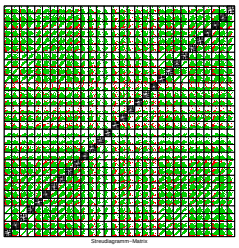
der Bundeswehr
Universität München

Dr. Robert Schmied



Professur für Mathematik
Fakultät für Elektrotechnik
und Informationstechnik
85577 Neubiberg
Tel.: 089/6004-3931
Fax: 089/6004-2615
robert.schmied@unibw.de

Explorative Statistik

WT 2016

6. Januar 2016

Inhaltsverzeichnis

I. Daten und deren Modellierung **3**

1. Ideen der Wahrscheinlichkeitstheorie **5**
1.1. Wahrscheinlichkeitsräume 5
1.2. Zufallsvariablen 10

2. Datenmatrix und Skalenarten **17**
2.1. Datenmatrix 17
2.2. Skalenarten 19
2.3. Absolute und relative Häufigkeiten 19

3. Merkmale mit Kardinalskala **23**
3.1. Kenngrößen für ein Merkmal 23
3.2. Kenngrößen für mehrere Merkmale 33

4. Merkmale mit Nominalskala **47**
4.1. Modus und Informationsentropie 47
4.2. Assoziationen 54

II. Zusammenhänge **63**

5. Hauptkomponentenanalyse **65**
5.1. Hauptachsentransformation 66
5.2. Hauptkomponentenmethode 71
5.3. Spezialfall: Kleine standardisierte Daten 72

6. Graphische Zusammenhangsanalyse **79**
6.1. Biplots 79
6.2. Nichtmetrische multidimensionale Skalierung 85

III. Abhängigkeiten **93**

7. Assoziationsregeln **95**
7.1. Modellierung von Assoziationsregeln 95
7.2. Support-Konfidenz-Ansatz zur Regelerzeugung 100
7.3. Erzeugung von Assoziationsregeln 103

8. Regressionsanalyse **111**
8.1. Exploration und Modellformulierung 111
8.2. Schätzung der Regressionsfunktion 112

8.3. Modell der multiplen Regression 117
8.4. Schätzung der Modellparameter 121

IV. Gruppierungen **127**

9. Klassifikation **129**
9.1. Bayessche Klassifikation 129
9.2. Support Vector Machines 132

10. Clusteranalyse **141**
10.1. Hierarchische Clusteranalyse 142
10.2. Klassifikatorische Clusteranalyse 146

Literatur **153**

Teil I.

Daten und deren Modellierung

1. Ideen der Wahrscheinlichkeitstheorie

Warum?

Konzepte und Methoden, die durch die Wahrscheinlichkeitstheorie bereitgestellt werden, helfen beim Beschreiben, Untersuchen und Beurteilen von erhobenen Daten. Sie sind eine wesentliche Grundlage für die Datenanalyse. Nur eine saubere Modellierung hilft dem Datenanalysten, Erkenntnisse zu gewinnen und zusammen mit Domänenexperten zu interpretieren. Ergebnisse von Analysen auf Basis wahrscheinlichkeitstheoretischer Modelle ohne entsprechende Hintergrundkenntnisse sind nicht sinnvoll.

1.1. Wahrscheinlichkeitsräume

Experimente,

- die nach einer bestimmten Vorschrift durchgeführt werden,
- die beliebig oft wiederholbar sind,
- deren Ausgang nicht vorhergesagt werden kann,

werden Zufallsexperimente genannt. Zu einem Zufallsexperiment gehören

- ein Ergebnisraum in Form einer Menge Ω,
- ein Elementarereignis $\omega \in \Omega$.
- Ereignisse.

Jedes Elementarereignis kann das Ergebnis des Zufallsexperiments sein. Ein Ereignis ist eine Menge $A \subseteq \Omega$ und tritt ein, wenn das Ergebnis ω des Zufallsexperiments in A enthalten ist, $\omega \in A$. Alle benötigten Ereignisse eines Zufallsexperiments werden zur Ereignismenge $\mathcal{F}$ zusammengefasst. Tabelle 1.1 zeigt einige oft benötigte Arten von Ereignissen. Die Sicherheit des Eintretens von Ereignissen wollen wir einschätzen. Dies soll über eine Funktion $\mathbb{P} : \mathcal{F} \to [0, 1]$ geschehen, an die wir gewisse Forderungen stellen wollen:

- $\mathbb{P}(\emptyset) = 0,\ \mathbb{P}(\Omega) = 1,\ \mathbb{P}(A) \geq 0$,
- $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup \bar{A}) = \mathbb{P}(A) + \mathbb{P}(\bar{A})$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ für $A \cap B = \emptyset$,
- $\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mathbb{P}(A_i)$ für paarweise disjunkte A_i.

Tabelle 1.1.: Beispiele für Ereignisse

Ereignis	Bezeichnung	Tore beim Fußball
Ω	Sicheres Ereignis	$\mathbb{N}_0$
$\emptyset$	Unmögliches Ereignis	$\emptyset$
$\bar{A} := \Omega \setminus A$	Das zu A komplementäre Ereignis	$A = \{0, 1, 2\}$, $\bar{A} = \{3, 4, 5, \ldots\}$
B mit $B \cap A = \emptyset$	Ein zu A disjunktes Ereignis	$A = \{1, 2\}$, $B = \{3, 4\}$
B mit $B \subseteq A$	Ein A implizierendes Ereignis	$A = \{0, 1, 2, 3, 4\}$, $B = \{0, 1, 2\}$

- Falls $B \subseteq A$, soll $\mathbb{P}(A) = \mathbb{P}(B \cup (A \setminus B)) = \mathbb{P}(B) + \mathbb{P}(A \setminus B)$ und damit $\mathbb{P}(B) \leq \mathbb{P}(A)$ sein.

Das unmögliche Ereignis tritt sicher nie, das sichere Ereignis immer ein. Alle anderen Ereignisse sollen dazwischenliegende Bewertungen erhalten. Die Bewertungen zweier komplementärer Ereignisse sollen addiert Eins ergeben, da eines der beiden Ereignisse sicher eintritt. Dies soll noch verallgemeinert werden. Für zwei disjunkte Ereignisse sollen sich die addierten Bewertungen so verhalten, wie wenn die beiden Ereignisse zusammen betrachtet werden. Das soll nicht nur für zwei sondern immer dann gelten, wenn die Ereignisse abgezählt werden können. Die Bewertung eines Ereignisses, das Teilmenge eines anderen Ereignisses ist, soll entsprechend kleiner sein. Jede diese Kriterien erfüllende Funktion ordnet Ereignissen eine Wahrscheinlichkeit zu.

Das Maßproblem
Leider gibt es für den wichtigen Standardfall $\Omega = \mathbb{R}$ und die Potenzmenge $\mathcal{F} = \mathcal{P}(\mathbb{R})$ keine solche Funktion $\mathbb{P}$. Dennoch ist der Ansatz der Abbildung von Ereignissen in das Intervall $[0, 1]$ nachvollziehbar. Die geforderten Eigenschaften scheinen plausibel. Deshalb scheint es eine sinnvolle Idee zu sein, die Ereignismenge $\mathcal{F}$ so einzuschränken, dass die Eigenschaften erhalten bleiben. Eine geeignete Ereignismenge ist die so genannte σ-Algebra.

Definition 1.1: σ-Algebra

Ein Mengensystem $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ bestehend aus Teilmengen des Ergebnisraums Ω heißt σ-Algebra, wenn es die drei folgenden Eigenschaften erfüllt:

- $\Omega \in \mathcal{F}$,
- $\bar{A} \in \mathcal{F}$ für alle $A \in \mathcal{F}$,
- $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$ für alle $A_i \in \mathcal{F}$.

Bemerkung.
Jede σ-Algebra ist eine Ereignismenge.

Auf Basis einer σ-Algebra können wir eine den Ereignissen eines Zufallsexperiments Wahrscheinlichkeiten zuordnende Funktion definieren in

Definition 1.2: Wahrscheinlichkeitsmaß

Sei $\mathcal{F}$ eine σ-Algebra eines Ereignisraums Ω. Eine Funktion $\mathbb{P} : \mathcal{F} \to \mathbb{R}$, die jedem Ereignis A aus $\mathcal{F}$ eine reelle Zahl zuordnet, heißt Wahrscheinlichkeitsmaß, wenn die drei folgenden Axiome erfüllt sind:

- $\mathbb{P}(A) \geq 0$ für alle $A \in \mathcal{F}$,
- $\mathbb{P}(\Omega) = 1$,
- $\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$ für paarweise disjunkte A_i.

Das Zufallsexperiment wird nun durch das Festlegen dreier Bestandteile modelliert. Dabei ist die Ergebnismenge der erste Baustein, auf den die anderen aufsetzen.

Definition 1.3: Wahrscheinlichkeitsraum, Träger

$(\Omega, \mathcal{F}, \mathbb{P})$ ist ein Wahrscheinlichkeitsraum, falls

- Ω ein Ergebnisraum,
- $\mathcal{F}$ eine σ-Algebra über Ω,
- $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $\mathcal{F}$ ist.

$T \subseteq \Omega$ mit $T = \{\omega \in \Omega;\ \mathbb{P}(\{\omega\}) > 0\} \subseteq \Omega$ heißt der Träger von $\mathbb{P}$. Ist T eine höchstens abzählbare Menge mit $\mathbb{P}(T) = 1$, so heißt der Wahrscheinlichkeitsraum diskret, ansonsten stetig.

Beispiel 1.4: Diskreter Wahrscheinlichkeitsraum

Die Anzahl erzielter Tore bei einem Fußballspiel kann über den Wahrscheinlichkeitsraum $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0), \mathbb{P})$ mit

$$\mathbb{P} : \mathcal{P}(\mathbb{N}_0) \to [0,1],\ A \mapsto \mathbb{P}(A) := \sum_{\omega\in A} e^{-\lambda} \cdot \frac{\lambda^\omega}{\omega!}$$

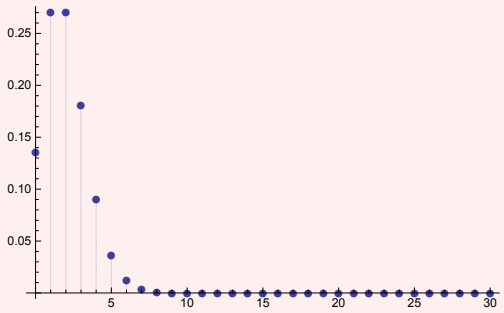
mit einem Parameter $\lambda > 0$, der die erwartete Trefferzahl beschreibt, modelliert werden.

Bemerkung.
(1) Das Wahrscheinlichkeitsmaß eines diskreten Wahrscheinlichkeitsraums heißt diskretes Wahrscheinlichkeitsmaß.
(2) Durch $f : T \to [0,1]$ mit $\mathbb{P}(\{\omega\}) = f(\omega)$, $\sum_{\omega\in T} f(\omega) = 1$ wird das diskrete Wahrschein-

lichkeitsmaß vollständig charakterisiert. Dies wird als Zähldichte oder diskrete Dichte bezeichnet. Im Beispiel wird die Zähldichte $\mathbb{P}(\{\omega\}) = f(\omega) := e^{-\lambda} \cdot \frac{\lambda^\omega}{\omega!}$ benutzt.
(3) Das diskrete Wahrscheinlichkeitsmaß im Beispiel heißt Poisson-Verteilung[1].

Beispiel 1.5

Annahme: Die erwartete Anzahl erzielter Treffer bei einem Fußballspiel ist 2. Wir setzen $\lambda = 2$ und betrachten die dazugehörige Zähldichte der Poisson-Verteilung.



$$\mathbb{P}(\{0,1,2,3\}) = \sum_{\omega=0}^{3} e^{-2} \cdot \frac{2^\omega}{\omega!} = \frac{19}{3e^2} = 0.86.$$

Sei T Träger von $\mathbb{P}$ eines diskreten Wahrscheinlichkeitsraums, dessen Elemente geordnet sind. Einer diskreten Dichte $f : T \to [0,1]$ lässt sich eine Funktion

$$F_\mathbb{P} : T \to [0,1],\ t \mapsto F_\mathbb{P}(t) := \sum_{\substack{\omega \leq t \\ \omega \in T}} f(\omega)$$

zuordnen. In diesem Fall sprechen wir von einer diskreten Verteilungsfunktion.

Kehren wir zu der Frage zurück, welche Ereignismenge für den Ergebnisraum $\mathbb{R}$ die Bildung eines Wahrscheinlichkeitsmaßes ermöglicht. Dazu bedarf es einer Menge, die alle „interessanten“ Mengen wie z.B. Einpunktmengen, beliebige Intervalle oder abzählbare Vereinigungen oder endliche Durchschnitte von Intervallen enthält. Solche Mengen werden oft in Anwendungen benötigt. Die Borelsche σ-Algebra $\mathcal{B}$ ist die Ereignismenge der reellen Zahlen und erfüllt diese Anforderungen. Ein Wahrscheinlichkeitsmaß $\mathbb{P} : \mathcal{B} \to [0,1]$ bestimmt z.B.

- $\mathbb{P}(]-\infty, a])$.
- $\mathbb{P}(]a,b]) = \mathbb{P}(]-\infty,b]\backslash]-\infty,a]) = \mathbb{P}(]-\infty,b]) - \mathbb{P}(]-\infty,a]),$

d.h. $]-\infty,a],]a,b] \in \mathcal{B}$ für $a,b \in \mathbb{R}$. Analog zu diskreten Wahrscheinlichkeitsmaßen heißt das Wahrscheinlichkeitsmaß eines stetigen Wahrscheinlichkeitsraums stetiges Wahrscheinlichkeitsmaß. Ebenso gibt es wie im diskreten Fall mit der diskreten Dichte eine stetige Dichte.

[1] S. D. Poisson, 1781-1840

Definition 1.6: Verteilungsfunktion, Dichte

Ist $\mathbb{P} : \mathcal{B} \to [0,1]$ ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$, so heißt

$$F_{\mathbb{P}} : \mathbb{R} \to [0,1],\ x \mapsto F_{\mathbb{P}}(x) := \mathbb{P}(]-\infty, x])$$

die Verteilungsfunktion von $\mathbb{P}$. Eine Funktion $f : \mathbb{R} \to \mathbb{R}_0^+$, für die $\int_{-\infty}^{\infty} f(x)dx = 1$ und

$$F_{\mathbb{P}}(x) = \int_{-\infty}^{x} f(x)dx$$

gilt, heißt stetige Dichte.

Bemerkung.
(1) $F_{\mathbb{P}}$ ist monoton wachsend, rechtsseitig stetig und es gelten die Grenzwerte $\lim_{x\to\infty} F_{\mathbb{P}}(x) = 1$ und $\lim_{x\to-\infty} F_{\mathbb{P}}(x) = 0$.
(2) Oft interessieren uns lediglich die Verteilungsfunktion oder die stetige Dichte und nicht das eigentliche Wahrscheinlichkeitsmaß. Dann schreiben wir F anstelle von $F_{\mathbb{P}}$.

Beispiel 1.7: Normalverteilung

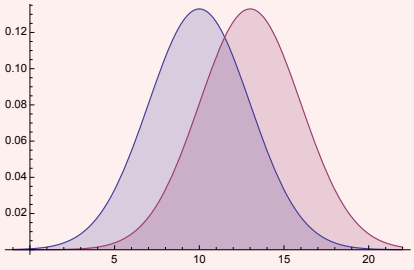
Eine der wichtigsten stetigen Dichten ist durch die Normalverteilung über

$$f : \mathbb{R} \to \mathbb{R}^+,\ x \mapsto f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}$$

mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$ gegeben. Für $\mu = 0$ und $\sigma^2 = 1$ sprechen wir von der Standardnormalverteilung. Für die Verteilungsfunktion $F = F_{(\mu,\sigma^2)}$ gibt es lediglich Näherungswerte. Es sei $\Phi = F_{(0,1)}$. Eine beliebige Normalverteilung lässt sich aus der Standardnormalverteilung mittels der Substitution

$$F_{(\mu,\sigma^2)}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

gewinnen. Das Bild zeigt den Graphen der Normalverteilung für die Parameter $(\mu, \sigma^2) = (10, 9)$ bzw. $(\mu, \sigma^2) = (13, 9)$.



