Weiterhin nennen wir für eine feste Matrix $W \in \mathbb{R}^{m\times(n+1)}$ eine Funktion der Form

$$\Phi_W : \mathbb{R}^n \to \mathbb{R}^m, a \mapsto \Phi_W(a) := \Phi(W, a) \tag{5}$$

eine *Potentialfunktion*. Schließlich nennen wir $z := \Phi_W(a)$ einen *Potentialvektor*.

Definition (Aktivierungsfunktion)

Wir nennen eine multivariate vektorwertige Funktion der Form

$$\Sigma : \mathbb{R}^n \to \mathbb{R}^n, z \mapsto \Sigma(z) := (\sigma(z_1), ..., \sigma(z_n))^T, \tag{6}$$

mit

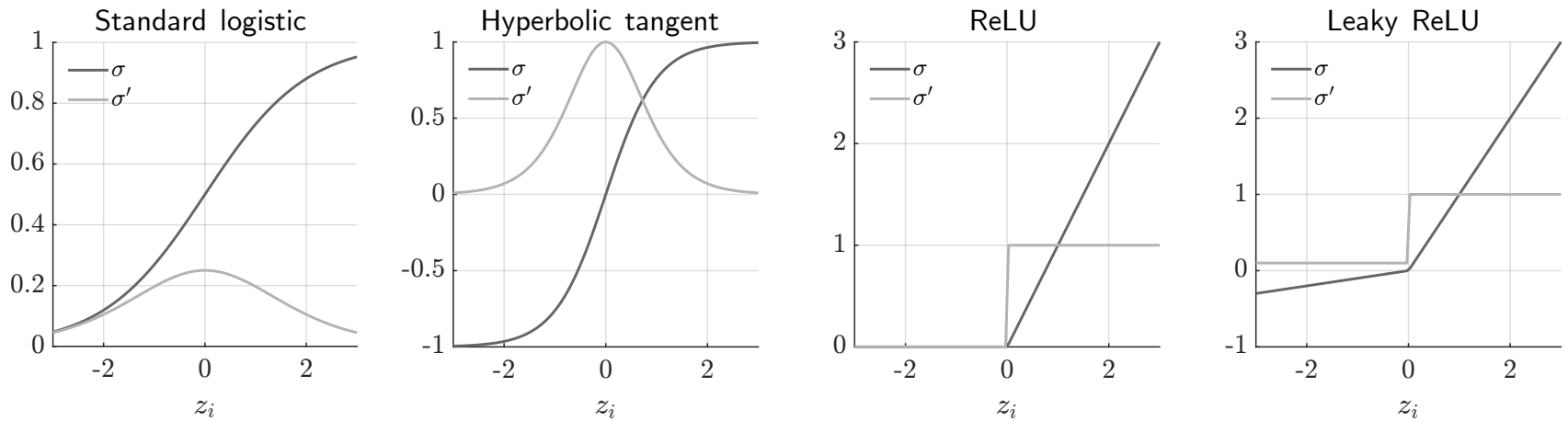
$$\sigma : \mathbb{R} \to \mathbb{R}, z_i \mapsto \sigma(z_i) =: a_i \text{ für alle } i = 1, ..., n, \tag{7}$$

eine *komponentenweise Aktivierungsfunktion* und die univariate reellwertige Funktion σ eine *Aktivierungsfunktion*.

Typische Aktivierungsfunktionen und ihre Ableitungen

Name	Definition	Ableitung
Standard logistic	$\sigma(z_i) := \frac{1}{1+\exp(-z_i)}$	$\sigma'(z_i) = \frac{\exp(z_i)}{(1+\exp(z_i))^2}$
Hyperbolic tangent	$\sigma(z_i) := \tanh(z_i)$	$\sigma'(z_i) = 1 - \tanh^2(z_i)$
ReLU	$\sigma(z_i) := \max(0, z_i)$	$\sigma'(z_i) = \begin{cases} 0, & z_i < 0 \\ \emptyset, & z_i = 0 \\ 1, & z_i > 0 \end{cases}$
Leaky ReLU	$\sigma(z_i) := \begin{cases} 0.1z_i, & z_i \leq 0 \\ z_i, & z_i > 0 \end{cases}$	$\sigma'(z_i) = \begin{cases} 0.01, & z_i \leq 0 \\ 1, & z_i > 0 \end{cases}$

Typische Aktivierungsfunktionen und ihre Ableitungen



Definition (k-schichtiges neuronales Netz)

Eine multivariate vektorwertige Funktion

$$f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_k}, x \mapsto f(x) =: y \tag{8}$$

heißt *k-schichtiges neuronales Netz*, wenn f von der Form

$$f : \mathbb{R}^{n_0} \xrightarrow{\Phi^1_{W^1}} \mathbb{R}^{n_1} \xrightarrow{\Sigma^1} \mathbb{R}^{n_1} \xrightarrow{\Phi^2_{W^2}} \mathbb{R}^{n_2} \xrightarrow{\Sigma^2} \mathbb{R}^{n_2} \xrightarrow{\Phi^3_{W^3}} \\ \dots \xrightarrow{\Phi^{k-1}_{W^{k-1}}} \mathbb{R}^{n_{k-1}} \xrightarrow{\Sigma^{k-1}} \mathbb{R}^{n_{k-1}} \xrightarrow{\Phi^{k}_{W^{k}}} \mathbb{R}^{n_k} \xrightarrow{\Sigma^{k}} \mathbb{R}^{n_k}, \tag{9}$$

ist, wobei für $l = 1, \dots, k$

$$\Phi^l_{W^l} : \mathbb{R}^{n_{l-1}} \to \mathbb{R}^{n_l}, a^{l-1} \mapsto \Phi^l_{W^l}(a^{l-1}) := W^l \cdot \begin{pmatrix} a^{l-1} \\ 1 \end{pmatrix} =: z^l \tag{10}$$

Potentialfunktionen und

$$\Sigma^l : \mathbb{R}^{n_l} \to \mathbb{R}^{n_l}, z^l \to \Sigma^l(z^l) =: a^l \tag{11}$$

komponentenweise Aktvierungsfunktionen sind. Für ein $x \in \mathbb{R}^{n_0}$ nimmt ein k-schichtiges neuronales Netz den Wert

$$f(x) := \Sigma^k(\Phi^k_{W^k}(\Sigma^{k-1}(\Phi^{k-1}_{W^{k-1}}(\Sigma^{k-2}(\cdots(\Sigma^1(\Phi^1_{W^1}(x))\cdots)))))) \in \mathbb{R}^{n_k}. \tag{12}$$

an.

Funktionale Architektur

Bemerkungen

- Die Vektoren $a^l = (a_1^l, ..., a_{n_l}^l)^T \in \mathbb{R}^{n_l}, l = 0, 1, ..., k$ heißen *Aktivierungsvektoren der $l$ten Schicht*.
- Die Komponenten $a_i^l \in \mathbb{R}, i = 1, ..., n_l, l = 0, 1, ..., k$ heißen *Aktivierungen der $l$ten Schicht*.
- Die Schicht mit Index $l = 0$ und Dimension n_0 heißt *Inputschicht*.
- Der Aktivierungsvektor mit Index $l = 0$ heißt *Input* und wird mit $x := a^0$ bezeichnet.
- Die Schicht mit Index $l = k$ und Dimension n_k heißt *Outputschicht*
- Der Aktivierungsvektor mit Index $l = k$ heißt Output und wird mit $y := a^k$ bezeichnet.
- Die Schichten mit den Indizes $l = 1, ..., k-1$ heißen *verdeckte Schichten (hidden layers)*.
- $w_{ij}^l \in \mathbb{R}$ sei der ijte Eintrag der $l$ten Wichtungsmatrix, d.h.

$$W^l = (w_{ij}^l)_{1 \leq i \leq n_l, 1 \leq j \leq n_{l-1}+1} \in \mathbb{R}^{n_l \times (n_{l-1}+1)} \text{ für } l = 1, ..., k. \tag{13}$$

- w_{ij}^l heißt *(synaptisches) Gewicht* der Verbindung von Neuron j in Schicht $l-1$ und Neuron i in Schicht l.
- Für $i = 1, ..., n_l$ heißt $w_{i,n_{l-1}+1}$ *Bias* von Neuron i in Schicht l.
- Die letzte Spalte von W^l enkodiert also die Biases für die Neuronen in Schicht l.

Bemerkungen

Auf der Ebene einzelner Neurone ergibt sich damit folgende Nomenklatur:

- Das *Potential* von Neuron i in Schicht l für $i = 1, \ldots, n_l$ und $l = 1, \ldots, k$ ist gegeben durch

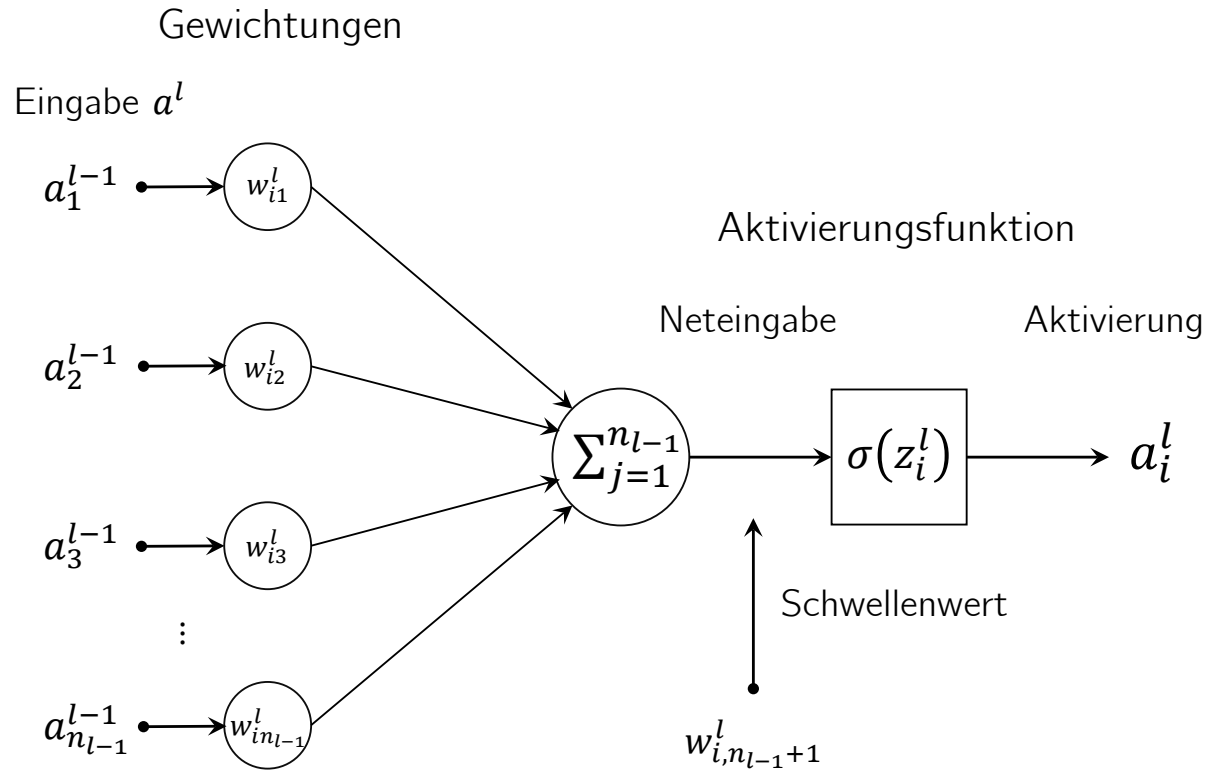
$$z_i^l = \sum_{j=1}^{n_{l-1}} w_{ij}^l a_j^{l-1} + w_{i,n_{l-1}+1} \in \mathbb{R}. \tag{14}$$

- Die *Aktivierung* von Neuron i in Schicht l für $i = 1, \ldots, n_l$ und $l = 1, \ldots, k$ ist gegeben durch

$$a_i^l = \sigma \left(\sum_{j=1}^{n_{l-1}} w_{ij}^l a_j^{l-1} + w_{i,n_{l-1}+1} \right) \in \mathbb{R}, \tag{15}$$

- Die Aktivierung a_i^l kann als die mittlere Feuerungsrate des $i$ten Neuron in der $l$ten Schicht verstanden werden.

Funktionale Architektur



Beispiel

$k = 3, n_0 = 2, n_1 = 3,\ n_2 = 3, n_3 = 2$, Standard logistic function σ

$$\begin{pmatrix} a^0 \\ 1 \end{pmatrix} = \begin{pmatrix} a_1^0 \\ a_2^0 \\ 1 \end{pmatrix} \qquad \begin{pmatrix} a^1 \\ 1 \end{pmatrix} = \begin{pmatrix} a_1^1 \\ a_2^1 \\ a_3^1 \\ 1 \end{pmatrix} \qquad \begin{pmatrix} a^2 \\ 1 \end{pmatrix} = \begin{pmatrix} a_1^2 \\ a_2^2 \\ a_3^2 \\ 1 \end{pmatrix} \qquad a^3 = \begin{pmatrix} a_1^3 \\ a_2^3 \end{pmatrix}$$

$$W^1 = \begin{pmatrix} w_{11}^1 & w_{12}^1 & w_{13}^1 \\ w_{21}^1 & w_{22}^1 & w_{23}^1 \\ w_{31}^1 & w_{32}^1 & w_{33}^1 \end{pmatrix} \quad W^2 = \begin{pmatrix} w_{11}^2 & w_{12}^2 & w_{13}^2 & w_{14}^2 \\ w_{21}^2 & w_{22}^2 & w_{23}^2 & w_{24}^2 \\ w_{31}^2 & w_{32}^2 & w_{33}^2 & w_{34}^2 \end{pmatrix} \quad W^3 = \begin{pmatrix} w_{11}^3 & w_{12}^3 & w_{13}^3 & w_{14}^3 \\ w_{21}^3 & w_{22}^3 & w_{23}^3 & w_{24}^3 \end{pmatrix}$$