Eine Verteilungsfunktion $F_{\mathbb{P}} : M \to [0,1]$ hilft uns dabei, verschiedene Fragestellungen zu beantworten. Neben dem Wert von $F_{\mathbb{P}}$ an verschiedenen Stellen $x \in M$, z.B. für die Standardnormalverteilung

$$\Phi(1) = 0.8413, \Phi(-1) = 0.1587 \Rightarrow \Phi(1) - \Phi(-1) = 0.6827, \tag{1.1}$$

lässt sich umgekehrt die Frage stellen, für welches $x \in M$ denn $F_{\mathbb{P}}(x) = \alpha$ für $0 < \alpha < 1$ gilt. Da es ein solches x nicht geben muss oder es nicht eindeutig bestimmt sein muss, nennen wir

$$F_{\mathbb{P}}^{-1} : (0,1) \to M,\ \alpha \mapsto F_{\mathbb{P}}^{-1}(\alpha) := \inf\{x \in M;\ F_{\mathbb{P}}(x) \geq \alpha\} \tag{1.2}$$

die Quantilfunktion von $\mathbb{P}$ und $F_{\mathbb{P}}^{-1}(\alpha)$ das α-Quantil von $\mathbb{P}$.

Beispiel 1.8: Median

Ein besonders wichtiges α-Quantil ist der Median $F_{\mathbb{P}}^{-1}(0.5)$. Für die Standardnormalverteilung gilt $\Phi^{-1}(0.5) = 0$. Für die allgemeine Normalverteilung ist

$$F_{(\mu,\sigma^2)}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = 0.5 \Leftrightarrow \frac{x-\mu}{\sigma} = 0 \Leftrightarrow x = \mu.$$

Für $x = \mu$ erhalten wir den maximalen Wert der Dichtefunktion der Normalverteilung. Denn, da die Dichtefunktion differenzierbar ist, folgt

$$\begin{aligned}
\frac{d}{dx}\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}\right) &= -\frac{1}{\sigma^2\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}} \cdot \frac{(x-\mu)}{\sigma} = 0 \Leftrightarrow x = \mu, \\
\frac{d^2}{dx^2}\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}}\right) &= \frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}} \cdot \left(\frac{(x-\mu)^2}{\sigma^2} - 1\right) \overset{x=\mu}{<} 0
\end{aligned}$$

Der maximale endliche Wert - falls es ihn gibt - einer diskreten oder stetigen Dichte heißt Modalwert, der oder die dazugehörigen Werte der Definitionsmenge der Dichte heißen Modus.

1.2. Zufallsvariablen

Oft sind wir nicht nur an einem Zufallsexperiment interessiert, sondern an einer Verknüpfung mehrerer möglicherweise identischer Zufallsexperimente. Dazu brauchen wir so genannte Zufallsvariablen.

Definition 1.9: Zufallsvariable

Es sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und $(\Psi, \mathcal{G})$ ein Tupel bestehend aus einer Menge Ψ und einer dazugehörigen σ-Algebra $\mathcal{G}$ auf Ψ. Eine Abbildung $X : \Omega \to \Psi$ heißt Zufallsvariable, wenn für jedes $G \in \mathcal{G}$ gilt $X^{-1}(G) \in \mathcal{F}$.

Bemerkung.
(1) Die Eigenschaft $X^{-1}(G) \in \mathcal{F}$ für jedes $G \in \mathcal{G}$ wird auch als $\mathcal{F}$-$\mathcal{G}$-Messbarkeit von X bezeichnet.

(2) Auf $\mathcal{G}$ ist ein Wahrscheinlichkeitsmaß $\mathbb{P}_X : \mathcal{G} \to [0,1]$ durch $\mathbb{P}_X(G) = \mathbb{P}(X^{-1}(G))$ definiert. $(\Psi, \mathcal{G}, \mathbb{P}_X)$ ist damit ein Wahrscheinlichkeitsraum.
(3) Eigenschaften eines Wahrscheinlichkeitsmaßes $\mathbb{P}_X : \mathcal{G} \to [0,1]$ werden als Eigenschaften der Zufallsvariablen X deklariert. Wir schreiben z.B.

- $\mathbb{P}_X(A) = \mathbb{P}(X \in A)$ für ein Ereignis $A \in \mathcal{G}$,
- $\mathbb{P}_X(\{a\}) = \mathbb{P}(X = a)$,
- $X \sim N(\mu, \sigma^2)$, wenn das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ auf $\mathcal{G}$ einer Normalverteilung mit Parametern μ und σ^2 entspricht,
- $X \sim Po(\lambda)$, wenn das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ auf $\mathcal{G}$ einer Poisson-Verteilung mit Parameter λ entspricht.

Wollen wir die Anzahl Tore zweier Fußballspiele untersuchen, so lässt sich jedes Spiel für sich alleine mittels einer Poisson-Verteilung modellieren. Gibt es eine Möglichkeit beide Spiele zusammen zu betrachten und zu bestimmen, mit welcher Wahrscheinlichkeit eine bestimmte Anzahl an Toren erzielt wird? Zunächst betrachten wir einen Wahrscheinlichkeitsraum

$$(\mathbb{N}_0^2, \mathcal{P}(\mathbb{N}_0^2), \mathbb{P}).$$

Ein Ergebnis könnte etwa $\omega = (2,3) \in \mathbb{N}_0^2$ sein, im ersten Spiel fallen zwei und im zweiten drei Tore. Nun interessiert uns die Summe der Tore. Wir modellieren dies mit $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$ über die Zufallsvariable

$$X : \mathbb{N}_0^2 \to \mathbb{N}_0,\ (\omega_1, \omega_2) \mapsto \omega_1 + \omega_2.$$

Das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ bekommen wir, wenn wir das Wahrscheinlichkeitsmaß $\mathbb{P}$ kennen. Doch wie sieht das aus? Um diese Frage zu beantworten benötigen wir noch etwas Vorarbeit. Zunächst klären wir, dass die Parameter einer Verteilung meist eine ganz bestimmte Bedeutung haben. Sie charakterisieren eine Verteilung.

Definition 1.10: Erwartungswert und Varianz von Zufallsvariablen

Es sei f eine diskrete oder stetige Dichte einer Zufallsvariablen X mit Parametern $\theta \in \Theta$ in einem Parameterraum Θ. Dann heißt

- diskret: $\mathbb{E}_\theta[X] := \sum_{x \in T} x \cdot f(x)$
- stetig: $\mathbb{E}_\theta[X] := \int_{-\infty}^{\infty} x \cdot f(x) dx$

der Erwartungswert und

- diskret: $\mathbb{V}_\theta[X] := \sum_{x \in T} (x - \mathbb{E}_\theta[X])^2 \cdot f(x) = \sum_{x \in T} x^2 \cdot f(x) - \mathbb{E}_\theta[X]^2$
- stetig: $\mathbb{V}_\theta[X] := \int_{-\infty}^{\infty} (x - \mathbb{E}_\theta[X])^2 \cdot f(x) dx = \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \mathbb{E}_\theta[X]^2$

die Varianz von X

Bemerkung.
Bei der Poisson-Verteilung haben wir $\Theta = \mathbb{R}^+$, bei der Normalverteilung ist $\Theta = \mathbb{R} \times \mathbb{R}^+$.

Definition 1.11: Kovarianz zweier Zufallsvariablen

Sind f, g zwei entweder diskrete oder stetige Dichten der Zufallsvariablen X und Y mit Parametern $\theta \in \Theta$ und $\psi \in \Psi$, so wird die Kovarianz zwischen X und Y durch

$$\begin{aligned}\mathrm{Cov}_{(\theta,\psi)}[X,Y] &:= \mathbb{E}_{(\theta,\psi)}[(X-\mathbb{E}_\theta[X])(Y-\mathbb{E}_\psi[Y])] \\ &= \mathbb{E}_{(\theta,\psi)}[XY]-\mathbb{E}_\theta[X]\mathbb{E}_\psi[Y].\end{aligned}$$

definiert.

Bemerkung.
Es ist $\mathbb{V}_\theta[X] = \mathrm{Cov}_\theta[X,X]$.

Für Erwartungswerte und Kovarianzen gelten folgende Eigenschaften:

- $\mathbb{E}_\theta[aX+b] = a\mathbb{E}_\theta[X]+b,\ a,b\in\mathbb{R}$,
- $\mathrm{Cov}_{(\theta,\psi)}[aX+b,cY+d] = ac\cdot\mathrm{Cov}_{(\theta,\psi)}[X,Y],\ a,b,c,d\in\mathbb{R}$,
- $\mathbb{V}_\theta[aX+b] = \mathrm{Cov}_\theta[aX+b,aX+b] = a^2\mathbb{V}_\theta[X],\ a,b\in\mathbb{R}$.

Satz 1.12

Sei $X\sim N(\mu,\sigma^2)$, d.h. X ist normalverteilt mit Parametern μ und σ^2. Für den Erwartungswert von X gilt $\mathbb{E}_{(\mu,\sigma^2)}[X]=\mu$.

Beweis.

Zunächst ist

$$\begin{aligned}\int x\cdot\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx &= \int(x-\mu+\mu)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx \\ &= \int(x-\mu)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx \\ &\quad+\int\mu\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx \\ &\overset{u=\left(\frac{x-\mu}{\sigma}\right)^2}{=} \int\frac{\sigma}{2\sqrt{2\pi}}e^{-\frac{1}{2}u}du \\ &\quad+\mu\int\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx \\ &= -\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{1}{2}u}+c \\ &\quad+\mu\int\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx.\end{aligned}$$

Damit folgt $\mathbb{E}_{(\mu,\sigma^2)}[X] = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = 0 + \mu \cdot 1 = \mu.$

□

Satz 1.13

Sei $X \sim N(\mu, \sigma^2)$, d.h. X ist normalverteilt mit Parametern μ und σ^2. Für die Varianz von X gilt $\mathbb{V}_{(\mu,\sigma^2)}[X] = \sigma^2$.

Beweis.

$$\begin{aligned}
\mathbb{V}_{(\mu,\sigma^2)}[X] &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&\overset{u=\frac{x-\mu}{\sigma}}{=} \int_{-\infty}^{\infty} \sigma^3 u^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cdot u e^{-\frac{1}{2}u^2} du \\
&\overset{p.I.}{=} \frac{\sigma^2}{\sqrt{2\pi}} \left[-u e^{-\frac{1}{2}u^2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} du \right] \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left[0 + \sqrt{2\pi} \right] = \sigma^2
\end{aligned}$$

□

Bemerkung.
Für $X \sim Po(\lambda)$ lässt sich zeigen[2], dass $\mathbb{E}_\lambda[X] = \lambda = \mathbb{V}_\lambda[X]$. Die Parameter einer Verteilung stehen oft für den Erwartungswert oder die Varianz der Zufallsvariablen.

Mehrdimensionale Zufallsvariablen

Sind $X_1, \ldots, X_n$ Zufallsvariablen, so heißt eine Funktion $f_{X_1 \ldots X_n} = f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$ mit

- diskreter Fall:
$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_i = x_i;\ i = 1, \ldots, n)$$
- stetiger Fall:
$$\int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_n \ldots dx_1 = \mathbb{P}(X_i \in [a_i, b_i];\ a_i < b_i,\ i = 1, \ldots, n)$$

[2] siehe etwa [4]

gemeinsame Wahrscheinlichkeitsdichtefunktion von $X_1, \ldots, X_n$. Sind

$$X_1 : U_1 \to V_1, \ldots, X_n : U_n \to V_n$$

Zufallsvariablen mit einer gemeinsamen Wahrscheinlichkeitsdichtefunktion

$$f_{X_1 \ldots X_n} = f_{X_1 \ldots X_n}(x_1, \ldots, x_n),$$

so heißt eine Funktion f_{X_i} mit

- diskreter Fall:

$$f_{X_i}(x_i) = \mathbb{P}(X_i = x_i) = \sum_{x_1 \in U_1} \ldots \sum_{x_{i-1} \in U_{i-1}} \sum_{x_{i+1} \in U_{i+1}} \ldots \sum_{x_n \in U_n} f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$$

- stetiger Fall:

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

Randverteilung von $X_i,\ i = 1, \ldots, n$.

Definition 1.14: Stochastische Unabhängigkeit

Die Zufallsvariablen $X_1, \ldots, X_n$ mit einer gemeinsamen Wahrscheinlichkeitsdichtefunktion $f_{X_1 \ldots X_n} = f_{X_1 \ldots X_n}(x_1, \ldots, x_n)$ heißen stochastisch unabhängig, wenn

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n)$$

gilt.

Bemerkung.
Oft wird angenommen, dass n Zufallsvariablen $X_1, \ldots, X_n$ stochastisch unabhängig und identisch verteilt sind (Schreibweise: u.i.v.).

Beispiel 1.15

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Mit den beiden Zufallsvariablen $X_1 : \Omega \to \mathbb{N}_0$ und $X_2 : \Omega \to \mathbb{N}_0$ wollen wir die Anzahl Tore bei zwei Fußballspielen modellieren. Wir nehmen an, dass die beiden Zufallsvariablen identisch Poisson-verteilt mit Parameter λ und stochastisch unabhängig sind, d.h. für die gemeinsame Wahrscheinlichkeitsdichtefunktion $f_{X_1 X_2}(x_1, x_2)$ von X_1 und X_2 gelte

$$f_{X_1 X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = e^{-\lambda} \cdot \frac{\lambda^{x_1}}{x_1!} \cdot e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!} = e^{-2\lambda} \cdot \frac{\lambda^{x_1 + x_2}}{x_1! \cdot x_2!}.$$

Sei

$$(\mathbb{N}_0^2, \mathcal{P}(\mathbb{N}_0^2), \mathbb{P})$$

der dazugehörige Wahrscheinlichkeitsraum mit $\mathbb{P}(x_1, x_2) = f_{X_1 X_2}(x_1, x_2)$. Ein Ergeb-

nis könnte etwa $\omega = (2,3) \in \mathbb{N}_0^2$ sein, im ersten Spiel fallen zwei und im zweiten drei Tore. Nun interessiert uns die Summe der Tore. Wir modellieren dies mit $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$ über die Zufallsvariable

$$X : \mathbb{N}_0^2 \to \mathbb{N}_0,\ (\omega_1, \omega_2) \mapsto \omega_1 + \omega_2.$$

Das Wahrscheinlichkeitsmaß $\mathbb{P}_X$ bekommen wir durch folgende Überlegung:

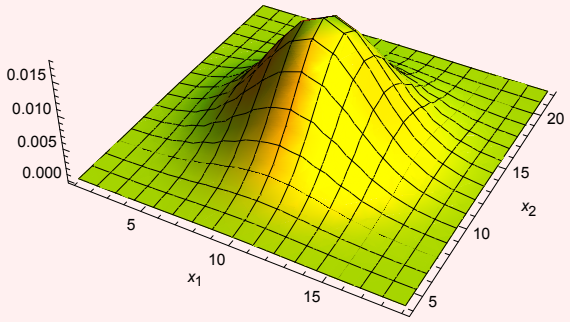
$$\begin{aligned}
\mathbb{P}(X = n) = \mathbb{P}_X(\{n\}) &= \mathbb{P}(X^{-1}(\{n\})) = \mathbb{P}(\{(x_1, x_2) \in \mathbb{N}_0^2;\ x_1 + x_2 = n\}) \\
&= \mathbb{P}(\{(x_1, n - x_1);\ x_1 \in \mathbb{N}_0\}) \\
&= \sum_{x_1=0}^{n} e^{-2\lambda} \cdot \frac{\lambda^n}{x_1! \cdot (n - x_1)!} \\
&= \frac{e^{-2\lambda}}{n!} \sum_{x_1=0}^{n} \frac{n!}{x_1! \cdot (n - x_1)!} \lambda^k \cdot \lambda^{n-k} \\
&= \frac{e^{-2\lambda}}{n!} \cdot (2\lambda)^n.
\end{aligned}$$

Es ist somit $X \sim Po(2\lambda)$.

Beispiel 1.16: Zweidimensionale Normalverteilung

Seien $X_1 \sim N(10, 9)$ und $X_2 \sim N(13, 9)$ stochastisch unabhängig. Die gemeinsame Wahrscheinlichkeitsdichtefunktion lautet

$$f(x_1, x_2) = \frac{1}{3 \cdot 3 \cdot 2\pi} e^{-\frac{1}{2}\left(\frac{(x_1-10)^2}{9} + \frac{(x_2-13)^2}{9}\right)}.$$



Sind (μ_1, σ_1^2) und (μ_2, σ_2^2) die Parameter von X_1 bzw. X_2, so kann mit einer etwas aufwändigeren Rechnung analog zu Beispiel 1.15 gezeigt werden, dass für die Zufallsvariable $X : \mathbb{R}^2 \to \mathbb{R}$, $(x_1, x_2) \mapsto X(x_1, x_2) = x_1 + x_2$ unter Annahme der stochastischen

Unabhängigkeit die stetige Dichte

$$f(x) = \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\frac{(x-(\mu_1+\mu_2))^2}{\sigma_1^2+\sigma_2^2}}$$

resultiert. Damit folgt $X \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Seien mit m der Erwartungswert und mit s die Standardabweichung vorgegeben. In R können wir die Normalverteilung (bzw. die Poisson-Verteilung mit Parameter l) über die Befehle

dnorm(x ,**mean**=m, **sd**=s)

Wert der Dichte der Normalverteilung bei x (analog: dpois(x,lambda=l)),

pnorm(**q** ,**mean**=m, **sd**=s)

Wert der Verteilungsfunktion der Normalverteilung bei q (ppois(q,lambda=l)),

qnorm(p ,**mean**=m, **sd**=s)

p-Quantil der Normalverteilung (qpois(p,lambda=l)),

rnorm(n ,**mean**=m, **sd**=s)

Eine Stichprobe vom Umfang n unabhängig identisch normalverteilter Zufallsvariablen (rpois(n,lambda=l)),

benutzen.