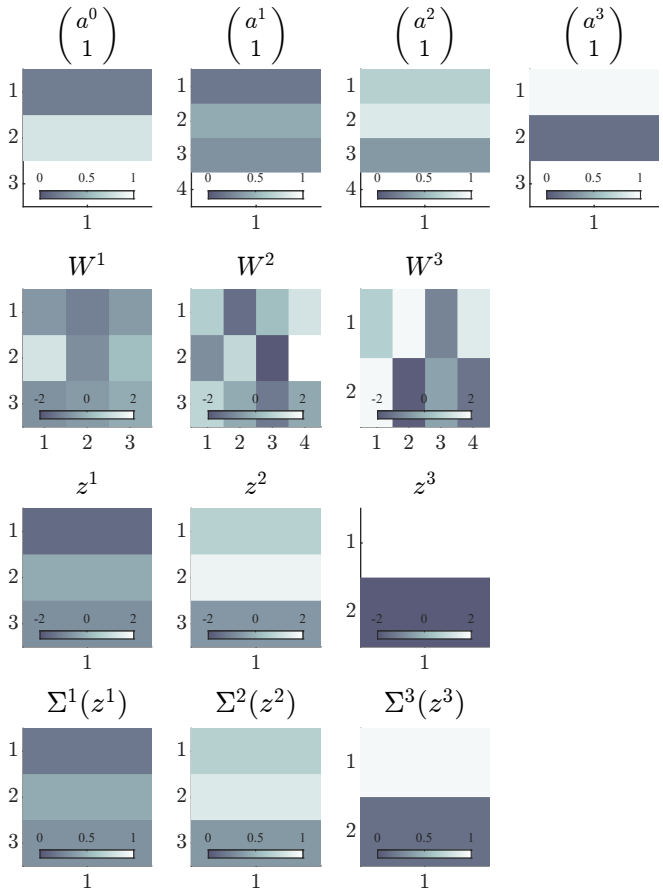
$$z^1 = \begin{pmatrix} z_1^1 \\ z_2^1 \\ z_3^1 \end{pmatrix} \qquad z^2 = \begin{pmatrix} z_1^2 \\ z_2^2 \\ z_3^2 \end{pmatrix} \qquad z^3 = \begin{pmatrix} z_1^3 \\ z_2^3 \end{pmatrix}$$

$$\Sigma^1(z^1) = \begin{pmatrix} \sigma(z_1^1) \\ \sigma(z_2^1) \\ \sigma(z_3^1) \end{pmatrix} \qquad \Sigma^2(z^2) = \begin{pmatrix} \sigma(z_1^2) \\ \sigma(z_2^2) \\ \sigma(z_3^2) \end{pmatrix} \qquad \Sigma^3(z^3) = \begin{pmatrix} \sigma(z_1^3) \\ \sigma(z_2^3) \end{pmatrix}$$

Es gilt $x = a^0$ und $a^3 = y$.

Beispiel

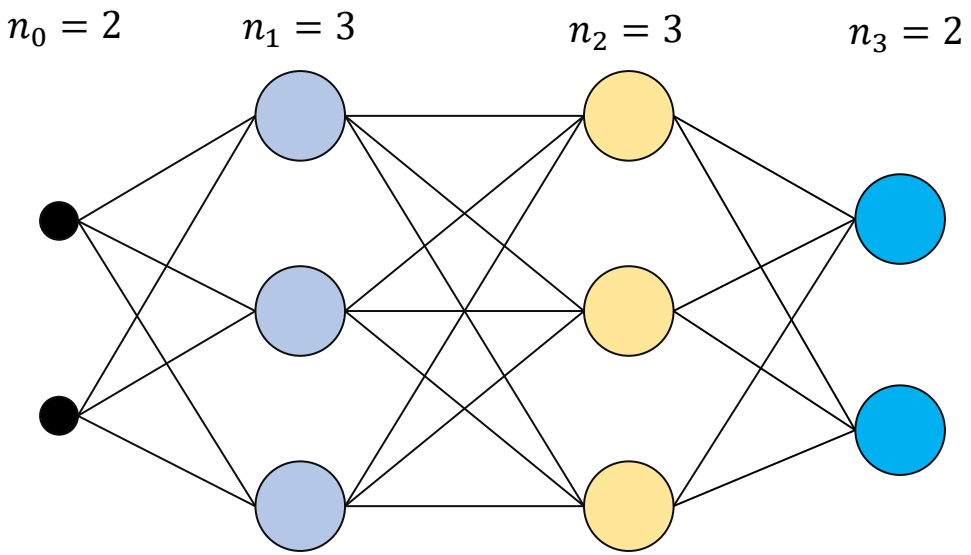
$k = 3, n_0 = 2, n_1 = 3,\ n_2 = 3, n_3 = 2$, Standard logistic function σ



Es gilt $x = a^0$ und $a^3 = y$.

Beispiel

$k = 3, n_0 = 2, n_1 = 3,\ n_2 = 3, n_3 = 2$



- Die Biases sind hier nicht visualiert.

Anwendungsszenario

Funktionale Architektur

Lernen

Backpropagation

Anwendungsbeispiel

Selbstkontrollfragen

Überblick

Anhand eines Trainingsdatensatzes werden die Parameter eines neuronalen Netzes wie folgt gelernt:

- Zunächst wird eine Funktion definiert, die misst, inwiefern sich bei einem gegebenen Inputvektor und Zielvektor der Output des neuronalen Netzes basierend auf einem Wert der Parameter unterscheidet. Diese Funktion nennt man eine *Kostenfunktion* oder *Zielfunktion*.
- Der summierte Wert der Kostenfunktion über alle Trainingsdatenpunkte wird dann durch Veränderung der Parameter minimiert, so dass Parameterwerte gefunden werden, für die die Abweichung zwischen Zielvektor und Output des neuronalen Netzes bei gegebenemen Inputvektor möglichst gering ist.
- Zur Minimierung der Kostenfunktion wird üblicherweise ein Gradientenverfahren benutzt.
- Zur Berechnung der in diesem Verfahren auftretenden Zielfunktionsgradienten wird ein komputational effizienter Algorithmus eingesetzt, der die spezielle Struktur neuronaler Netze ausnutzt und unter dem Namen *Backpropagation Algorithmus* bekannt ist.

In der Folge wollen wir die Aspekte dieses Lernprozesses genauer betrachten.

Definition (Trainingsdatensatz)

Ein *Trainingsdatensatz* für ein neuronales Netz ist eine Menge

$$\mathcal{D} := \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}, \tag{16}$$

wobei $x^{(i)} \in \mathbb{R}^{n_0}$ *Featurevektor* und $y^{(i)} \in \mathbb{R}^{n_k}$ *Zielvektor* genannt werden.