2. Datenmatrix und Skalenarten

Warum?

Die Grundlage aller Modelle in der Datenanalyse sind Daten. Daten sind die erhobenen Werte festgelegter Eigenschaften. Oft werden Daten heute ungeordnet erhoben (Stichwort: Big Data). Zur Weiterverarbeitung müssen die Daten jedoch vorbereitet und in eine bestimmte Form gebracht werden. Diese Form nennen wir eine Datenmatrix. Dabei steht jede Spalte für eine Eigenschaft und jede Zeile für die gemeinsamen Werte der jeweiligen Eigenschaften bei einem Untersuchungsobjekt. Dabei ist zu beachten, dass die Werte einer Eigenschaft aus einer festgelegten Wertemenge stammen und mit dieser Wertemenge zudem festgelegt werden muss, wie deren Elemente weiterverarbeitet werden dürfen.

Die Grundannahme der Mathematischen Statistik, die laut [9] dadurch gegeben ist, die „Beobachtungen als Realisierungen von Zufallsgrößen aufzufassen und damit (zu) unterstellen, dass sich der Vorgang durch eine Wahrscheinlichkeitsverteilung beschreiben lässt“, liefert eine Schnittstelle zwischen einer daten- und modellgetriebenen Untersuchung in der Stochastik. Eine Beobachtung, z.B. eine Messung bei der Durchführung eines physikalischen Experiments, ist demnach dem Zufall unterworfen. Zufallseinflüsse werden durch einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$, der nicht näher spezifiziert wird, modelliert.

2.1. Datenmatrix

Für jede Datenerhebung gibt es eine Gesamtheit aller Information tragenden Objekte. Die Gesamtheit der Objekte wird durch Identifikationsmerkmale d.h. sachliche, räumliche oder zeitliche Kriterien eindeutig festgelegt. Die Anzahl Objekte ist in realen Situationen oftmals so groß, dass nicht für jedes einzelne Objekt Daten erhoben werden können. Dann wird eine Teilmenge herangezogen, die repräsentativ für die Gesamtheit aller Objekte ist. Ein einzelnes Objekt der Untersuchung muss sich von anderen Objekten ebenfalls durch Identifikationsmerkmale eindeutig unterscheiden. Ein einzelnes Objekt, das die für eine Untersuchung interessierende Information trägt, heißt Merkmalsträger $g_{i\cdot}$, $i \in I$, wobei I eine beliebige Indexmenge sei. Der Punkt im Index deute an, dass es zu jedem Merkmalsträger Ausprägungen zu verschiedenen Eigenschaften geben kann. Wir fassen alle durch festgelegte Identifikationsmerkmale zu einer virtuellen Gesamtheit zusammengebrachten Merkmalsträger konkret in der Grundgesamtheit $G := \{g_{i\cdot};\ i \in I\}$ zusammen. Wie die einzelnen Merkmalsträger erfasst werden, spielt dabei keine Rolle. Sie werden nun als Einheit in der Grundgesamtheit betrachtet. Werden die Merkmalsträger in der Grundgesamtheit G zusammengefasst, so nehmen wir an, dass die Merkmalsträger gleiche jedem von ihnen gegebene Eigenschaften haben, die empirisch beobachtet oder gemessen werden können.

Die Beobachtungsdaten werden nach bestimmten Gesichtspunkten entsprechend einer vorgegebenen Fragestellung gesammelt. Eine zentrale Aufgabe besteht in der Modellierung der durch die Fragestellung gestellten Aufgabe. Die Modellierung umfasst dabei nicht nur die Festlegung von zu erhebenden Merkmalen, der Träger der Merkmale und die Überlegung, auf welche Weise die Informationen abgegriffen und quantifiziert werden, sondern auch je nach geplanter Weiterverarbeitung der Daten die Festlegung des zugrunde liegenden stochastischen Modells. Dies erfordert auch einen gewissen Mut, sich auf ein bestimmtes Modell festzulegen. Das hat aber die Konsequenz, dass die Ergebnisse nachfolgender Datenanalysen bereits aufgrund der Modellierung in eine bestimmte Richtung gelenkt werden können. Für jedes Beobachtungsdatum sind drei Sichten zu unterscheiden: Welcher Wertemenge entstammt das Datum, welche Eigenschaft repräsentiert das Datum und welchem Objekt ist das Datum zugeordnet? Ein Datum ohne Zufallseinfluss kann damit als Wert einer Abbildung von der Menge der Objekte in die vorgesehene Wertemenge, den Merkmalsraum M, angesehen werden, $e : G \to M$. Die unbekannte Abbildung $e \in \mathbb{E}$ entstammt einem Funktionenraum $\mathbb{E}$ und steht dabei für die festgelegte gemeinsame Eigenschaft der Objekte.

Da nicht immer alle Merkmalsträger einer Grundgesamtheit untersucht werden können, muss eine Auswahl getroffen werden. Es wird also lediglich von einer Teilmenge der Grundgesamtheit eine Datenerhebung vorgenommen (eine so genannte Teilerhebung). Eine solche Teilmenge $S \subseteq G$ heißt Stichprobenmenge der Grundgesamtheit G. $S = \{s_{1\cdot}, \ldots, s_{n\cdot}\} \subseteq G$ mit $n \in \mathbb{N}$ sei als endlich vorausgesetzt. n heißt der Stichprobenumfang. Seien nun $\mathbb{E} = \{e_1, \ldots, e_k\}$, $e_j : S \to M_j$ $(j = 1, \ldots, k)$ und $M = \bigcup_{j=1}^{k} M_j$. Eine zufallsunabhängige Beobachtung kann nun als Abbildung $\psi_S : \mathbb{E} \to M^n$, $e \mapsto \psi_S(e) := (x_1, \ldots, x_n)^T := (e(s_{1\cdot}), \ldots, e(s_{n\cdot}))^T$ beschrieben werden. Wird die Grundgesamtheit herangezogen, wird dann anstelle von ψ_S entsprechend ψ_G geschrieben mit $\psi_G(e) = (e(g_{1\cdot}), e(g_{2\cdot}), \ldots)^T$. Der Vektor $\mathbf{x} := (x_1, \ldots, x_n)^T$ heißt Stichprobe vom Umfang n.

Die Abbildung e, die eine Eigenschaft der Merkmalsträger repräsentiert, wird deswegen auch als Merkmal bezeichnet. Die Vektoren von k zufalls(un)abhängigen Beobachtungen mit Stichprobenumfang n werden in der Datenmatrix

$$\begin{array}{c} \\ s_{1\cdot} = g_{i_1\cdot} \\ s_{2\cdot} = g_{i_2\cdot} \\ \vdots \\ s_{n\cdot} = g_{i_n\cdot} \end{array} \begin{array}{c} \begin{array}{cccc} e_1 & e_2 & \ldots & e_k \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \end{array} = (x_{ij}) =: X$$

zusammengefasst. Jedes x_{ij} heißt Merkmalswert bzw. Datum des Merkmalsträgers $s_{i\cdot}$ für das Merkmal e_j. Liegt eine Datenmatrix vor, sind eine Vielzahl verschiedener Fragestellungen möglich. Eine Auswahl daraus soll in den weiteren Kapiteln betrachtet werden. Abkürzend schreiben wir meist X_j und meinen dabei sowohl das Merkmal e_j als auch den Vektor $\mathbf{x}_{\cdot\mathbf{j}}$ der Merkmalswerte aller Merkamlsträger für das Merkmal e_j.

2.2. Skalenarten

Es gibt verschiedene Arten von Merkmalen:

- Merkmale mit Nominalskala: Einzelne Merkmalswerte können lediglich hinsichtlich ihrer Gleichheit oder Ungleichheit unterschieden werden. Als Merkmalsraum verwenden wir endliche Mengen, deren Elemente Kategorien genannt werden. Wir sprechen hier von qualitativen Merkmalen.
- Merkmale mit Ordnialskala: Die Merkmalswerte können zudem in eine Reihenfolge „$\preceq$“ gebracht werden. Es sind keine Abstände zwischen den einzelnen Merkmalswerten interpretierbar.
- Merkmale mit Kardinalskala: Es können zusätzlich Abstände zwischen Merkmalswerten gebildet und interpretiert werden. Wir werden für solche Merkmale als Merkmalsraum die reellen Zahlen benutzen, $M = \mathbb{R}$. Sie sind die in den Ingenieurwissenschaften am häufigsten anzutreffende Merkmalsart.

Jedes Merkmal mit Kardinalskala kann als Merkmal mit Ordinalskala und jedes Merkmal mit Ordinalskala als Merkmal mit Nominalskala aufgefasst werden. Dies ist jeweils mit einem Informationsverlust verbunden. Nicht-qualitative Merkmale nennen wir auch quantitative Merkmale.

Es bezeichne

$$B := \{c \in M;\ c = x_i \text{ für ein } i = 1, \ldots, n\} \tag{2.1}$$

die Menge der beobachteten Ausprägungen $c \in M$ einer Stichprobe **x** vom Umfang n des Merkmals X. Liegt eine stochastische Modellierung der Datenmatrix zugrunde, sprechen wir auch von den Realisierungen anstelle der beobachteten Ausprägungen. Wir werden beide Begriffe synonym verwenden.

2.3. Absolute und relative Häufigkeiten

Allen Skalenarten ist gemein, dass bei einer Stichprobe die beobachteten Ausprägungen zusammengefasst und ihre Anzahl ausgezählt werden kann. Insbesondere bei qualitativen Merkmalen ist eine sinnvolle und wichtige Beschreibung über das Auszählen der Realisierungen, das Bestimmen von Häufigkeiten, möglich. Wir betrachten zunächst ein qualitatives Merkmal X mit dem endlichen Merkmalsraum $M = \{c_1, \ldots, c_m\}$. Bei qualitativen Merkmalen steht keine Ordnung und keine sinnvolle Abstandsdefinition zweier Merkmalsausprägungen zur Verfügung. Wir unterscheiden zwei Begriffe.

Auszählen von Häufigkeiten

Qualitative Merkmale können in einer Häufigkeitstabelle erfasst werden. Dabei werden für insgesamt n Merkmalsträger $s_1, \ldots, s_n$ einer Stichprobe S und $m = |M|$ Kategorien $c_1, \ldots, c_m$ eines qualitativen Merkmals X entweder absolute Häufigkeiten in Form der Anzahl der Fälle n_l in der Kategorie c_l oder relative Häufigkeiten in Form der Proportion h_l in der Kategorie c_l erfasst. Um die Anzahl auszuzählen, benötigen wir eine Indikatorfunktion

$$I_X^L : M \to \{0,1\}, \quad x \mapsto I_X^L(x) := \begin{cases} 1, & x \in L \text{ (hier: } L = \{c_l\}), \\ 0, & \text{sonst.} \end{cases} \tag{2.2}$$

Die Anzahl der Fälle und die Proportion einer Stichprobe ergeben sich dann über

$$aH : M^n \to \mathbb{N}_0^m,\ \mathbf{x} \mapsto aH(\mathbf{x}) := \left(\sum_{i=1}^{n} I_X^{\{c_1\}}(x_i), \ldots, \sum_{i=1}^{n} I_X^{\{c_m\}}(x_i) \right)^T,$$
$$rH : M^n \to \mathbb{N}_0^m,\ \mathbf{x} \mapsto rH(\mathbf{x}) := \left(\frac{1}{n} \sum_{i=1}^{n} I_X^{\{c_1\}}(x_i), \ldots, \frac{1}{n} \sum_{i=1}^{n} I_X^{\{c_m\}}(x_i) \right)^T.$$

Der Wert der Abbildung $n : M \to \mathbb{N}_0$, $c_l \mapsto n_l := n(c_l) := aH(x_1, \ldots, x_n)_l$ heißt absolute Häufigkeit der Kategorie c_l und entsprechend heißt der Wert der Abbildung $h : M \to \mathbb{N}_0$, $c_l \mapsto h_l := h(c_l) := rH(x_1, \ldots, x_n)_l$ relative Häufigkeit der Kategorie c_l.

Beispiel 2.1: Passagierdaten vom Untergang der Titanic

Wir betrachten Daten von den 2201 Passagieren zum Untergang der Titanic. Es werden die Klasse, das Alter und Geschlecht sowie das Überleben der jeweiligen Person erfasst. Hier ein Ausschnitt der Datentabelle:

Datentabelle „Titanic“

Class	Age	Sex	Survived
1st	Adult	Male	Yes
1st	Adult	Male	Yes
1st	Adult	Male	Yes
1st	Adult	Male	Yes
⋮	⋮	⋮	⋮
Crew	Adult	Female	Yes
Crew	Adult	Female	No
Crew	Adult	Female	No
Crew	Adult	Female	No

Bei allen vier Merkmalen handelt es sich um qualitative Merkmale und wir können die Fälle auszählen:

Class

Kategorie	n_l	h_l
1st	325	$325/2201$
2nd	285	$285/2201$
3rd	706	$706/2201$
Crew	885	$885/2201$
gesamt	2201	1

Age

Kategorie	n_l	h_l
Adult	2092	$2092/2201$
Child	109	$109/2201$
gesamt	2201	1

Sex

Kategorie	n_l	h_l
Female	470	$^{470}/_{2201}$
Male	1731	$^{1731}/_{2201}$
gesamt	2201	1

Survived

Kategorie	n_l	h_l
Yes	1490	$^{1490}/_{2201}$
No	711	$^{711}/_{2201}$
gesamt	2201	1

Einen einfachen Datensatz können wir noch direkt abzählen. Doch schon beim Titanic-Datensatz mit 2201 Merkmalsträgern ist eine Abzählung offenbar zu aufwändig. Wir nutzen stattdessen R[1].

```
setwd("/Daten/Datensatz")
titanic<-read.table("Titanic.txt",sep="\t",head=TRUE)
attach(titanic)
head(titanic)
str(titanic)
```

In der ersten Zeile setzen wir das Arbeitsverzeichnis. Der Titanic-Datensatz wird aus der entsprechenden Textdatei eingelesen, die Spalten sind durch einen Tabulator getrennt und es existiert eine Kopfzeile. Die einzelnen Variablen werden durch die dritte Zeile zugänglich, d.h. das Merkmal Age kann nun entweder indirekt über „titanic$Age“ oder direkt angesprochen werden. Die vierte Zeile führt zu folgender Ausgabe.

```
  Class   Age  Sex Survived
1 First Adult Male      Yes
2 First Adult Male      Yes
3 First Adult Male      Yes
4 First Adult Male      Yes
5 First Adult Male      Yes
6 First Adult Male      Yes
```

Der str-Befehl liefert Informationen über die Datenstruktur eines Objekts.

```
'data.frame':   2201 obs. of  4 variables:
 $ Class    : Factor w/ 4 levels "Crew","First",...:
                          2 2 2 2 2 2 2 2 2 2 ...
 $ Age      : Factor w/ 2 levels "Adult","Child":
                          1 1 1 1 1 1 1 1 1 1 ...
 $ Sex      : Factor w/ 2 levels "Female","Male":
                          2 2 2 2 2 2 2 2 2 2 ...
 $ Survived : Factor w/ 2 levels "No","Yes":
                          2 2 2 2 2 2 2 2 2 2 ...
```

Um die Häufigkeiten eines Merkmals zu bestimmen, können wir den table-Befehl nutzen.

```
table(Class)
prop.table(table(Class))
```

[1] Ein interessantes Buch zur Anwendung von R ist [3]

```
Class
  Crew   First  Second   Third
   885     325     285     706

Class
      Crew       First      Second       Third
0.4020900  0.1476602  0.1294866  0.3207633
```

3. Merkmale mit Kardinalskala

Warum?