Bemerkungen

- Im Kontext der zuvor betrachteten multivariaten Verfahren gilt hier $n_0 = m$.
- Typische Zielvektorformate beim Training neuronaler Netze sind

  $y^{(i)} \in \{0, 1\}$ für binäre Klassifikationsprobleme,

  $y^{(i)} \in \{0, 1\}^{n_k}$ mit $\sum_{i=1}^{n_k} y_i = 1, n_k > 1$ für n_k-fache Klassifikationsprobleme,

  $y^{(i)} \in \mathbb{R}^{n_k}, n_k > 1$ für Regressionsprobleme.

Definition (Trainieren eine neuronalen Netzes)

f sei ein k-schichtiges neuronales Netz und $\mathcal{D}$ sei ein Trainingsdatensatz. Dann bezeichnet der Begriff des *Trainierens* den Prozess der Adaptation der Wichtungsmatrizen $W^1, ..., W^k$ des neuronalen Netzes mit dem Ziel, ein Abweichungskriterium zwischen der Ouptutaktivierung $f(x^{(i)})$ und dem assoziierten Wert des Zielvektors $y^{(i)}$ über alle Trainingsdatenpunkte $(x^{(i)}, y^{(i)}), i = 1, ..., n$ des Trainingsdatensatzes $\mathcal{D}$ hinweg zu minimieren.

Bemerkungen

- Wir erinnern an das Ziel $f := \text{argmin}_{\tilde{f} \in F} \|\upsilon - \tilde{f}(\xi)\|$ der prädiktiven Modellierung.
- Das erwähnte Abweichungskriterium wird in Form von *Kostenfunktionen* definiert.
- Wir benötigen noch den Begriff der *Wichtungsmatrix-variaten neuronalen Netzfunktion*.

Definition (Wichtungsmatrix-variate neuronale Netzfunktion)

f sei ein k-schichtiges neuronales Netz und x sei ein Input von f. Dann ist *Wichtungsmatrix-variate neuronale Netzfunktion f_x von f* definiert als die Funktion

$$f_x : \mathbb{R}^{n_1 \times (n_0+1)} \times \cdots \times \mathbb{R}^{n_k \times (n_{k-1}+1)} \to \mathbb{R}^{n_k}, (W^1, \ldots, W^k) \mapsto f_x(W^1, \ldots, W^k)$$
$$:= \Sigma^k(\Phi^k(W^k, \Sigma^{k-1}(\Phi^{k-1}(W^{k-1}, \cdots (W^2, \Sigma^1(\Phi^1(W^1, x))) \cdots)))), \quad (17)$$

wobei für $l = 1, \ldots, k$, Φ^l die bivariate Potentialfunktion bezeichnet, die der Potentialfunktion $\Phi^l_{W^l}$ in der Definition des neuronalen Netzes entspricht. Weiterhin definieren wir für $l = 1, \ldots, k$, die *Wichtungsmatrix-variate neuronale Netzfunktion der $l$ten Schicht* f_x^l für festes fixed $W^\ell \in \mathbb{R}^{n_\ell \times (n_{\ell-1}+1)}$m mit $\ell = 1, \ldots, k$ und $\ell \neq l$ als

$$f_x^l : \mathbb{R}^{n_1 \times (n_{l-1}+1)} \to \mathbb{R}^{n_k}, W^l \mapsto f_x^l(W^l) := f_x^l(W^1, \ldots, W^k). \quad (18)$$

Bemerkungen

- Die Definition von f in der Definition eines k-schichtiges neuronales Netzes ist eine Funktion des Inputs x bei festen Wichtungsmatrizen $W^1, \ldots, W^l$. Zum Trainieren eines neuronalen Netzes ist es aber entscheidend, bei festem Input den Output des neuronalen Netzes bei Variation der Parameter $W^1, \ldots, W^l$ zu monitoren. Dies motiviert den Begriff der Wichtungsmatrix-variaten neuronalen Netzfunktion: Die Definition von f_x in (17) ist eine Funktion der Wichtungsmatrizen $W^1, \ldots, W^l$ bei festem Input x.

Definition (Output-spezifische Kostenfunktionen)

f sei ein k-schichtiges neuronales Netz und y sei ein Zielvektor von f. Dann wird eine multivariate reelwertige Funktion der Form

$$c_y : \mathbb{R}^{n_k} \to \mathbb{R}, a^k \mapsto c_y(a^k) \tag{19}$$

Output-spezifische Kostenfunktion genannt.

Bemerkung

- Eine Output-spezifische Kostenfunktion c_y misst die Abweichung des Outputs a^k eines neuronalen Netzes von einem Zielvektor y. Untenstehende Tabelle führt zwei typische Beispiele für Output-spezifische Kostenfunktionen und ihre Gradienten, die in der Folge wichtig werden, auf.

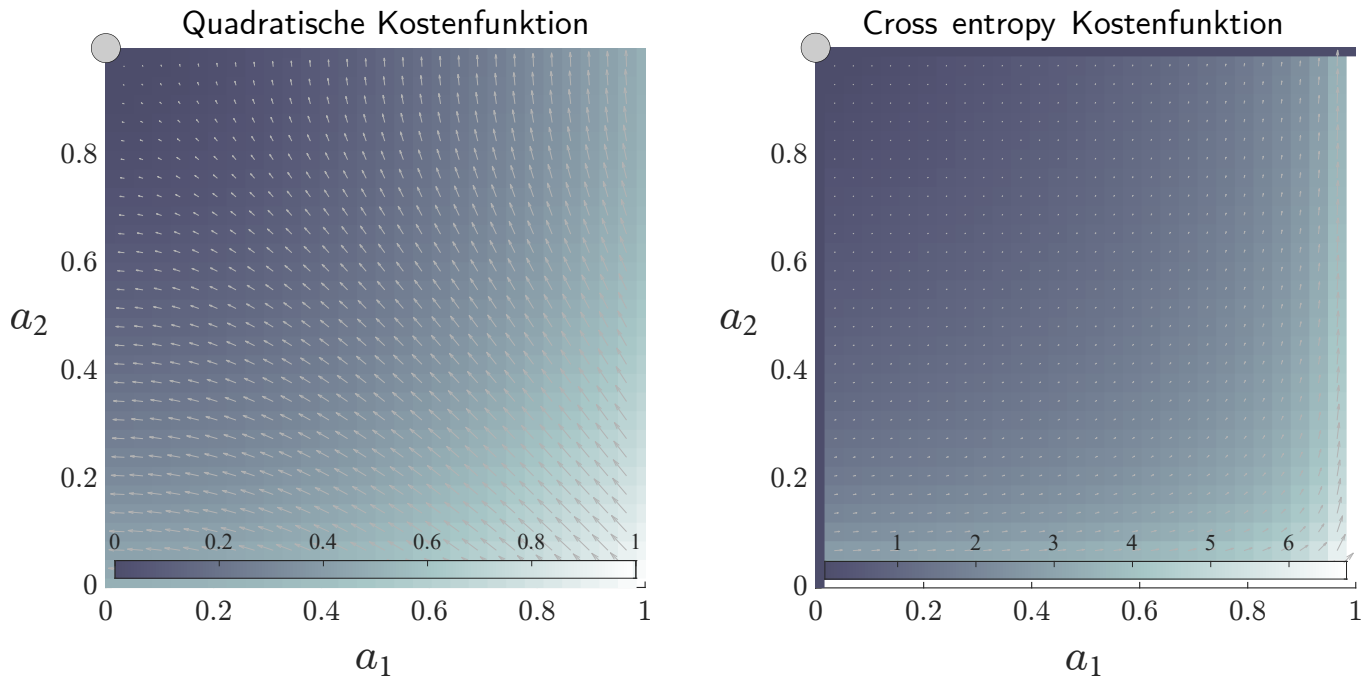
Quadratische Kostenfunktion

Definition	Gradient
$c_y(a^k) := \frac{1}{2}\sum_{j=1}^{n_k}(a_j^k - y_j)^2$	$\nabla c_y(a^k) := \left(a_j^k - y_j\right)_{j=1,\ldots,n_k}$