Sehr oft sind die Wertemengen von Eigenschaften eine Teilmenge der Menge der reellen Zahlen. Deswegen wird zur vereinfachten Weiterverarbeitung der Werte die Menge $\mathbb{R}$ als Wertemenge angenommen ohne etwas dabei zu verlieren. Daten auf Basis der Menge $\mathbb{R}$ lassen sich durch Informationsverdichtung durch einzelne Kenngrößen beschreiben. Das ist Gegenstand der deskriptiven Statistik und hat den Vorteil, dass ein schneller Überblick über eine große Menge an Daten gewonnen werden kann. Dies wird häufig durch eine visuelle Darstellung unterstützt. Bei mehreren solchen Eigenschaften kann zudem der Zusammenhang zwischen diesen untersucht werden.

Meist treten in den Anwendungen mehrere Merkmale gleichzeitig auf. Jedoch ist es sinnvoll, Merkmale auch einzeln zu untersuchen. Das erfolgt nicht zuletzt deshalb, weil es für eine Vielzahl von Verfahren in der angewandten Statistik Annahmen hinsichtlich der Beschaffenheit einzelner Merkmale - wie etwa Verteilungsannahmen - gibt. Weiter können so Untersuchungen hinsichtlich Ausreißern (z.B. Messfehler oder Eingabefehler) durchgeführt werden.

3.1. Kenngrößen für ein Merkmal

Sei $\mathbf{x} = (x_1, \ldots, x_n)^T$ eine Stichprobe eines kardinalen Merkmals X. Die drei zumeist benutzten Kenngrößen zur Beschreibung kardinaler Merkmale sind der empirische Mittelwert $\bar{x}$, die empirische Varianz s^2 und die empirische Standardabweichung s,

$$\bar{x} : \mathbb{R}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto \bar{x} := \bar{x}(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{3.1}$$

$$s^2 : \mathbb{R}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto s^2 := s^2(\mathbf{x}) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{3.2}$$

$$s : \mathbb{R}^n \to \mathbb{R}, \qquad \mathbf{x} \mapsto s := s(\mathbf{x}) := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{3.3}$$

Die drei Funktionen stellen Schätzfunktionen dar. Der durch Einsetzen der Stichprobenelemente erhaltene Wert ist der Schätzwert.

Satz und Definition 3.1: Schätzfunktion und Schätzwert

Sei $(x_1, \ldots, x_n)^T \in M^n$ eine Stichprobe, wobei jedes x_i aus einer u.i.v. Zufallsvaria-

blen mit parametrisierter Dichtefunktion stamme. Sei Θ der entsprechende Parameterraum und $\theta \in \Theta$ der wahre, aber unbekannte Parameter.

- $\gamma : \Theta \to \Gamma$ sei eine Abbildung vom Parameterraum in Γ,
- eine Zufallsvariable $g : M^n \to \Gamma$ heißt Schätzfunktion für $\gamma(\theta)$,
- $g(x_1, \ldots, x_n)$ heißt Schätzwert.

Sei G eine Menge von Schätzfunktionen und $g \in G$. Dann heißt g

- erwartungstreu, wenn $\mathbb{E}_\theta[g] = \gamma(\theta)$,
- effizient, wenn $\mathbb{V}_\theta[g] \leq \mathbb{V}_\theta[\tilde{g}]$ für alle $\tilde{g} \in G$,
- konsistent, wenn $\mathbb{V}_\theta[g] \overset{n\to\infty}{\longrightarrow} 0$.

Die drei Kenngrößen sind demnach erhaltene Schätzwerte der angegebenen Schätzfunktionen. Der empirische Mittelwert ist ein Lageparameter, während die empirische Varianz und die Standardabweichung Streuparameter sind.

Beispiel 3.2

Sei $\mathbf{x} = (5, 10, 4, 13, 8)^T$. Dann gilt

$$\begin{aligned} \bar{x} &= \frac{1}{5}(5 + 10 + 4 + 13 + 8) = 8, \\ s^2 &= \frac{1}{4}\left((5-8)^2 + (10-8)^2 + (4-8)^2 + (13-8)^2 + (8-8)^2\right) = \frac{54}{4} = 13.5, \\ s &= \sqrt{13.5} = 3.67. \end{aligned}$$

Beispiel 3.3: Wisconsin Brustkrebs-Daten

Bei einer Fine-needle aspiration biopsy (FNA)[a] werden durch eine feine Nadel dem Körper Zellen entnommen und unter dem Mikroskop untersucht. Im vorliegenden Datensatz[b] wurden jedem der 569 Patienten 10-40 Zellen entnommen und deren Zellkerne untersucht. Dabei wurden 10 Merkmale erfasst: radius, texture, peri, area, smooth, comp, scav, ncav, symt und fracd. Über alle Kerne wurden der empirische Mittelwert, der größte Wert und die empirische Standardabweichung gebildet, so dass für jeden Patienten diese 30 Merkmale zur Verfügung stehen. Darüber hinaus wird erfasst, ob das Gewebe bös- oder gutartig ist. Für das Merkmal radius.mv erhalten wir beispielsweise

$$\begin{aligned} \bar{x} &= 14.13, \\ s^2 &= 12.42, \\ s &= 3.52. \end{aligned}$$

[a] http://en.wikipedia.org/wiki/Fine-needle_aspiration
[b] siehe https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Äquivarianz

Bei Merkmalen spielt möglicherweise die Messskala eine wichtige Rolle. So können bei einer Messung von Temperaturen die Messwerte in Grad Celsius, Kelvin oder Fahrenheit angegeben werden. Ein Lage- oder Streuparameter sollte einen Wert unabhängig von der gewählten Messskala liefern. Schätzfunktionen für Lageparameter $m : \mathbb{R}^n \to \mathbb{R}$ werden als äquivariant bezeichnet, wenn für beliebige $a, b \in \mathbb{R}$

$$m(a \cdot x_1 + b, \ldots, a \cdot x_n + b) \quad = \quad a \cdot m(x_1, \ldots, x_n) + b$$

gilt, während Schätzfunktionen für Streuparameter $s : \mathbb{R}^n \to \mathbb{R}$ als äquivariant bezeichnet werden, wenn für beliebige $a, b \in \mathbb{R}$

$$s(a \cdot x_1 + b, \ldots, a \cdot x_n + b) \quad = \quad |a| \cdot s(x_1, \ldots, x_n)$$

gilt.

Satz 3.4

Die Schätzfunktionen für den empirischen Mittelwert und die empirische Standardabweichung sind äquivariant.

Beweis.

$$\frac{1}{n}\sum_{i=1}^{n}(a \cdot x_i + b) \quad = \quad a \cdot \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n} \cdot n \cdot b = a \cdot \frac{1}{n}\sum_{i=1}^{n} x_i + b$$

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a \cdot x_i + b - (a \cdot \bar{x} + b))^2} \quad = \quad |a| \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

□

Robustheit

Manchmal finden sich bei Messungen Werte, die sich deutlich von den anderen unterscheiden. Sie werden Ausreißer genannt. Kenngrößen sollten möglichst wenig durch einzelne Ausreißer beeinflusst werden. Solche Kenngrößen werden als robust bezeichnet. Um den Einfluss einzelner Beobachtungen bestimmen zu können, gibt es die Möglichkeit der Betrachtung des so genannten Sensitivitätsdiagramms. Dabei wird für jede Realisierung x_j eine skalierte Differenz SC zwischen der Kenngröße m_n der gesamten Stichprobe und der Kenngröße $m_{n(j)}$ der um den einen Merkmalswert reduzierten Stichprobe ermittelt,

$$SC(x_j, m) := k \cdot |m_n - m_{n(j)}|. \tag{3.4}$$

Die Skalierung mit k hängt von der jeweils untersuchten Kenngröße ab. Betrachten wir den Graphen

$$\{(x_j, SC(x_j, m));\ j = 1, \ldots, n\} \text{ bzw. } \{(j, SC(x_j, m));\ j = 1, \ldots, n\},$$

so werden grundsätzlich besonders große Werte als Ausreißer identifiziert und interpretiert. Mit $\mathbf{1} := (1, \ldots, 1)^T$ werde der Vektor (mit entsprechender Dimension) bezeichnet, dessen jede Komponente 1 ist, mit $\tilde{\mathbf{1}} := (1, \ldots, 1, 0, 1, \ldots, 1)^T$ der Vektor, dessen j-te Komponente 0, alle anderen Komponenten 1 sind. Es ist $\bar{x} = \frac{1}{n}\mathbf{1}^T\mathbf{x}$. Sei

$$\bar{x}_j := \frac{1}{n-1}\sum_{\substack{i=1\\ i\neq j}}^{n} x_i = \frac{1}{n-1}\mathbf{x}^T\tilde{\mathbf{1}}. \tag{3.5}$$

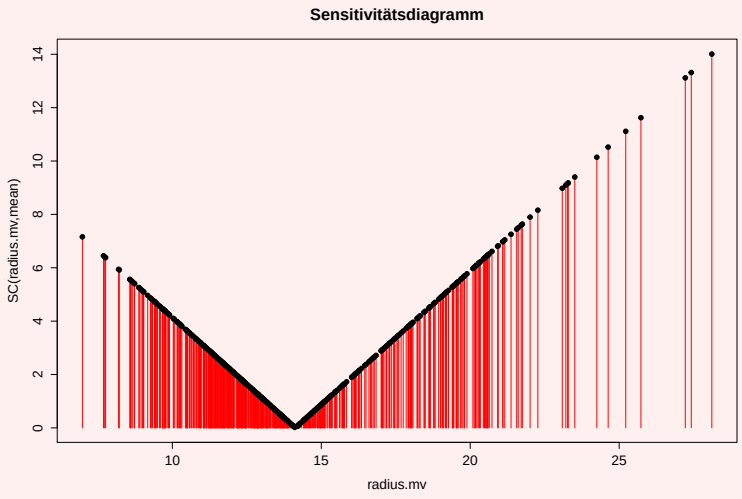
Mit $k = n$ erhalten wir zunächst für den empirischen Mittelwert

$$\begin{aligned}
SC(x_j, \bar{x}) &= n \cdot |\bar{x} - \bar{x}_j| \\
&= n \cdot \left|\frac{1}{n}\mathbf{x}^T\mathbf{1} - \frac{1}{n-1}\mathbf{x}^T\tilde{\mathbf{1}}\right| \\
&= \left|\mathbf{x}^T(\mathbf{1} - \frac{n}{n-1}\tilde{\mathbf{1}})\right| \\
&= \left|x_j + \mathbf{x}^T(\tilde{\mathbf{1}} - \frac{n}{n-1}\tilde{\mathbf{1}})\right| \\
&= \left|x_j - \frac{1}{n-1}\mathbf{x}^T\tilde{\mathbf{1}}\right| \\
&= |x_j - \bar{x}_j| \\
&= \left|\frac{n}{n-1}(x_j - \bar{x})\right|.
\end{aligned}$$

SC ist damit ein Maß für die absolute Abweichung von x_j zum empirischen Mittelwert der anderen Realisierungen.

Beispiel 3.5

Für das Merkaml radius.mv und den empirischen Mittelwert erhalten wir folgendes Sensitivitätsdiagramm:



Es ist linear in x_j und es ergeben sich Werte nahe 0 im Bereich der empirischen Mittelwerte (um 14.1).

Wir untersuchen noch die skalierte Differenz für die empirische Varianz. Zur einfacheren Berechnung führen wir noch zwei Bezeichnungen ein. Mit

$$\mathbf{x} - \bar{x} \cdot \mathbf{1} = \mathbf{x} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{x} = \underbrace{I - \frac{1}{n}\mathbf{1}\mathbf{1}^T}_{H}\mathbf{x}$$

sei $H := I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Entsprechend sei $\tilde{H} := \tilde{I} - \frac{1}{n-1}\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T$, wobei $\tilde{I} = I - \mathbf{e}_j\mathbf{e}_j^T$ eine modifizierte Einheitsmatrix sei, deren j,j-te Komponente 0 ist. Damit lässt sich die empirische Varianz schreiben als

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\mathbf{x}^T H^T H \mathbf{x} \overset{(3.18)}{=} \frac{1}{n-1}\mathbf{x}^T H \mathbf{x}.$$

Analog zu (3.5) ist $s_j^2 = \frac{1}{n-2}\mathbf{x}^T\tilde{H}\mathbf{x}$. Setzen wir $k = n-1$, so bekommen wir

$$\begin{aligned}
SC(x_j, s^2) &= (n-1)\cdot|s^2 - s_j^2| \\
&= \left|\mathbf{x}^T H\mathbf{x} - \frac{n-1}{n-2}\mathbf{x}^T\tilde{H}\mathbf{x}\right| \\
&= \left|\mathbf{x}^T\left(H - \frac{n-1}{n-2}\tilde{H}\right)\mathbf{x}\right| \\
&= \left|\mathbf{x}^T\left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) - \frac{n-1}{n-2}\left(\tilde{I} - \frac{1}{n-1}\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\right)\right)\mathbf{x}\right| \\
&= \left|\mathbf{x}^T\mathbf{x} - \frac{1}{n}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{x} - \frac{n-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{1}{n-2}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \\
&= \left|\frac{-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + x_j^2 - \frac{1}{n}\underbrace{(\mathbf{x}^T\tilde{\mathbf{1}} + \mathbf{x}^T\mathbf{e}_j)(\mathbf{x}^T\tilde{\mathbf{1}} + \mathbf{x}^T\mathbf{e}_j)^T}_{\mathbf{x}^T\tilde{\mathbf{1}}\mathbf{e}_j^T\mathbf{x} + \mathbf{x}^T\mathbf{e}_j\mathbf{e}_j^T\mathbf{x} + \mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x} + \mathbf{x}^T\mathbf{e}_j\tilde{\mathbf{1}}^T\mathbf{x}} + \frac{1}{n-2}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \\
&\overset{x_j = \mathbf{e}_j^T\mathbf{x}}{=} \left|\frac{n-1}{n}x_j^2 - \frac{2}{n}\mathbf{x}^T\tilde{\mathbf{1}}\cdot x_j + \frac{-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{2}{n(n-2)}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \qquad (3.6) \\
&= \left|\frac{n}{n-1}(x_j - \bar{x})^2 - s_j^2\right|
\end{aligned}$$

Wir untersuchen, für welchen Wert von x_j die skalierte Differenz gleich 0 ist. Da SC quadratisch in x_j ist, erhalten wir allgemein die beiden folgenden Lösungen:

$$\begin{aligned}
SC(x_j, s^2) &= 0 \\
\Leftrightarrow x_j &= \frac{n}{2(n-1)}\cdot\left(\frac{2}{n}\mathbf{x}^T\tilde{\mathbf{1}} \pm \left(\frac{4}{n^2}(\mathbf{x}^T\tilde{\mathbf{1}})^2\right.\right. \\
&\quad \left.\left. -\frac{4(n-1)}{n}\left(-\frac{1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{2}{n(n-2)}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right)\right)^{\frac{1}{2}}\right) \\
&= \frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1} \pm \left((\mathbf{x}^T\tilde{\mathbf{1}})^2\cdot\frac{-n}{(n-1)^2(n-2)} + \mathbf{x}^T\tilde{I}\mathbf{x}\cdot\frac{n}{(n-1)(n-2)}\right)^{\frac{1}{2}} \\
&= \frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1} \pm \sqrt{\frac{n}{n-1}}\cdot s_j. \qquad (3.7)
\end{aligned}$$

Bei der Untersuchung des empirischen Mittelwerts hat sich ergeben, dass Werte von x_j nahe des empirischen Mittelwerts $\bar{x}_j$ eine kleine Differenz ergeben. Setzen wir $x_j = \frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1}$ an, ergibt sich als Differenz

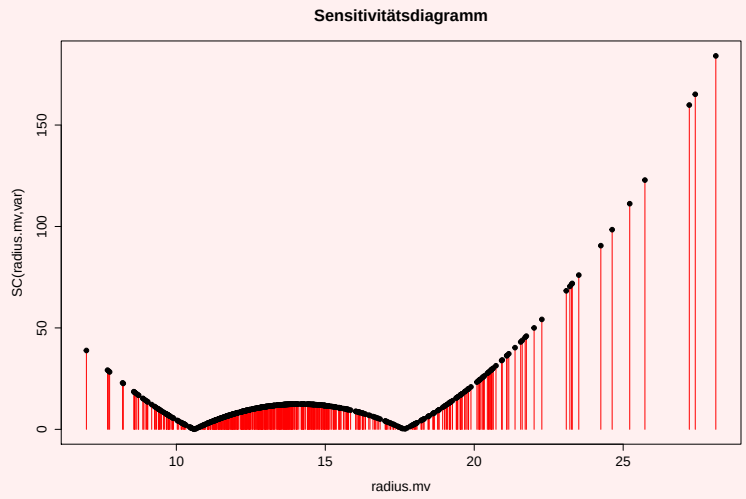
$$\begin{aligned} SC\left(\frac{\mathbf{x}^T\tilde{\mathbf{1}}}{n-1}, s^2\right) &= \left|\frac{(\mathbf{x}^T\tilde{\mathbf{1}})^2}{n(n-1)} - \frac{2(\mathbf{x}^T\tilde{\mathbf{1}})^2}{n(n-1)} + \frac{-1}{n-2}\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{2}{n(n-2)}\mathbf{x}^T\tilde{\mathbf{1}}\tilde{\mathbf{1}}^T\mathbf{x}\right| \\ &= \left|\frac{1}{n-2}\left(-\mathbf{x}^T\tilde{I}\mathbf{x} + \frac{1}{n-1}(\mathbf{x}^T\tilde{\mathbf{1}})^2\right)\right| \\ &= s_j^2. \end{aligned} \tag{3.8}$$

Damit verringert sich die Varianz in diesem Fall zu

$$s^2 = s_j^2 - \frac{1}{n-1}s_j^2 = \frac{n-2}{n-1}s_j^2.$$

Innerhalb des Intervalls $[\bar{x}_j - s_j, \bar{x}_j + s_j]$ verringert sich die Varianz, danach vergrößert sie sich.

Beispiel 3.6




```
x1=radius.mv
mean(x1)
var(x1)
sd(x1)
u=c()
for(i in 1:l)
{   u=append(u,l*abs(mean(v1)-mean(v1[-i])))
}
plot(x1,u,pch=19,type="h",col=2)
```

Im letzten Beispiel ändert sich die empirische Varianz durch Weglassen des größten beobachteten Merkmalswertes relativ um 2.7 Prozent. Um extreme Werte bei Kenngrößen

zu berücksichtigen, können die größten oder kleinsten α-Prozent der Werte bei der Berechnung eines Lageparameters weggelassen werden. Dazu müssen wir zunächst die Merkmalswerte einer Stichprobe sortieren. Wir betrachten die Permutation

$$T : \mathbb{R}^n \to \mathbb{R}^n, \quad \mathbf{x} = (x_1, \ldots, x_n)^T \mapsto \left(x_{(1)}, \ldots, x_{(n)}\right)^T := T(x_1, \ldots, x_n)$$
$$\text{mit } x_{(i)} \leq x_{(j)} \text{ für alle } i < j.$$

T heißt Ordnungsstatistik auf $\mathbb{R}^n$. Die Abbildung

$$T_i : \mathbb{R}^n \to \mathbb{R}, \ \mathbf{x} = (x_1, \ldots, x_n) \mapsto T_i(x_1, \ldots, x_n) := T(x_1, \ldots, x_n)_i = x_{(i)}$$

heißt i-te Ordnungsstatistik auf $\mathbb{R}^n$. Das α-Quantil $c_{(\alpha)}$ wird dann für $0 \leq \alpha \leq 1$ durch das Element mit dem kumulierten Gewicht von mindestens $n \cdot \alpha$ bestimmt durch

$$c_{(\alpha)} : \mathbb{R}^n \to \mathbb{R}, \ c_{(\alpha)} := c_{(\alpha)}(\mathbf{x}) := \begin{cases} T_{\lceil n\cdot\alpha \rceil}(\mathbf{x}) = x_{(\lceil n\cdot\alpha \rceil)}, & 0 < \alpha \leq 1, \\ x_{(1)}, & \alpha = 0. \end{cases} \tag{3.9}$$

Das getrimmte Mittel ist nun für $\frac{1}{n} < \alpha < \frac{1}{2}$ definiert durch

$$\bar{x}_\alpha : \mathbb{R}^n \to \mathbb{R}, \ \mathbf{x} \mapsto \bar{x}_\alpha := \bar{x}_\alpha(\mathbf{x}) := \frac{1}{n - 2\lfloor \alpha \cdot n \rfloor + 1} \sum_{i=\lfloor \alpha\cdot n \rfloor}^{\lceil (1-\alpha)\cdot n \rceil} x_{(i)} \tag{3.10}$$

Ein weiterer wichtiger robuster Lageparameter ist der Median $\tilde{x}$,

$$\text{med} : \mathbb{R}^n \to \mathbb{R}, \ \mathbf{x} \mapsto \tilde{x} := \text{med}(\mathbf{x}) := c_{(0.5)}. \tag{3.11}$$

Der Median ist demnach das 50-Prozent Quantil. Auf Basis der Quantile lassen sich auch weitere Streuparameter bestimmen wie etwa der Interquartilsabstand IQR oder der Median der absoluten Abweichung vom Median MAD,

$$IQR : \mathbb{R}^n \to \mathbb{R}, \ \mathbf{x} \mapsto \text{IQR} := IQR(\mathbf{x}) := c_{(0.75)} - c_{(0.25)}, \tag{3.12}$$
$$MAD : \mathbb{R}^n \to \mathbb{R}, \ \mathbf{x} \mapsto \text{MAD} := MAD(\mathbf{x}) := \text{med}(|x_1 - \tilde{x}|, \ldots, |x_n - \tilde{x}|). \tag{3.13}$$

Der untere Whisker errechnet sich über $W_u := c_{(0.25)} - 1.5 \cdot \text{IQR}$, der obere Whisker über $W_o := c_{(0.75)} + 1.5 \cdot \text{IQR}$. Mit Hilfe der Ausreißer-Regel von Tukey lassen sich ebenfalls mögliche Ausreißer entdecken. Dabei werden solche x_i gesucht, für die

$$x_i < W_u \text{ bzw. } x_i > W_o$$

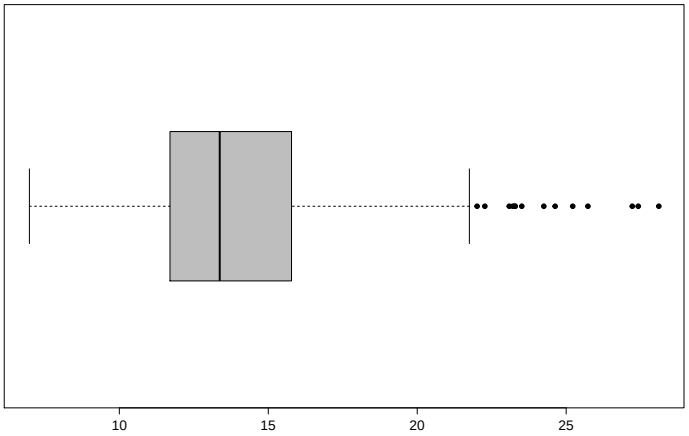
gilt. Mit Hilfe des Medians, der Whisker und der für die Whisker verwendeten Quantile lässt sich ein Merkmal übersichtlich graphisch darstellen.

fivenum (x1)

Boxplot

Ein Boxplot hilft bei der Entdeckung potentieller Ausreißer. Dabei wird nicht jeder beobachtete Merkmalswert einzeln gezeichnet, sondern die Werte werden zu Boxen zusammengefasst und der Median wird gekennzeichnet. In der Abbildung ist das als Querstrich zu erkennen. Sämtliche Werte, die sich zwischen oberem und unterem Quartil $(c_{(0.75)}, c_{(0.25)})$ befinden, werden in der inneren Box um den Median zusammengefasst. Die schwarzen

Linien umfassen alle Werte vom oberen Quartil bis zum oberen Whisker bzw. vom unteren Quartil zum unteren Whisker. Diejenigen Werte, die noch nicht erfasst sind, werden außerhalb einzeln dargestellt. Sie lassen sich als Ausreißer interpretieren. Die obere Graphik zeigt einen Boxplot für das Merkmal aus dem Beispiel 3.3 und deutet mehrere mögliche Ausreißer hin.




```
boxplot(radius.mv, horizontal=TRUE, col="gray",pch=19)
```

Histogramm

Wir unterteilen den Bereich zwischen dem kleinsten und dem größten Wert der Stichprobe in m nicht notwendigerweise gleichbreite Intervalle $[b_i, b_{i+1}[$, $i = 1, \ldots, m$, mit $b_1 \leq \min\{x_1, \ldots, x_n\}$ und $b_{m+1} > \max\{x_1, \ldots, x_n\}$. Jedes Intervall $[b_i, b_{i+1}[$ enthält eine bestimmte Anzahl n_i an Merkmalswerten. Um die Anzahl auszuzählen, benötigen wir für eine beliebige Menge L eine Indikatorfunktion gemäß (2.2). Es ist $n_i = \sum_{j=1}^{n} I^{[b_i, b_{i+1}[}(x_j)$.
In einem Histogramm werden die Intervalle auf einer Achse angetragen. Jedem Intervall $[b_i, b_{i+1}[$ wird entsprechend der relativen Häufigkeit $h_i = \frac{n_i}{n}$ der enthaltenen Merkmalswerte eine Höhe k_i so zugeordnet, dass für die Fläche $(b_{i+1} - b_i) \cdot k_i = h_i$ gilt. Fassen wir ein Histogramm als Graph der (reellen) Funktion

$$h : \mathbb{R} \to \mathbb{R},\ b \mapsto h(b) := \begin{cases} \frac{h_i}{b_{i+1} - b_i}, & b \in [b_i, b_{i+1}[, \\ 0, & \text{sonst,} \end{cases}$$

auf, so gilt

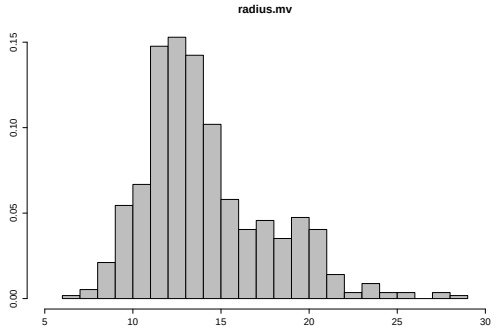
$$\int_{-\infty}^{\infty} h(x)dx = \sum_{i=1}^{m} (b_{i+1} - b_i) \cdot \frac{h_i}{b_{i+1} - b_i} = \sum_{i=1}^{m} \frac{n_i}{n} = 1.$$

h ist stets nicht-negativ und damit wird h zu einer Dichtefunktion. Wir können ein Histogramm als Schätzer der Dichte interpretieren.

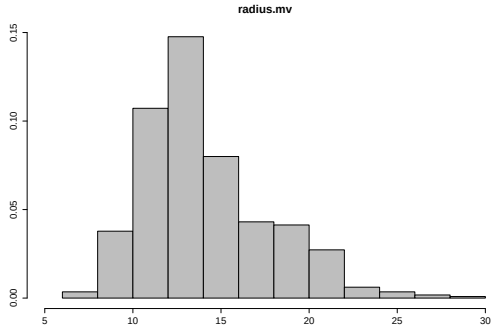
Es gibt die Faustregel nach Freedman und Diaconis, dass die Breite der Intervalle (Binbreite) eines Histogramms gleich $\frac{2 \cdot \text{IQR}}{\sqrt[3]{n}}$ sein soll. Dies ist aber lediglich eine Richtschnur.

```
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(5,30), breaks="FD")
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(5,30))
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(5,30), breaks=seq(4,32,4))
```

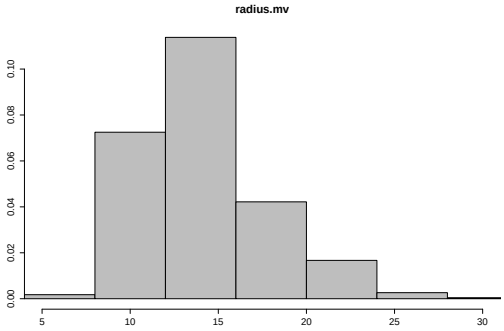
Die folgende Abbildung zeigt ein Histogramm für das Merkmal radius.mv aus Beispiel 3.3.



Ein weiterer wichtiger Aspekt ist die Festlegung der linken Intervallgrenze b_1. Standardmäßig wird der kleinste Merkmalswert genommen. Dies kann jedoch zu Fehlinterpretationen führen. Zur Exploration sollten mehrere Intervallzahlen und verschiedene Startwerte für b_1 gewählt werden.



Es erweist sich als Vorteil, wenn die Intervallbreite identisch gewählt wird. Dennoch ist bei der Interpretation und der Wahl der Intervallbreite Vorsicht geboten. Oft wird ein Ergebnis suggeriert, das lediglich auf einer unglücklichen Wahl der Intervallbreite zurückzuführen ist. Dies tritt insbesondere bei der Untersuchung von Lücken in Erscheinung.



Eine Merkregel ist, dass eine große Zahl an Intervallen eine gleichmäßige Verteilung erzeugt, bei einer geringen Intervallzahl Details verschluckt werden.
Eine weitere Interpretationsmöglichkeit besteht in der Beurteilung der Schiefe der Verteilung. Vergleichen wir den empirischen Mittelwert mit dem Median, so können wir Aussagen bezüglich der Schiefe der empirischen Verteilung treffen, wie folgende Übersicht zeigt.

Schiefe der Verteilung

Vergleich	Interpretation
$\bar{x} = \tilde{x}$	Symmetrische empirische Verteilung
$\bar{x} > \tilde{x}$	Rechtsschiefe empirische Verteilung
$\bar{x} < \tilde{x}$	Linksschiefe empirische Verteilung

Hier liegt eine rechtsschiefe empirische Verteilung vor, was sich durch Rechnung belegen lässt: $\bar{x} = 14.13 > 13.37 = \tilde{x}$.

Kern-Dichteschätzung

Wir haben für das Histogramm die Eigenschaft einer Dichte nachgewiesen. Allerdings sind für die Höhe des Histogramms lediglich Ausprägungen im jeweiligen Intervall entscheidend. Um das zu vermeiden, kann ein so genanntes gleitendes Histogramm verwendet werden. Für ein beliebiges $b \in \mathbb{R}$ approximieren wir die Dichte mittels $\tilde{h} : \mathbb{R} \to \mathbb{R}$,

$$\tilde{h}(b) := \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} I^{[b-h,b+h[}(x_i)}{2h}.$$

Seien $x_1, \ldots, x_n$ die Merkmalswerte. Wir notieren die Funktion $\tilde{h}$ in der Form

$$\tilde{h}(b) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{b - x_i}{h}\right), \quad K\left(\frac{b - x_i}{h}\right) := \begin{cases} \frac{1}{2}, & x_i - h < b \leq x_i + h, \\ 0, & \text{sonst.} \end{cases}$$

Dann gilt

$$\int_{-\infty}^{\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{b - x_i}{h}\right) db = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K(x) \cdot h dx = \frac{1}{nh} \cdot nh = 1.$$

Die bei der Integration durchgeführte Substitution $x := \frac{b-x_i}{h}$ liefert eine Funktion

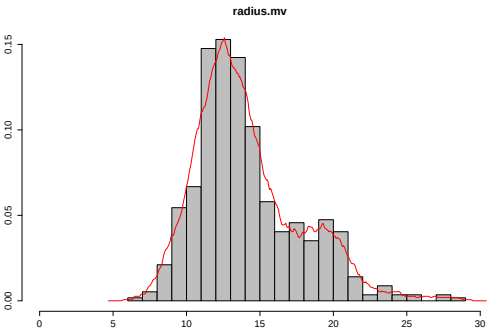
$$K(x) = \begin{cases} \frac{1}{2}, & -1 < x \leq 1, \\ 0, & \text{sonst}, \end{cases}$$

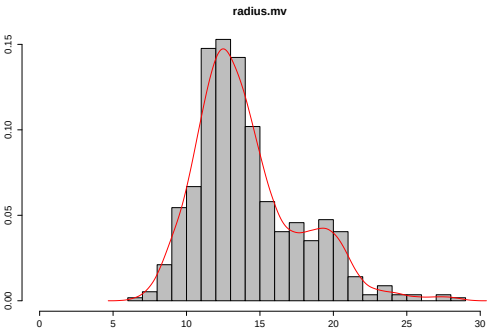
die zwischen -1 und 1 nur positive Werte annimmt und den Integralwert 1 hat. Wir nennen eine solche Funktion eine Kernfunktion und die Dichte $\tilde{h}(b)$ einen Kerndichteschätzer. Die hier benutzte Rechteckskernfunktion führt zu einer nichtstetigen Dichtefunktion.