Cross-entropy Kostenfunktion

Definition	Gradient
$c_y(a^k) := -\sum_{i=1}^{n_k} y_j \ln a_j^k + (1-y_j)\ln(1-a_j^k)$	$\nabla c_y(a^k) := \left(-\frac{y_j}{a_j^k} + \frac{1-y_j}{1-a_j^k}\right)_{j=1,\ldots,n_k}$

Output-spezifische Kostenfunktionswerte für $y = (0, 1)^T$ bei logistischer Aktivierungsfunktion



- Beide Funktionen haben ihr Minimum bei $a = y$.
- Die Pfeile bilden die skalierten Gradientenwerte der jeweiligen Funktion ab.

Definition (Trainingsdatenpunkt-spezifische Kostenfunktionen)

f sei ein k-schichtiges neuronales Netz, f_x sei die zugehörige wichtungsmatrix-variate neuronale Netzfunktion, x und y seien Inputs und Outputs des neuronalen Netzes, $\mathcal{D}$ sei ein Trainingsdatensatz und c_y sei eine Output-spezifische Kostenfunktion. Dann heißt eine multimatrixvariate reellwertige Funktion der Form

$$c_{xy} : \mathbb{R}^{n_1 \times (n_0+1)} \times ... \times \mathbb{R}^{n_k \times (n_{k-1}+1)} \to \mathbb{R},$$
$$(W^1, ..., W^k) \mapsto c_{xy}(W^1, ..., W^k) := c_y(f_x(W^1, ..., W^k)) \quad (20)$$

Trainingsdatenpunkt-spezifische Kostenfunktion.

Bemerkung

- Eine Trainingsdatenpunkt-spezifische Kostenfunktion c_{xy} misst die Abweichung des Outputs a^k eines neuronalen Netzes von einem Zielvektor y mithilfe einer Output-spezifischen Kostenfunktion c_y für einen festen Input x als Funktion der (also bei variablen) Wichtungsmatrizen.

Definition (Gewichtsvektor)

f sei ein k-schichtige neuronales Netz mit $n_l \times n_{l-1} + 1$-dimensionalen Gewichtsmatrizen $W^l, l = 1, .., k$ und es sei

$$p := \sum_{l=1}^{n_k} n_l(n_{l-1} + 1). \tag{21}$$

die Anzahl der Gewichtsparameter des neuronalen Netzes. Dann heißt

$$\mathcal{W} := \left(\text{vec}\left(W^l\right)\right)_{1 \leq l \leq k} \in \mathbb{R}^p \tag{22}$$

der *Gewichtsvektor* des neuronalen Netzes.

Bemerkung

- Die Vektorisierung und Konkatenation der Wichtungsmatrizen im Sinne des Gewichtsvektors erlaubt es, dass Trainieren eines neuronalen Netzes als ein Standardoptimierungsproblem einer multivariaten (nicht multimatrixvariaten) reellwertigen zu formulieren.

Definition (Additive Kostenfunktion)

$\mathcal{D}$ sei ein Trainingsdatensatz und c_{xy} sei eine Trainingsdatenpunkt-spezifische Kostenfunktion. Dann nennt man eine multivariate reellwertige Funktion der Form

$$c_{\mathcal{D}} : \mathbb{R}^p \to \mathbb{R}, \mathcal{W} \mapsto c_{\mathcal{D}}(\mathcal{W}) := \frac{1}{n} \sum_{i=1}^{n} c_{x^{(i)}y^{(i)}}(W^1, ..., W^k) \tag{23}$$

eine *additive Kostenfunktion*.

Bemerkung

- Die additive Kostenfunktion ist die zentrale Zielfunktion beim Trainieren eines neuronalen Netzes.
- $c_{\mathcal{D}}$ ist eine multivariate reellwertige Funktion, es liegt also ein Standardoptimierungsproblem vor.
- Wir nehmen dabei stillschweigend an, dass die sinnvolle Aufteilung des Gewichtsvektors $\mathcal{W}$ auf die Wichtungsmatrizen $W^1, ..., W^k$ in der Auswertung der Funktion $c_{\mathcal{D}}$ geschieht.

Definition (Batch Gradientenverfahren für neuronale Netze)

f sei ein k-schichtiges neuronales Netz mit Gewichtsvektor $\mathcal{W}$, $\mathcal{D}$ sei ein Trainingsdatensatz bestehend aus n Trainingsdatenpunkten, und $c_{\mathcal{D}}$ sei eine additive Kostenfunktion mit assoziierter Trainingsexemplar-spezifischer Kostenfunktion c_{xy}. Dann ist ein Gradientenverfahren zur Minimierung der additiven Kostenfunktion $c_{\mathcal{D}}$ (und damit zum Lernen der Parameter von f) definiert durch

Initialisierung

Wahl eines Startpunktes $\mathcal{W}^{(0)}$ und einer Lernrate $\alpha > 0$.

Iterationen

Für $j = 1, 2, \ldots$ setze

$$\mathcal{W}^{(j)} := \mathcal{W}^{(j-1)} - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla c_{x^{(i)}y^{(i)}}(W^1, \ldots, W^k), \tag{24}$$

wobei

$$\nabla c_{x^{(i)}y^{(i)}}(W^1, \ldots, W^k) = \left(\nabla_{W^l} c_{x^{(i)}y^{(i)}}(W^1, \ldots, W^k) \right)_{1 \leq l \leq k} \tag{25}$$

für $i = 1, \ldots, n$ den Gradienten der $i$ten Trainingsexemplar-spezifischen Kostenfunktion bezeichnet

Bemerkungen

- $\mathcal{W}^{(j)}$ wird in (22) in die negative Richtung des Gradientenmittelwerts über Trainingsdatenpunkte adaptiert. Wird der Gradientenmittelwert dagegen nur über eine zufällig gewählte Teilmenge der Trainingsdatenpunkte berechnet, so spricht man von einem *stochastischen Gradientenverfahren*.