Durch eine andere Wahl der Kernfunktion können stetige Dichten erzeugt werden. Hierzu müssen die Kernfunktionen stetig sein. Die additive Überlagerung führt dann zu einer stetigen Funktion. Als Beispiel sei die Bisquare-Kernfunktion

$$K(x) := \begin{cases} \frac{15}{16}(1-x^2)^2, & -1 < x \leq 1, \\ 0, & \text{sonst}, \end{cases}$$

vorgestellt. In der Abbildung sind die Rechtecks- und Bisquare-Kernfunktion zu sehen. Betrachten wir wieder die Daten aus Beispiel 3.3 und schätzen die Dichte von radius.mv mit Hilfe der Rechtecks- bzw. der Bisquare-Kernfunktion. Problematisch bei der Kerndichteschätzung ist ähnlich zu den Histogrammen die Wahl der Bandbreite h. Es gibt verschiedene Ansätze zur optimalen Bandbreitenwahl, jedoch sind diese jeweils nicht unproblematisch. Je größer die Bandbreite, desto glatter wird die geschätzte Dichtefunktion.






```
hist(radius.mv, freq=FALSE, right=FALSE, xlab="", ylab="",
     main="radius.mv", col="gray", xlim=c(0,30), breaks="FD")
lines(density(radius.mv, kernel=c("rectangular")), col=2)
lines(density(radius.mv, kernel=c("biweight")), col=2)
```

3.2. Kenngrößen für mehrere Merkmale

Die Lageparameter lassen sich grundsätzlich auf den zwei- oder mehrdimensionalen Fall übertragen[1]. Seien k Merkmale $X_1, \ldots, X_k$ und eine Datenmatrix $X \in \mathbb{R}^{n,k}$ gegeben. Dann werden der Schwerpunkt $\bar{\mathbf{x}}$ und der Medianpunkt $\tilde{\mathbf{x}}$ definiert als

$$\bar{\mathbf{x}} : \mathbb{R}^{n,k} \to \mathbb{R}^k, \qquad X \mapsto \bar{\mathbf{x}} := \bar{\mathbf{x}}(X) := \frac{1}{n}\mathbf{1}^T X = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_k), \tag{3.14}$$

$$\tilde{\mathbf{x}} : \mathbb{R}^{n,k} \to \mathbb{R}^k, \qquad X \mapsto \tilde{\mathbf{x}} := \tilde{\mathbf{x}}(X) := (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k). \tag{3.15}$$

[1] vgl. auch [6]

Für weitere Betrachtungen sind die Vektoren als Zeilenvektoren geschrieben. Nun lässt sich die Datenmatrix X mit Hilfe des Schwerpunkts zentrieren,

$$\begin{pmatrix} x_{11} - \bar{\mathbf{x}}_1 & \dots & x_{1k} - \bar{\mathbf{x}}_k \\ x_{21} - \bar{\mathbf{x}}_1 & \dots & x_{2k} - \bar{\mathbf{x}}_k \\ \vdots & & \vdots \\ x_{n1} - \bar{\mathbf{x}}_1 & \dots & x_{nk} - \bar{\mathbf{x}}_k \end{pmatrix} = X - \mathbf{1}\bar{\mathbf{x}} = X - \mathbf{1}\frac{1}{n}\mathbf{1}^T X = \underbrace{\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)}_{H} X.$$

$H \in \mathbb{R}^{n,n}$ wird Zentrierungsmatrix genannt und ist identisch zur Matrix H nach Beispiel 3.5.

Definition 3.7: Orthogonalprojektion

Eine lineare Abbildung $\mathbb{R}^n \to \mathbb{R}^n$, die durch die Matrix $M \in \mathbb{R}^{n,n}$ repräsentiert wird, heißt Orthogonalprojektion auf den Unterraum $U \subseteq \mathbb{R}^n$, falls

$$M\mathbf{x} \in U \qquad \text{für alle } \mathbf{x} \in \mathbb{R}^n, \tag{3.16}$$

$$(\mathbf{x} - M\mathbf{x})^T\mathbf{u} = 0 \qquad \text{für alle } \mathbf{u} \in U,\ \mathbf{x} \in \mathbb{R}^n. \tag{3.17}$$

Satz 3.8

Die Zentrierungsmatrix H ist eine Orthogonalprojektion auf den Unterraum

$$U = \{\mathbf{x} \in \mathbb{R}^n;\ \mathbf{1}^T\mathbf{x} = 0\}.$$

Beweis.

Zunächst gilt

$$\begin{aligned} H^T &= (I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T)^T = (I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T) = H, \\ H^T H &= H^2 = (I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T)(I - \tfrac{1}{n}\mathbf{1}\mathbf{1}^T) = I - \tfrac{2}{n}\mathbf{1}\mathbf{1}^T + \tfrac{1}{n^2}n\mathbf{1}\mathbf{1}^T = H. \end{aligned}$$

Damit ist H erstens symmetrisch. Zweitens ist H idempotent und so eine Projektion. Um den Unterraum U bestimmen zu können, untersuchen wir die Eigenwerte und Eigenräume von H.

Wir behaupten, dass H die Eigenwerte 0 und 1 besitzt. Das gilt wegen $(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{1} = \mathbf{0}$ und $(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(1, -1, 0, \dots, 0)^T = (1, -1, 0, \dots, 0)^T$. Wir bestimmen für $\lambda = 0$ bzw. $\lambda = 1$ die Lösungen der Gleichungen $H\mathbf{v} = \lambda\mathbf{v}$. Zunächst für $\lambda = 0$:

$$H\mathbf{v} = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{v} = \mathbf{v} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{v} \stackrel{!}{=} \mathbf{0} \Leftrightarrow v_i = \bar{\mathbf{v}} \text{ für alle } i = 1, \dots, n, v_i \neq 0.$$

Damit ist $\lambda = 0$ Eigenwert von H zum eindimensionalen Eigenraum $E_0 = \{\mathbf{v} \in$

$\mathbb{R}^n;\ v_i = v_j$ für alle $i, j\}$. Weiter ist für $\lambda = 1$:

$$H\mathbf{v} = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{v} = \mathbf{v} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{v} \stackrel{!}{=} \mathbf{v} \Leftrightarrow \mathbf{1}\mathbf{1}^T\mathbf{v} = \mathbf{0} \text{ und } \mathbf{v} \neq \mathbf{0}.$$

Damit ist $\lambda = 1$ Eigenwert von H zum $n-1$-dimensionalen Eigenraum $E_1 = \{\mathbf{v} \in \mathbb{R}^n;\ \mathbf{1}^T\mathbf{v} = 0\}$.

Da zudem für beliebiges $\mathbf{x} \in \mathbb{R}^n$ und $\mathbf{v} \in E_1$ dann $(\mathbf{x} - H\mathbf{x})^T\mathbf{v} = \left(\frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{x}\right)^T\mathbf{v} = \frac{1}{n}\mathbf{x}^T\mathbf{1}\mathbf{1}^T\mathbf{v} = 0$ ist, wird H zu einer Orthogonalprojektion.

□

Mit den Überlegungen lässt sich die empirische Varianz eines Merkmals schreiben als

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}||H\mathbf{x}||^2 = \frac{1}{n-1}\mathbf{x}^T H\mathbf{x}. \tag{3.18}$$

Dabei ist $||.||$ die euklidische Norm, d.h. für $\mathbf{x} \in \mathbb{R}^n$ ist

$$||\mathbf{x}|| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{x_1^2 + \ldots + x_n^2}.$$

Assoziationsmaße für zwei Merkmale

Sind zwei Merkmale gegeben, kann die empirische Varianz jeweils als Streuparameter herangezogen werden. Darüber hinaus interessiert aber auch ein gemeinsames Streuverhalten und damit verbunden auch ein gemeinsames Verhalten der Merkmale. Wir suchen eine Kenngröße, die uns etwas über den Zusammenhang der beiden Merkmale aussagen kann.

Wir bestimmen ein Assoziationsmaß für zwei quantitative Merkmale X_1 und X_2 mit jeweils n Realisierungen. Für ein quantitatives Merkmal lassen sich jeweils die empirischen Varianzen $s^2_{X_1}$ bzw. $s^2_{X_2}$ bestimmen. Seien $\bar{x}_1$ und $\bar{x}_2$ die beiden empirischen Mittelwerte und x_{ij} mit $i \in \{1, \ldots, n\}$ und $j \in \{1, 2\}$ die Realisierungen. Für das Produkt der Abstände zu den Mittelwerten gilt

$$\begin{aligned}
(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) > 0 \quad &\Leftrightarrow \quad \begin{cases} \text{(I)} & (x_{i1} - \bar{x}_1) > 0 \wedge (x_{i2} - \bar{x}_2) > 0 \text{ oder} \\ \text{(III)} & (x_{i1} - \bar{x}_1) < 0 \wedge (x_{i2} - \bar{x}_2) < 0 \end{cases} \\
(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) < 0 \quad &\Leftrightarrow \quad \begin{cases} \text{(II)} & (x_{i1} - \bar{x}_1) > 0 \wedge (x_{i2} - \bar{x}_2) < 0 \text{ oder} \\ \text{(IV)} & (x_{i1} - \bar{x}_1) < 0 \wedge (x_{i2} - \bar{x}_2) > 0 \end{cases}
\end{aligned} \tag{3.19}$$

Dieser Sachverhalt kann graphisch dargestellt werden durch die zweidimensionale Punktmenge $\{(x_{i1}, x_{i2}) | 1 \leq i \leq n\} \subset \mathbb{R}^2$. Wir unterteilen das x-y-Koordinatensystem durch die Geraden $y = \bar{x}_2$ und $x = \bar{x}_1$ in vier Bereiche (I) bis (IV). Jede Realisierung (x_{i1}, x_{i2}) liegt je nach Zusammenhang in (3.19) im entsprechenden Bereich. Werden nun sämtliche Abstandsprodukte addiert und ergibt sich ein deutlich positiver Wert, so müssen die Realisierungen überwiegend in den Bereichen (I) und (III) liegen. Ist die Summe deutlich negativ, müssen die Realisierungen überwiegend in den Bereichen (II) und (IV) liegen. Dementsprechend liegt ein positiver bzw. negativer Zusammenhang zwischen X_1 und X_2 vor. Wir halten fest

Definition 3.9: Empirische Kovarianz

Seien X_1 und X_2 quantitative Merkmale mit jeweils n Realisierungen und den arithmetischen Mittelwerten $\bar{x}_1$ bzw. $\bar{x}_2$. Dann heißt der Schätzwert der Schätzfunktion $s^2_{X_1X_2} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$(\mathbf{x}_1, \mathbf{x}_2) \mapsto s^2_{X_1X_2} := s^2_{X_1X_2}(\mathbf{x}_1, \mathbf{x}_2) := \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

empirische Kovarianz zwischen den Merkmalen X_1 und X_2.

Bemerkung.
(I) Für $X_1 = X_2$ erhalten wir die empirische Varianz s^2 von X_1.
(II) Offensichtlich gilt $s^2_{X_1X_2} = s^2_{X_2X_1}$.

Fassen wir für k Merkmale X_1 bis X_k die paarweisen empirischen Kovarianzen in einer Matrix zusammen, erhalten wir die symmetrische empirische Varianz-Kovarianzmatrix $S_X = \left(s^2_{X_iX_j}\right) \in \mathbb{R}^{k,k}$. Sie lässt sich bestimmen über

$$S_X = \frac{1}{n-1}(HX)^T HX = \frac{1}{n-1}X^T H^T HX = \frac{1}{n-1}X^T HX.$$

Doch offen bleibt die Frage, was ein deutlich positiver Wert der Kovarianz zweier Merkmale ist. Denn die Summe hängt stark von den Realisierungen ab. Werden nämlich sämtliche Realisierungen von X_1 mit einem Faktor multipliziert, erhöht sich auch die Kovarianz um diesen Faktor. Entsprechendes gilt für das zweite Merkmal und für beide zusammen. Mit dem Faktor u für das erste und dem Faktor v für das zweite Merkmal ergibt sich

$$\frac{1}{n-1}\sum_{i=1}^{n}(u \cdot x_{i1} - u \cdot \bar{x}_1)(v \cdot x_{i2} - v \cdot \bar{x}_2) = \frac{u \cdot v}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

Diesen Nachteil können wir durch eine Normierung mit dem Produkt der empirischen Standardabweichungen beider Merkmale ausgleichen. Damit kann für je zwei kardinal skalierte Merkmale ein vergleichbarer Wert für den Zusammenhang bestimmt werden. Wir erhalten

Definition 3.10: Empirischer Korrelationskoeffizient nach Bravais-Pearson

Für zwei Merkmale X_1 und X_2 mit jeweils n Realisierungen und den empirischen Mittelwerten $\bar{x}_1$ bzw. $\bar{x}_2$ heißt der Schätzwert der Schätzfunktion $\rho_{X_1X_2} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$\rho_{X_1X_2} := \rho_{X_1X_2}(\mathbf{x}_1, \mathbf{x}_2) := \frac{s^2_{X_1X_2}}{s_{X_1}s_{X_2}} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} \cdot \sqrt{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} \tag{3.20}$$

der empirische Korrelationskoeffizient der Merkmale X_1 und X_2.

Bemerkung.
Es gilt: $-1 \leq \rho_{X_1X_2} \leq 1$, da mit $\mathbf{y}_1 = \mathbf{x}_1 - \bar{x}_1\mathbf{1} = H\mathbf{x}_1$ und $\mathbf{y}_2 = \mathbf{x}_2 - \bar{x}_2\mathbf{1} = H\mathbf{x}_2$

$$\rho_{X_1X_2} = \frac{\mathbf{y}_1^T\mathbf{y}_2}{||\mathbf{y}_1|| \cdot ||\mathbf{y}_2||} = \cos(\phi) \in [-1, 1] \tag{3.21}$$

für einen Winkel ϕ ist.

Alle Korrelationen in einer Matrix zusammengefasst erhalten wir die empirische Korrelationsmatrix $R_X = (\rho_{X_iX_j}) \in \mathbb{R}^{k,k}$. Auch sie lässt sich in Matrixschreibweise darstellen. Dazu überlegen wir uns, dass mit (3.21) gilt:

$$\rho_{X_iX_j} = \frac{\mathbf{y}_i^T\mathbf{y}_j}{||\mathbf{y}_i|| \cdot ||\mathbf{y}_j||} = \frac{1}{n-1} \cdot \frac{\mathbf{x}_i^T H \mathbf{x}_j}{s_{X_i}s_{X_j}} = \frac{1}{n-1} \cdot \left(\frac{\mathbf{x}_i}{s_{X_i}}\right)^T H \left(\frac{\mathbf{x}_j}{s_{X_j}}\right).$$

Somit ist $R_X = D_X^{-\frac{1}{2}} S_X D_X^{-\frac{1}{2}}$ mit $D_X^{-\frac{1}{2}} = \text{diag}\left((s_{X_i}^2)^{-\frac{1}{2}}\right)$.

Beispiel 3.11

Gegeben sei die Datenmatrix

$$X = \begin{pmatrix} 22 & 5 \\ 25 & 10 \\ 21 & 4 \\ 28 & 13 \\ 24 & 8 \end{pmatrix} \in \mathbb{R}^{5,2} \text{ mit } \bar{\mathbf{x}} = (24, 8) \text{ und } \mathbf{s}^2 = (7.5, 13.5).$$

Wir erhalten als empirische Kovarianz $s_{X_1X_2}^2 = \frac{40}{4} = 10$ und damit $\rho_{X_1X_2} = \frac{10}{\sqrt{7.5}\sqrt{13.5}} = 0.994$. Es liegt ein fast perfekter positiver Zusammenhang vor.

Beispiel 3.12

Für die Merkmale radius.mv und peri.mv erhalten wir als empirische Kovarianz $s_{X_1X_2}^2 = 85.45$ und damit $\rho_{X_1X_2} = 0.998$. Es liegt ein nahezu perfekter positiver Zusammenhang vor.