Anwendungsszenario

Funktionale Architektur

Training

Backpropagation

Anwendungsbeispiel

Selbstkontrollfragen

Wesen und Motivation des Backpropagation Algorithmus

Der Backpropagation (BP) Algorithmus dient der numerischen Bestimmung der Komponenten

$$\frac{\partial}{\partial w_{ij}^l} c_{xy}(W^1, ..., W^k) \text{ für alle } i = 1, ..., n_l, j = 1, ..., n_{l-1} + 1, \text{ und } l = 1, ..., k. \tag{26}$$

des Gradienten $\nabla c_{x^{(i)}y^{(i)}}(W^1, ..., W^k)$ der $i$ten Trainingsexemplar-spezifischen Kostenfunktion.

Prinzipiell können diese partiellen Ableitungen numerisch durch

$$\frac{\partial}{\partial w_{ij}^l} c_{xy}(W^1, ..., W^k) \approx \frac{c_{xy}(W^1, ..., \tilde{W}^l, ..., W^k) - c_{xy}(W^1, ..., W^l, ..., W^k)}{\epsilon}, \tag{27}$$

mit

- $\tilde{W}^l := W^l + 1_{ij}^l \epsilon$,
- einer Matrix $1_{ij}^l \in \mathbb{R}^{n_l \times (n_{l-1}+1)}$ aus 0en mit einer 1 an der w_{ij}^l Stelle in W^l und
- einem Schrittweitenparameter $\epsilon > 0$

approximiert werden (vgl. Definition der partiellen Ableitung).

Wesen und Motivation des Backpropagation Algorithmus

Dieses Vorgehen würde für jede Iteration des Gradientenverfahren und für jeden Trainingsdatenpunkt

$$K := 1 + \sum_{l=1}^{k} n_l(n_{l-1} + 1) \tag{28}$$

Auswertungen der Trainingsdatenpunkt-spezifischen Kostenfunktion c_{xy} und somit von f erfordern. Man nennt die Auswertung von f für einen Trainingsdatenpunkt x einen *Forward Pass*.

Die zentrale Eigenschaft des Backpropagation Algorithmus ist es, für die Auswertung von ∇c_{xy} die Anzahl der notwendigen *Forward Passes* pro Gradientenverfahrensiteration von K auf 1 zu reduzieren.

Um dies zu erreichen, nutzt der Backpropagation Algorithmus einen sogenannten *Backward Pass*, der die gleiche komputationale Komplexität wie der *Forward Pass* hat und auf einer multivariate Version der Kettenregel der Differentialrechnung sowie der repetitiven funktionalen Architektur neuronaler Netze beruht.

Der Backpropagation Algorithmus reduziert die Anzahl nötiger *Passes* zur Auswertung von ∇c_{xy} also von K *Forward Passes* auf 1 *Forward Pass* und 1 *Backward Pass*.

Theorem (Backpropagation Algorithmus)

f sei ein k-schichtiges neuronales Netz, $W_\bullet^l \in \mathbb{R}^{n_l \times n_{l-1}}$ seien für $l = 1, \ldots, k$ Matrizen, die durch das Entfernen der letzten Spalte der Wichtungsmatrizen $W^l \in \mathbb{R}^{n_l \times n_{l-1}+1}$ entstehen, c_{xy} sei eine Trainingsdatenpunkt-spezifische Kostenfunktion, $\nabla c_y(a^k)$ sei der Gradient der Output-spezifischen Kostenfunktion, $\tilde{\Sigma}^l(z^l) := (\sigma'(z_1^l), \ldots, \sigma'(z_{n_l}^l))^T$ sei der Vektor der Aktivierungsfunktionenableitungen ausgewertet an der Stelle z^l und $\Sigma^l(z^l)$ die komponentenweise Aktivierungsfunktion evaluiert an der Stelle z^l. Dann können die partiellen Gradienten von c_{xy} hinsichtlich der Wichtungsmatrizen W^l für $l = k, k-1, \ldots, 1$ mit folgendem Algorithmus berechnet werden:

Initialisierung

Setze $W^{k+1} := (1 \quad 0)$ und $\delta^{k+1} := \nabla c_y(a^k)$.

Iterationen

Für $l = k, k-1, k-2, \ldots, 1$, setze

$$\delta^l := \left(\left(W_\bullet^{l+1} \right)^T \cdot \delta^{l+1} \right) \circ \tilde{\Sigma}^l(z^l) \tag{29}$$

und

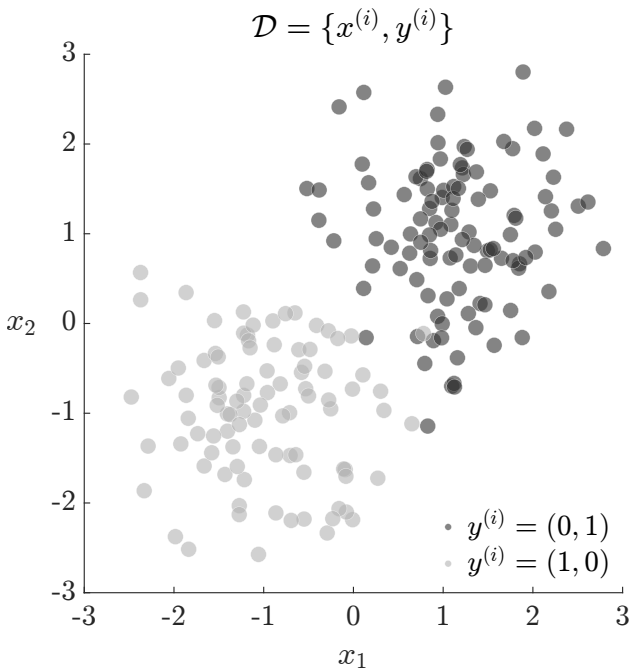
$$\nabla_{W^l} c_{xy}(W^1, \ldots, W^k) := \text{vec}\left(\delta^l \cdot \left(\Sigma^{l-1}(z^{l-1})^T \quad 1 \right) \right), \tag{30}$$

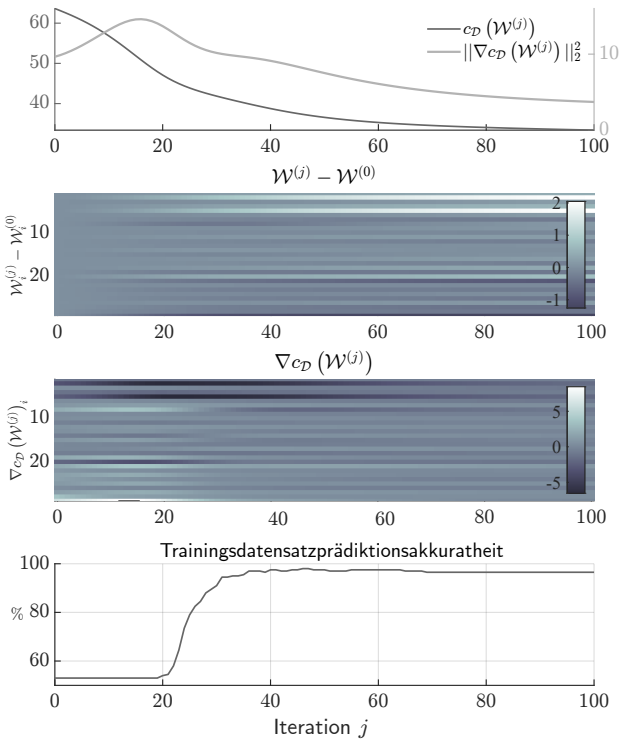
mit Rekursionstermination durch $\Sigma^0(z^0) := x^T$ und dem Hadamard-Produkt $\circ$.

Für weitere Details und einen Beweis verweisen wir auf Ostwald and Usée (2021).

Simulation und Analyse mit Matlab Implementation (Ostwald and Usée (2021))

- Neuronales Netz mit $k = 3, n_0 = 2, n_1 = 3,\ n_2 = 3, n_3 = 2$.
- Trainingsdatensatz anhand eines LDA Modells simuliert.





Anwendungsszenario

Funktionale Architektur

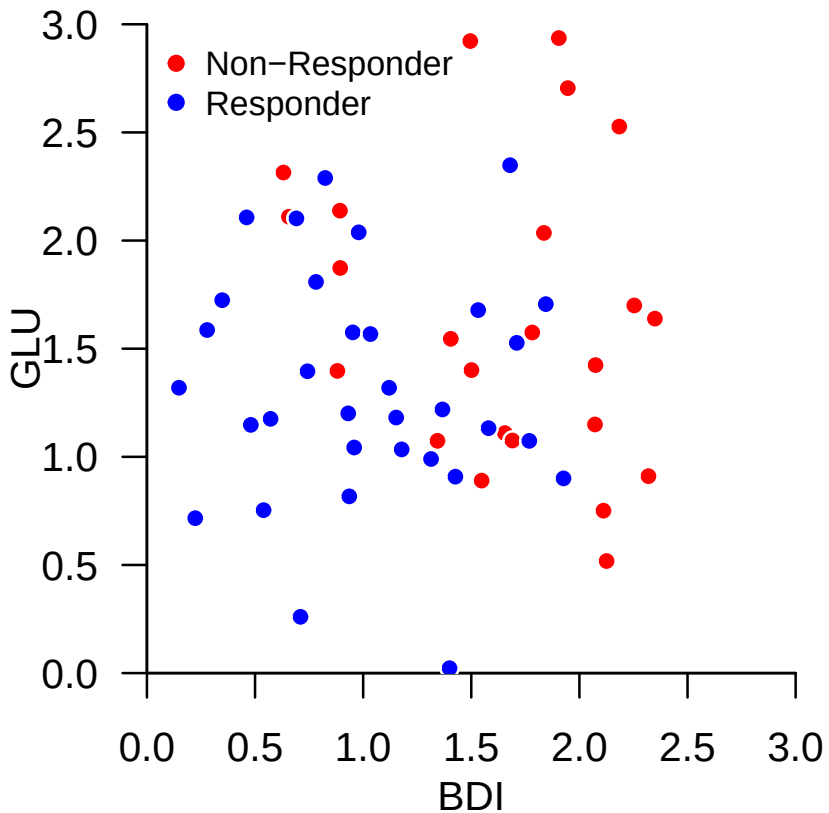
Training

Backpropagation

Anwendungsbeispiel

Selbstkontrollfragen

BDI und GLU Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg



Anwendungsbeispiel

`BDI` und `GLU` Werte bei Depressionsdiagnose als Prädiktoren von CBT Erfolg `RES`

BDI	GLU	RES
0.74	1.40	1
0.22	0.72	1
0.82	2.29	1
2.07	1.15	0
1.71	1.53	1
1.77	1.07	1
1.95	2.70	0
2.18	2.53	0
0.93	1.20	1
1.34	1.07	0
2.35	1.64	0
1.43	0.91	1
1.66	1.11	0
0.28	1.59	1
2.13	0.52	0
1.37	1.22	1
0.89	2.14	0
0.88	1.40	0
0.98	2.04	1
1.93	0.90	1

Prädiktive Modellierung mit `neuralnet()` (Günther and Fritsch (2010))

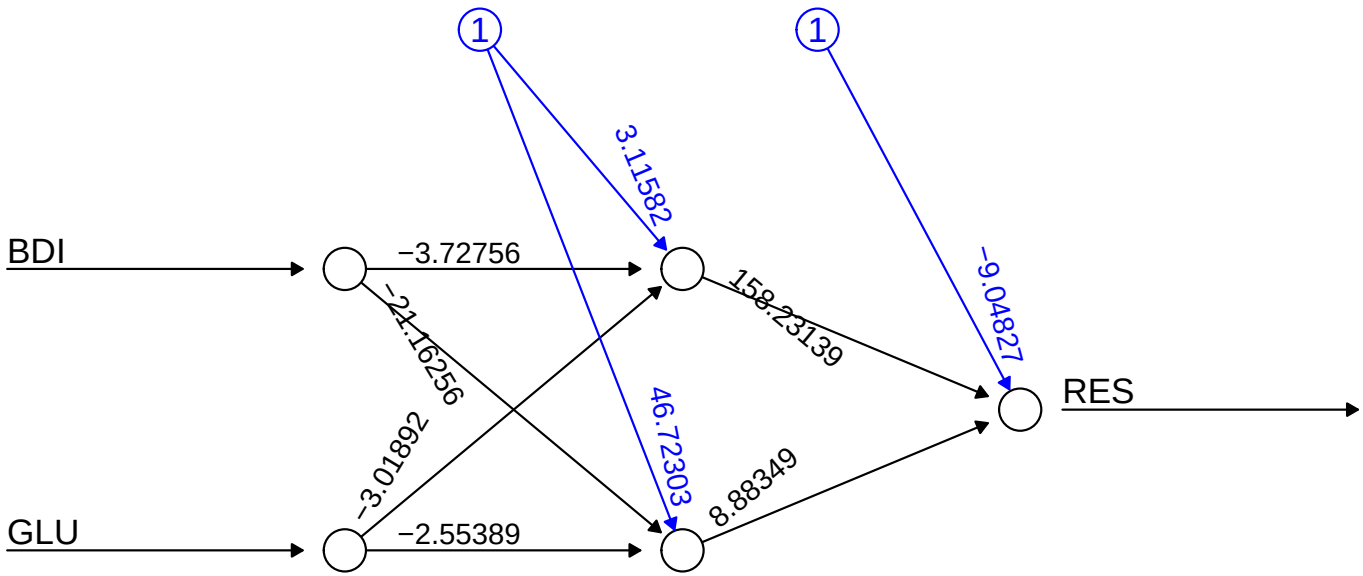
- Eine verdeckte Schicht mit zwei Neuronen

```
library(neuralnet)                                                  # R Paket
set.seed(1)                                                         # random number seed
D    = read.csv("./13_Daten/13_Neuronale_Netze.csv")                # Datensatz
nn   = neuralnet(RES           ~ BDI + GLU,                         # y^{(i)}, x^{(i)} Definitionen
                 data          = D,                                 # Datensatz
                 hidden        = 2,                                 # 2 Neurone in 1 verdeckte Schicht
                 err.fct       = "ce",                              # Cross-Entropie Kostenfunktion
                 linear.output = FALSE)                             # Sigmoide Aktivierungsfunktion
R    = data.frame(BDI           = nn$covariate[,1],                  # BDI
                  GLU           = nn$covariate[,2],                  # GLU
                  RES           = nn$response,                       # RES
                  PRE           = as.numeric(nn$net.result[[1]] > 0.5))  # Trainingsdatenklassifikation
print(sprintf("Prädiktionsakkuratheit = %0.2f", mean(R$PRE == R$RES)))  # Trainingsdatenklassifikationsaccuracy
```

```
[1] "Prädiktionsakkuratheit = 0.72"
```