Robustheit

Auch bei mehreren Merkmalen ist darauf zu achten, mögliche Ausreißer zu identifizieren. Ausreißer müssen hier zusätzlich im Zusammenhang aller Merkmale gesehen werden. Um eine solche Betrachtung durchzuführen, erweitern wir die Idee des Sensitivitätsdiagramms auf mehrere Merkmale. In Erweiterung zu (3.4) legen wir die Differenz SC fest über

$$SC(\mathbf{x}_{i\cdot}, \mathbf{m}) := n \cdot ||\mathbf{m}_n - \mathbf{m}_{n(i)}||. \tag{3.22}$$

Ist $\mathbf{m}$ etwa der Schwerpunkt, so kann als Norm z.B. die euklidische benutzt werden. Ist $\mathbf{m}$ die empirische Varianz-Kovarianzmatrix, so muss mit einer Matrixnorm gearbeitet werden. Als natürliche Erweiterung der euklidischen Norm kann die Frobeniusnorm einer Matrix $A = (a_{ij}) \in \mathbb{R}^{n,k}$

$$||A|| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij}^2}.$$

benutzt werden. Zur Exploration von einflussreichen Datenpunkten untersuchen wir wiederum einen Graphen der Form

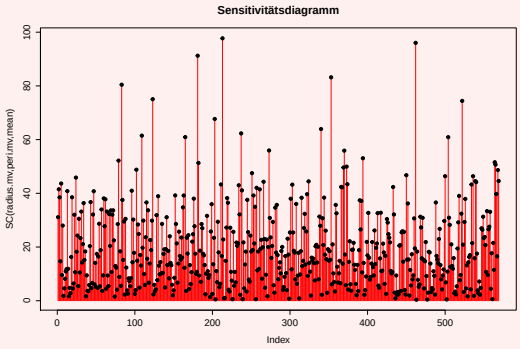
$$\{(x_{i1}, x_{i2}, SC(\mathbf{x}_{i\cdot}, \mathbf{m})); \ i = 1, \ldots, n\}$$

oder

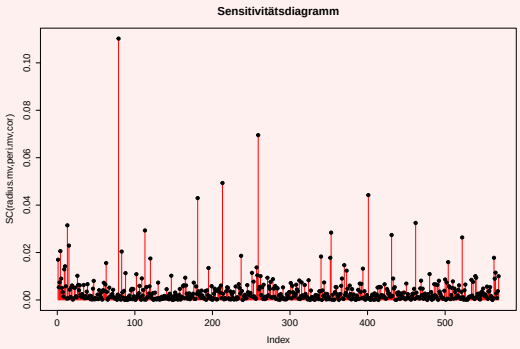
$$\{(i, SC(\mathbf{x}_{i\cdot}, \mathbf{m})); \ i = 1, \ldots, n\}.$$

Beispiel 3.13

Gegeben seien die Merkmale radius.mv und peri.mv. Das Sensitivitätsdiagramm zum Schwerpunkt



zeigt ein paar Kandidaten für Ausreißer. Betrachten wir das Sensitivitätsdiagramm zur empirischen Varianz-Kovarianzmatrix,

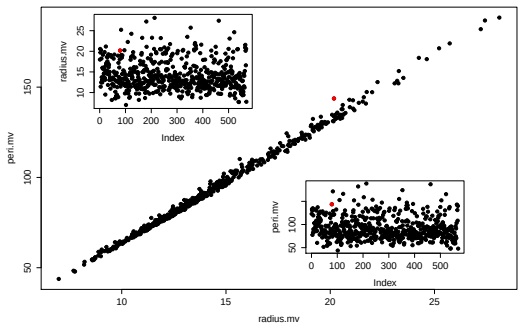


so ist ein Wert besonders auffällig, der auch zu den Kandidaten im ersten Diagramm zählt.

```
u=c()
for(i in 1:l)
{
        u=append(u,l*abs(sqrt(sum(
            (c(mean(radius.mv),mean(peri.mv))
            -c(mean(radius.mv[-i]),mean(peri.mv[-i])))^2)
            )))
}
plot(u,pch=19,main="Sensitivitaetsdiagramm",type="h",
    ylab="SC(radius.mv,peri.mv,mean)",xlab="Index",col=2)
points(u,pch=19)
u=c()
for(i in 1:l)
{
        u=append(u,l*abs( sqrt(2*(cor(radius.mv,peri.mv)
                -cor(radius.mv[-i],peri.mv[-i]))^2) ))
}
plot(u,pch=19,main="Sensitivitaetsdiagramm",type="h",
    ylab="SC(radius.mvmperi.mv,mean)",xlab="Index",col=2)
points(u,pch=19)
```

Scatterplot

Im letzten Beispiel haben wir einen Merkmalsträger identifiziert, der Einfluss auf die empirische Varianz-Kovarianzmatrix und den empirischen Mittelwert nimmt. Um ihn besser beurteilen zu können, betrachten wir einen so genannten Scatterplot der beiden Merkmale radius.mv und peri.mv.



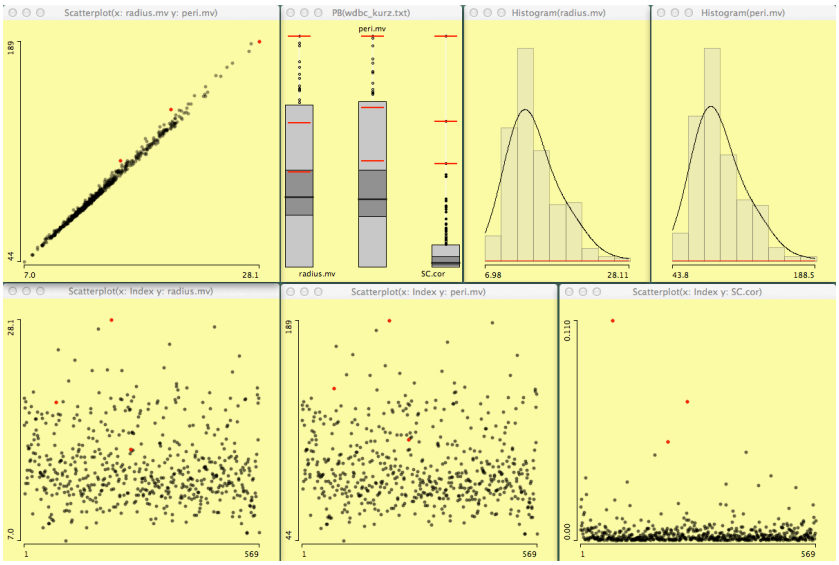

```
plot(radius.mv,peri.mv,pch=19)
```

Dabei werden die Merkmalswertepaare $\{(x_{i1}, x_{i2});\ i = 1, \ldots, n\}$ in ein Koordinatensystem angetragen. Grundsätzlich lässt sich hier der positiv lineare Zusammenhang erkennen. Es ist gedanklich möglich, eine Gerade durch die Daten zu legen. Die Modellierung und Interpretation eines solchen Zusammenhangs erfolgt im Kapitel 8 über die Regressionsanalyse. Der mögliche Ausreißer ist rot gekennzeichnet. Klein sind Scatterplots Index gegen

das jeweilige Merkmal gezeichnet. In keinem der drei Scatterplots würde der rote Merkmalsträger als Ausreißer identifiziert werden.

Interaktivität in statistischer Graphik

Die bisherigen graphischen Darstellungen haben wir separat voneinander betrachtet. Es wäre aber sicher interessant, die beiden erkannten „Ausreißer“ im Sensitivitätsdiagramm aus Beispiel 3.13 mit den einzelnen Boxplots oder dem Scatterplot zu vergleichen. Dies ist mit Hilfe interaktiver Elemente in statistischer Software möglich, wie es von Tukey in [7] erstmals gefordert wurde. Besonders wichtig dabei ist die Technik des gelinkten Highlighting. Dabei werden alle Graphiken eines Datensatzes verbunden, so dass sich Änderungen, Auswahlen oder Anpassungen in einer Graphik oder dem zugrunde liegenden Datensatz auf alle Graphiken übertragen werden (Linking). Eine auf den Merkmalsträger bezogene Teilauswahl der Daten (Selektion) wird durch eine entsprechende farbliche Darstellung der ausgewählten Objekte (Highlighting) illustriert. Durch die Kombination von Linking, Selektion und Highlighting wird eine Selektion in einer Graphik auf alle anderen Graphiken exakt übertragen. Ein weiterer Aspekt ist die Versorgung mit zusätzlichen Informationen auf Wunsch (Abfrage). Im Beispiel selektieren wir die beiden „Ausreißer“ und erfahren, wo sich diese in den Boxplots bzw im Scatterplot wiederfinden.



Selektieren wir die drei größten Werte im Sensitivitätsdiagramm für die empirische Korrelation, können wir durch Highlicghting in allen anderen Plots die entsprechenden Merkmalsträger betrachten. Die beiden größten Werte werden nicht als univariate (für ein Merkmal) Ausreißer erkannt. Lediglich der dritte Wert ist der größte Werte für beide Merkmale.

Ein Histogramm, in dem für ein Intervall keine Werte vorliegen, wird ein Rechteck der Höhe 0 gezeichnet. Das sollte durch einen roten Querstrich gekennzeichnet werden, es erfolgt eine Warnung, eine durch das Tool automatisch erzeugte Darstellung eines besonderen Sachverhalts.

Monotone Zusammenhänge

Der Einfluss eines Merkmalsträgers auf die Kovarianz bzw. Korrelation kann beachtlich groß sein. Die Schätzwerte für den linearen Zusammenhang sind also sehr ausreißeranfällig. Wir können stattdessen bestimmen, wie gut eine monotone Funktion den Zusammenhang beschreiben kann. Vorausgesetzt wird lediglich eine monotone Folge von Werten, auf denen eine Metrik definiert werden kann. Da diese Überlegung auch für komparative Merkmale gilt, betrachten wir kurz Eigenschaften komparativer Merkmale.

Realisierungen eines komparativen Merkmals können der Größe nach geordnet und in eine Rangfolge gebracht werden. Damit eröffnet sich eine erste Beschreibungsmöglichkeit für die Daten, indem die relativen Häufigkeiten aufsummiert werden. Sei B gemäß (2.1) die Menge der beobachteten Ausprägungen, eine Teilmenge des Merkmalsraums M. Wir erhalten

Definition 3.14: Empirische Verteilungsfunktion

Sei X ein komparatives Merkmal mit Merkmalsraum M und m beobachteten Ausprägungen $c_1, \dots, c_m \in B \subseteq M$, wobei $c_i \preceq c_j$ für $i < j$ gelte. Seien weiter durch $h(c_i) = h_i,\ i = 1, \dots, m$, die relativen Häufigkeiten einer beobachteten Anzahl von Realisierungen gegeben. Dann heißt die Abbildung $H : M \to [0, 1]$ mit

$$H(c) := \sum_{c_i \preceq c} h(c_i)$$

empirische Verteilungsfunktion des Merkmals X.

Bemerkung.
(I) Die empirische Verteilungsfunktion ist eine rechtsseitig stetige Treppenfunktion.
(II) Definition 3.14 schreibt keineswegs $m = |M|$ vor und damit auch nicht $c = c_l$ für ein $l \in \{1, \dots, m\}$.

Jede beobachtete Merkmalsausprägung eines Merkmals X erhält einen so genannten Rang $r : B \to \mathbb{R}$

$$c_l \mapsto r(c_l) := \begin{cases} \sum_{i=1}^{l-1} n_i + \frac{n_l(n_l+1)}{2n_l} = n \cdot H(c_{l-1}) + \frac{n_l+1}{2}, & l \neq 1, \\ \frac{n_1+1}{2}, & l = 1, \end{cases}$$

wobei $n_l = n(c_l)$ die absoluten Häufigkeiten und $H(c_l)$ die empirische Verteilungsfunktion des Merkmals X ist. Der letzte Summand ist notwendig, um so genannte Bindungen zu berücksichtigen. Eine Bindung entsteht, wenn mehr als ein Merkmalsträger den gleichen Merkmalswert eines Merkmals X besitzt. Damit erhalten die sortierten Merkmalsträger einer Stichprobenmenge S den Rang

$$r_X : S \to \mathbb{R},\ s_{(i)\cdot} \mapsto R_X(s_{(i)\cdot}) := r\left(x_{(i)}\right).$$

Der „Rangdurchschnitt“ $\bar{r}_X$ eines Merkmals X mit n Realisierungen und m beobachteten

Merkmalsausprägungen ist $\bar{r}_X = \frac{n+1}{2}$.

$$\begin{aligned}
\bar{r}_X &= \frac{1}{n}\sum_{i=1}^{n} r_X(\tilde{\omega}_{i\cdot}) = \frac{1}{n}\sum_{i=1}^{n} r(X(\tilde{\omega}_{i\cdot})) \\
&= \frac{1}{n}\sum_{l=1}^{m} n_l r(c_l) = \frac{1}{n}\sum_{l=1}^{m} n_l \left(\sum_{i=1}^{l-1} n_i + \frac{n_l+1}{2}\right) \\
&= \frac{1}{n}\sum_{l=1}^{m} \frac{1}{2}\left(n_l^2 + n_l + 2\sum_{i=1}^{l-1} n_l n_i\right) \\
&= \frac{1}{2n}\left(\sum_{l=1}^{m} n_l^2 + \sum_{l=1}^{m} n_l + 2\sum_{l=1}^{m}\sum_{i=1}^{l-1} n_l n_i\right) \\
&= \frac{1}{2n}\left(\sum_{l=1}^{m} n_l + \left(\sum_{l=1}^{m} n_l\right)^2\right) \\
&= \frac{1}{2n}\left(n + n^2\right) \\
&= \frac{n+1}{2}
\end{aligned} \tag{3.23}$$

Die gemittelten quadrierten Abstände der Ränge vom Rangdurchschnitt ergeben die Rangvarianz und diese errechnet sich über

$$\sigma^2_{r_X} = \frac{1}{n-1}\sum_{i=1}^{n}(r_X(s_{(i)\cdot}) - \bar{r}_X)^2.$$

Mit diesen Voraussetzungen können wir ein Assoziationsmaß festlegen in

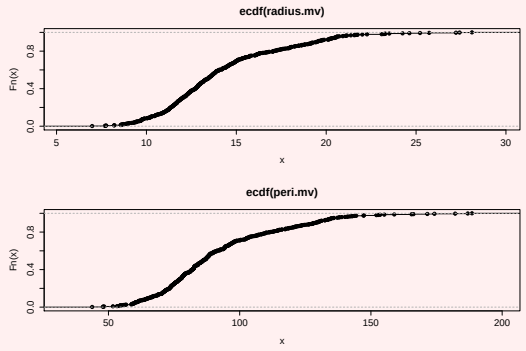
Definition 3.15: Korrelationskoeffizient nach Spearman

Der Spearmansche Rangkorrelationskoeffizient zwischen zwei komparativen Merkmalen X_1 und X_2 ist gegeben durch

$$\rho_{X_1,X_2} := \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{r_{X_1}(s_{(i)\cdot}) - \bar{r}_{X_1}}{\sqrt{\sigma^2_{r_{X_1}}}} \cdot \frac{r_{X_2}(s_{(i)\cdot}) - \bar{r}_{X_2}}{\sqrt{\sigma^2_{r_{X_2}}}}\right). \tag{3.24}$$

Beispiel 3.16

Die empirischen Verteilungsfunktionen bei den Merkmalen radius.mv und peri.mv bei den Brustkrebsdaten sehen folgendermaßen aus:



Der Rangdurchschnitt ist $\bar{R}_{X_1} = \bar{R}_{X_2} = 285$. Es gibt bei beiden Merkmalen einzelne Bindungen. Die Rangvarianzen ergeben sich zu $\sigma^2_{R_{X_1}} = 27027.37$, $\sigma^2_{R_{X_2}} = 27027.45$ und der Spearmansche Rangkorrelationskoeffizient ist $\rho_{X_1,X_2} = 0.998$. Es besteht eine starke monotone Beziehung zwischen radius.mv und peri.mv.

```
rank(radius.mv)
rank(peri.mv)
var(rank(radius.mv))
var(rank(peri.mv))
cor(rank(radius.mv),rank(peri.mv))
cor(radius.mv,peri.mv,method = c("spearman"))
plot(ecdf(radius.mv))
plot(ecdf(peri.mv))
```

Verallgemeinerte Varianz

Eine Verallgemeinerung der Varianz auf die Situation mit mehr als einem Merkmal führt auf zwei Ansätze: Die Totalvariation s_t^2 und die verallgemeinerte empirische Varianz s_v^2.

$$s_t^2 : \mathbb{R}^{n,k} \to \mathbb{R}, X \mapsto s_t^2(X) := \text{sp}(S) = \sum_{i=1}^{k} s^2_{X_i X_i}, \tag{3.25}$$

$$s_v^2 : \mathbb{R}^{n,k} \to \mathbb{R}, X \mapsto s_v^2(X) := \det(S). \tag{3.26}$$

Beide Werte verdichten die vorhandene Information sehr stark. Die Idee der Totalvarianz hat den Vorteil, dass sie sich bei einer orthogonalen Transformation nicht ändert. Für einen Punkt $\mathbf{x} = (x_1, \ldots, x_k)^T \in \mathbb{R}^k$ gilt $\mathbf{x} = (x_1, \ldots, x_k)^T = \sum_{i=1}^{k} x_i \mathbf{e}_i$ mit der kanonischen Basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$. Der Punkt soll mit Hilfe einer anderen Orthonormalbasis $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ darge-

stellt werden. Sei Q die Matrix, deren Spalten die Basisvektoren $\mathbf{q}_1, \ldots, \mathbf{q}_k$ sind. Es folgt

$$\sum_{i=1}^{k} x_i \mathbf{e}_i = \sum_{i=1}^{k} y_i \mathbf{q}_i \Leftrightarrow I^{k,k}\mathbf{x} = Q\mathbf{y} \Leftrightarrow \mathbf{y} = Q^T\mathbf{x}.$$

Ist $X \in \mathbb{R}^{n,k}$ eine Datenmatrix, so entsprechen die Zeilen den Objektpositionen (Punkten) im $\mathbb{R}^k$. Für das j-te Objekt gilt $\sum_{i=1}^{k} x_{ji}\mathbf{e}_i = \sum_{i=1}^{k} y_{ji}\mathbf{q}_i$ genau dann, wenn $I^{k,k}\mathbf{x}_{j\cdot}^T = Q\mathbf{y}_{j\cdot}^T$, somit $\mathbf{y}_{j\cdot} = \mathbf{x}_{j\cdot}Q$. Stellen wir alle Zeilen von X in der neuen Basis dar, so erhalten wir $Y = XQ$.