Prädiktive Modellierung mit `neuralnet()` (Günther and Fritsch (2010))

- Eine verdeckte Schicht mit zwei Neuronen



Error: 23.421166 Steps: 5771

Prädiktive Modellierung mit `neuralnet()` (Günther and Fritsch (2010))

- Eine verdeckte Schicht mit zwei Neuronen

```
# Leave-one-out cross-validation
set.seed(1)                                                                  # random number generator seed
D       = read.csv("./13_Daten/13_Neuronale_Netze.csv")                      # Datensatz
K       = nrow(D)                                                            # Anzahl Cross Folds
y_pred  = matrix(rep(NaN, K*2), nrow = K)                                    # Prädiktionsperformancearray
for(k in 1:K){                                                               # K-fold LOOCV
    D_train      = D[-k,]                                                    # Trainingsdatensatz
    D_test       = D[ k,]                                                    # Testdatensatz
    y_pred[k,1]  = t(D[k,3])                                                 # Testdatensatzfeaturevektorlabel
    nn           = neuralnet(RES~BDI+GLU,                                    # y^{(i)}, x^{(i)} Definitionen
                             D_train,                                        # Trainingsdatensatz
                             hidden = 2,                                     # 1 verdeckte Schicht, 2 Neurone
                             err.fct = "ce",                                 # Cross-Entropie Kostenfunktion
                             linear.output = FALSE)                          # Sigmoide Aktivierungsfunktion
    pred         = predict(nn, D_test)                                       # Testdatenpunktprädiktion
    y_pred[k,2]  = as.numeric(pred[, 1] > 0.5)}                              # Klassifikationsregel
rp      = sum(y_pred[y_pred[,1] == 1,2] == 1)                                # |(1,1)|
rn      = sum(y_pred[y_pred[,1] == 0,2] == 0)                                # |(0,0)|
fp      = sum(y_pred[y_pred[,1] == 0,2] == 1)                                # |(0,1)|
fn      = sum(y_pred[y_pred[,1] == 1,2] == 0)                                # |(1,0)|
ACC     = (rp+rn)/(rp+fp+rn+fn)                                              # Accuracy
SEN     = rp/(rp+fn)                                                         # Sensitivity
SPE     = rn/(rn+fp)                                                         # Specificity
cat("Accuracy   : "  , ACC, ", Sensitivity: " , SEN, ", Specificity: " , SPE)   # Ergebnisausgabe
```

```
Accuracy   :  0.65 , Sensitivity:  0.7647059 , Specificity:  0.5
```

Anwendungsszenario

Funktionale Architektur

Lernen

Backpropagation

Anwendungsbeispiel

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie die zentralen Ideen Universeller Approximationstheoreme im Kontext neuronaler Netze.
2. Geben Sie die Formel für das Potential z_i^l eines Neurons i in einer Schicht l eines neuronalen Netzes wieder und erläutern Sie die verschiedenen Komponenten dieser Formel und ihre intuitive Bedeutung.
3. Geben Sie die Formel für die Aktivierung a_i^l eines Neurons i in einer Schicht l eines neuronalen Netzes wieder und erläutern Sie ihre Bestandteile und deren intuitive Bedeutung.
4. Erläutern Sie das prinzipielle Vorgehen zum Trainieren eines neuronalen Netzes.
5. Geben Sie die Definition der Quadratischen Kostenfunktion wieder und erläutern Sie ihre Bestandteile.
6. Geben Sie das Batch Gradientenverfahren zum Trainieren neuronaler Netze wieder.
7. Differenzieren Sie die Begriffe Batch und Stochastischen Gradientenverfahren zum Trainieren neuronaler Netze.
8. Erläutern Sie Wesen und Motivation des Backpropagation Algorithmus.

Referenzen I

Cybenko, G. 1989. "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals, and Systems*, 12.

Friedman, Avner. 1970. *Foundations of Modern Analysis*. Dover Publications.

Günther, Frauke, and Stefan Fritsch. 2010. "Neuralnet: Training of Neural Networks." *The R Journal* 2 (1): 30. https://doi.org/10.32614/RJ-2010-006.

Hanin, Boris, and Mark Sellke. 2018. "Approximating Continuous Functions by ReLU Nets of Minimal Width." *arXiv:1710.11278 [Cs, Math, Stat]*, March. https://arxiv.org/abs/1710.11278.

Hopfield, J. J. 1982. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proceedings of the National Academy of Sciences* 79 (8): 2554–58. https://doi.org/10.1073/pnas.79.8.2554.

Hornik, Kurt. 1991. "Approximation Capabilities of Multilayer Feedforward Networks." *Neural Networks* 4 (2): 251–57. https://doi.org/10.1016/0893-6080(91)90009-T.

Kidger, Patrick, and Terry Lyons. 2020. "Universal Approximation with Deep Narrow Networks." *arXiv:1905.08539 [Cs, Math, Stat]*, June. https://arxiv.org/abs/1905.08539.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

Leshno, Moshe, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. 1993. "Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function." *Neural Networks* 6 (6): 861–67. https://doi.org/10.1016/S0893-6080(05)80131-5.

Lu, Zhou, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. 2017. "The Expressive Power of Neural Networks: A View from the Width." *arXiv:1709.02540 [Cs]*, November. https://arxiv.org/abs/1709.02540.

McCulloch, Warren S, and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115–33.

Minsky, Marvin, and Seymour A. Papert. 1969. *Perceptrons: An Introduction to Computational Geometry*. 2. print. with corr. Cambridge/Mass.: The MIT Press.

Referenzen II

Ostwald, Dirk, and Franziska Usée. 2021. "An Induction Proof of the Backpropagation Algorithm in Matrix Notation." *arXiv:2107.09384 [Cs, Math, q-Bio, Stat]*, July. https://arxiv.org/abs/2107.09384.

Pinkus, Allan. 1999. "Approximation Theory of the MLP Model in Neural Networks." *Acta Numerica* 8 (January): 143–95. https://doi.org/10.1017/S0962492900002919.

Rosenblatt, F. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386–408. https://doi.org/10.1037/h0042519.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature*, no. 323: 533–36.

Schmidhuber, Jürgen. 2015. "Deep Learning in Neural Networks: An Overview." *Neural Networks* 61 (January): 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.