Wie ändert sich die empirische Varianz-Kovarianzmatrix bei einer orthogonalen Transformation der Daten? Mit $S_X = \frac{1}{n-1}X^THX$ gilt

$$S_Y = \frac{1}{n-1}(XQ)^TH(XQ) = Q^T\frac{1}{n-1}X^THXQ = Q^TS_XQ.$$

Satz und Definition 3.17: Spur einer Matrix

Sei $A = (a_{ij}) \in \mathbb{R}^{k,k}$. Die Spur von A ist die Summe der Diagonalelemente von A, $\mathrm{sp}(A) := \sum_{i=1}^{k} a_{ii}$. Mit $B = (b_{ij}) \in \mathbb{R}^{k,k}$ gilt $\mathrm{sp}(AB) = \mathrm{sp}(BA)$.

Beweis.

$$\mathrm{sp}(AB) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}b_{ji} = \sum_{j=1}^{k}\sum_{i=1}^{k} b_{ji}a_{ij} = \mathrm{sp}(BA).$$

□

Mit Satz 3.17 folgt

$$\mathrm{sp}(S_Y) = \mathrm{sp}(Q^TS_XQ) = \mathrm{sp}(QQ^TS_X) = \mathrm{sp}(S_X). \tag{3.27}$$

Dies werden wir uns bei der Hauptkomponentenanalyse in Kapitel 5 zunutze machen. Dafür berücksichtigt die Kenngröße keinerlei Information über Kovarianzen zwischen den Merkmalen. Zwar geschieht dies bei der verallgemeinerten empirischen Varianz, dennoch kann der einzelne Wert keine unterschiedlichen Strukturen in den Kovarianzen erklären.

Standardisierung

Dividieren wir die zentrierten Daten eines Merkmals X durch die Standardabweichung (hier mit Index X), $w_i := \frac{x_i - \bar{x}}{s_X}$, ergibt sich ein neues Merkmal $W = HXD_X^{-\frac{1}{2}}$ mit empirischem Mittelwert von 0 und einer empirischen Varianz von 1:

$$\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i-\bar{x}}{s_X} = 0,$$
$$s_W^2 = \frac{1}{n-1}\sum_{i=1}^{n} w_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_X}\right)^2 = \frac{1}{s_X^2}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{s_X^2}{s_X^2} = 1.$$

Deren Varianz-Kovarianzmatrix ist gemäß (3.21) zugleich ihre Korrelationsmatrix:

$$\begin{aligned} S_W &= \frac{1}{n-1}(HXD_X^{-\frac{1}{2}})^T H(HXD_X^{-\frac{1}{2}}) = D_X^{-\frac{1}{2}} S_X D_X^{-\frac{1}{2}} = R_X. \\ R_W &= S_W = R_X. \end{aligned}$$

Die aus X durch

$$\mathbf{z} := \frac{H\mathbf{x}}{||H\mathbf{x}||\sqrt{n-1}} \tag{3.28}$$

mit $z_i = \frac{x_i - \bar{x}}{s_X\sqrt{n-1}}$ gewonnenen Daten heißen kleine standardisierte Daten. Um eine Datenmatrix X so zu standardisieren, ist

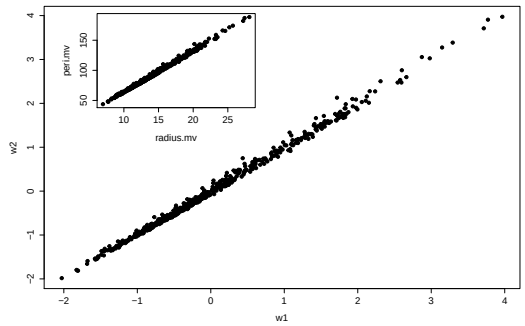
$$Z = \frac{1}{\sqrt{n-1}} HXD_X^{-\frac{1}{2}} = \frac{1}{\sqrt{n-1}} W$$

zu berechnen. Hierbei gilt

$$\begin{aligned} S_Z &= \frac{1}{n-1}\frac{1}{\sqrt{n-1}}(HXD_X^{-\frac{1}{2}})^T H \cdot \frac{1}{\sqrt{n-1}}(HXD_X^{-\frac{1}{2}}) = D_X^{-\frac{1}{2}} S_X D_X^{-\frac{1}{2}} = \frac{1}{n-1} R_X, \\ R_Z &= \text{diag}(\sqrt{n-1})\frac{1}{n-1} R_X \text{ diag}(\sqrt{n-1}) = R_X, \\ Z^T Z &= \frac{1}{n-1} D_X^{-\frac{1}{2}} X^T H^T H X D_X^{-\frac{1}{2}} = R_X. \end{aligned}$$

```
w1=(radius.mv-mean(radius.mv))/sd(radius.mv)
w2=(peri.mv-mean(peri.mv))/sd(peri.mv)
mean(w1)
var(w1)
cov(cbind(w1,w2))
cor(cbind(w1,w2))
plot(w1,w2,pch=19)
```

Die Abbildung zeigt den Scatterplot der standardisierten Daten, oben eingeklinkt ist der Scatterpelot der Originaldaten zu sehen.



4. Merkmale mit Nominalskala

Warum?

Es gibt manchmal Wertemengen zu Eigenschaften, die sich nicht in das Konzept der Menge der reellen zahlen einbetten lassen. Mit Konzept ist dabei gemeint, wie sich die erhobenen Daten weiterverarbeiten lassen. Dennoch ist es sinnvoll und notwendig, auch solche Daten durch Kenngrößen und Visualisierungen zu beschreiben. Ebenso ist es häufig von großem Interesse, ob es eine Beziehung zwischen zwei oder mehreren solchen Eigenschaften gibt.

Die Analyse gemeinsam auftretender Werte (Items) hat durch die rasante Entwicklung des Internets enorm an Bedeutung gewonnen. Ausgehend von vereinzelten Marketing-basierten Verhaltensanalysen von Kunden anfang der 1990er Jahre, der so genannten Warenkorbanalyse, sind heute derartige Ansätze weit verbreitet. Ein oft anzutreffender Text auf diversen Webportalen ist etwa

> „Kunden, die dieses Produkt gekauft haben, kauften auch . . . “

Die Analyse des Nutzerverhaltens (Click-Verhalten) auf einzelnen Webseiten oder die Untersuchung von Texten auf Zusammenhänge zwischen vorkommenden Worten und einzelnen Schlüsselbegriffen sind Beispiele für die Untersuchung gemeinsam auftretender Werte. Auch in den Ingenieurwissenschaften finden entsprechende Verfahren Anwendung.

Beispiel 4.1: Diagnostik von Störungen

Oft ist es schwierig, die Ursache für eine Störung an technischen Anlagen zu finden. Diagnosesysteme nutzen physikalische Modelle und Erfahrungen, um die Ursachenfindung zu unterstützen. Physikalische Eigenschaften der technischen Anlage und konkrete Messungen einzelner Größen werden zu so genannten Assoziationsregelmodellen verarbeitet. Die Wirkung (Störung) wird dabei auf Basis gewisser Kriterien durch Ursachen (Werte physikalischer Größen) erklärt.

Grundlegend für eine Vielzahl an Verfahren zur Zusammenhangsanalyse ist der so genannte Item-Support. Das heißt nichts anderes als dass wir das Vorkommen von (Merkmals)Werten abzählen müssen.

4.1. Modus und Informationsentropie

Bevor wir die Item-Analyse beginnen können, benötigen wir noch etwas Vorwissen für die Untersuchung von Merkmalen mit Nominalskala.

Lageparameter

Bei einer Realisierung eines qualitativen Merkmals gibt es die Möglichkeit, die Ausprägungen auszuzählen. Interessant dabei ist, welche Ausprägung am häufigsten auftritt. Diese Kenngröße legen wir fest in

Definition 4.2: Modus

Sei X ein qualitatives Merkmal mit $m = |M|$ Kategorien $c_1, \ldots, c_m$ und repräsentiere das Tupel $(n_1, \ldots, n_m)$ die absoluten Häufigkeiten der Kategorien $c_1, \ldots, c_m$ einer Stichprobe von X. Die Abbildung

$$\begin{aligned} md: \quad & \mathbb{N}^m \to \mathcal{P}(\{c_1, \ldots, c_m\}), \\ & (n_1, \ldots, n_m) \mapsto Mod := md(n_1, \ldots, n_m) := \{c_i;\ n_i \geq n_l\ \forall l \in \{1, \ldots, m\}\} \end{aligned}$$

liefert die am häufigsten beobachteten Kategorien und jede Kategorie $c_l \in Mod$ heißt Modus des Merkmals X.

Beispiel 4.3

Der Modus des Merkmals Class aus dem Ttanic-Beispiel ist die Kategorie „Crew“. Der Modus des Merkmals Age ist die Kategorie „Adult“.

In R kann der Modus über folgende Funktion bestimmt werden.

```
modus<-function(X)
{       a<-table(X, exclude=NULL)
        c<-which(a==max(a))
        return(c)
}
modus(Class)
```

Wir erzeugen zunächst eine Tabelle der absoluten Häufigkeiten des Merkmals, wobei auch fehlende Werte mit aufgenommen werden. Über den which-Befehl werden diejenigen Indizes und Kategoriebezeichnungen gespeichert, für welche die Häufigkeit am größten ist.

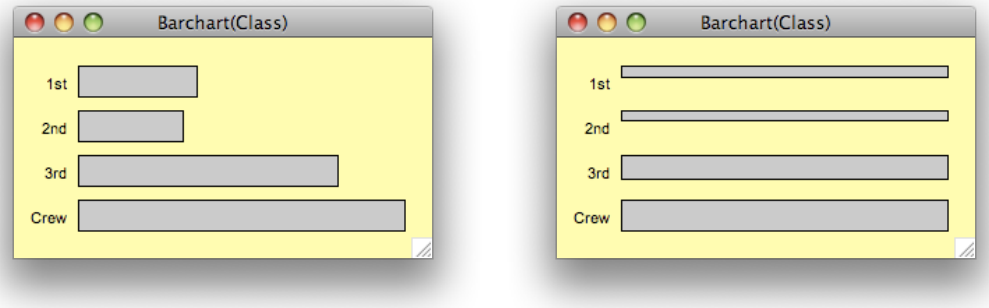
```
Crew
    1
```

Die erste Kategorie der Crew ist der Modus des Merkmals Class im Titanic-Beispiel.

Barchart und Spineplot

Barcharts und Spineplots werden verwendet, um die absoluten oder relativen Häufigkeiten von Kategorien darzustellen. Für jede Kategorie wird eine Rechtecksfläche (Balken) erzeugt. Dabei ist die Häufigkeit proportional zur dargestellten Fläche. Um die Häufigkeiten vergleichen zu können, wird beim Barchart die gleiche Breite gewählt. Die Höhe der

Balken ist damit maßgebend für die Häufigkeiten.




```
dat=read.table("Titanic.txt",sep="\t",head=TRUE)
attach(dat)
barplot(table(Class))
mosaicplot(table(Class))
```

Das linke Bild zeigt einen Barchart der Kategorien des Merkmals Class der Titanic-Daten. In einem Barchart lässt sich oft ohne zusätzlichen Aufwand der Modus erkennen. In unserem Fall ist es die Kategorie Crew.

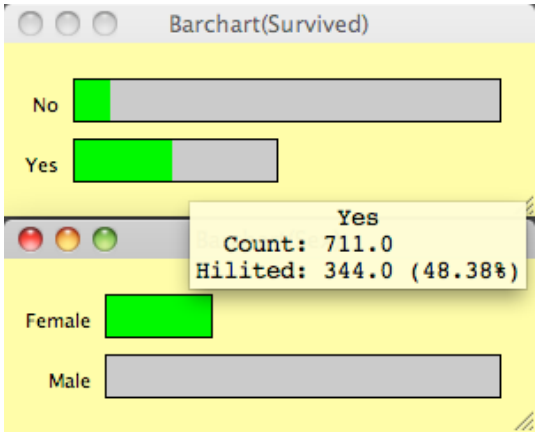
Beim Spineplot ist dagegen, wie im rechten Bild (trotz der Beschriftung mit Barchart) zu sehen, die Höhe der Balken gleich und die Breite der Balken ist damit proportional zu den Häufigkeiten. Oftmals interessiert uns ein gemeinsames Auftreten von Merkmalswerten unterschiedlicher Merkmale. Wenn wir wissen wollen wie viele Frauen unter den Überlebenden waren, so können wir dies zunächst mit Hilfe einer Kreuztabelle (siehe Abschnitt 4.2) untersuchen.

```
ftable(Sex,Survived)
```

Wir nutzen den ftable-Befehl, da er in der Darstellung Vorteile hat, wie wir noch sehen werden.

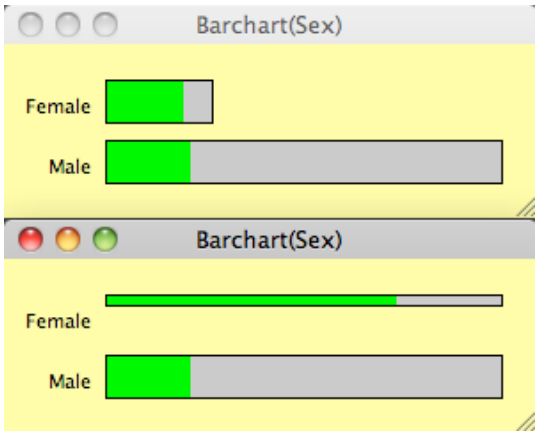
```
        Survived   No  Yes
Sex
Female            126  344
Male             1364  367
```

Wir sehen, dass 344 Frauen überlebt haben. Doch die Frage nach dem Anteil Frauen unter den Überlebenden können wir nicht direkt beantworten. Dazu müssen wir den Anteil der Frauen unter allen Überlebenden betrachten, also $\frac{344}{344+367} = 48\%$. Hier gibt es mit Hilfe interaktiver Techniken die Möglichkeit, schneller ans Ziel zu gelangen. Durch gelinktes Higlighting in Verbindung mit einer Abfrage können wir die Antwort sehen.



Durch Highlighting wird der Anteil der Häufigkeiten in jeder Kategorie andersfarbig dargestellt, der beispielsweise durch die Kategorie Yes des Merkmals Survived repräsentiert wird.

Der Spineplot hat Vorteile gegenüber dem Barchart, wenn wir Anteile in den Häufigkeiten vergleichen wollen. Wir fragen uns, ob anteilig mehr Männer oder Frauen überlebt haben.

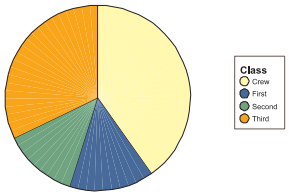


Wir sehen den entscheidenden Vorteil bei der Anwendung des gelinkten Highlighting, da wir sofort erkennen können, dass anteilig mehr Frauen überlebt haben als Männer. Der Barchart gibt die absoluten Zahlen wider, die bei beiden Kategorien annähernd gleich sind.

Pie charts

Ein Pie chart ist eine Graphik, bei der die Fläche eines Kreises entsprechend der Häufigkeiten der Kategorien eines qualitativen Merkmales in Segmente aufgeteilt wird.

Nebenstehendes Bild zeigt einen Pie chart des Merkmals Class. Die Intention des Pie charts entspricht dem des Barcharts. Pie charts sind jedoch eine weniger vorteilhafte Art und Weise, um Informationen über Häufigkeiten darzustellen. Das Auge ist gut darin, lineare Maße abzuschätzen und schlecht darin, relative Flächen zu beurteilen. Ist in dem Pie chart die grüne (Second) oder die blaue (First) Fläche größer?



Streuparameter und Konzentrationsmessung

Die Häufigkeiten eines qualitativen Merkmals können verwendet werden, um zu untersuchen, wie gleichmäßig sich die Daten auf die einzelnen Kategorien verteilen. Entfallen sämtliche Ausprägungen auf einen Wert, so liegt keine Streuung in den Daten, d.h. wir nehmen sicher an, dass bei einer weiteren Realisierung derselbe Wert auftritt. Ist dies nicht der Fall, sind die Daten nicht konzentriert und es liegt eine Unsicherheit bzgl. der Vorhersage der Merkmalsausprägung einer neuen Realisierung vor. Wir können uns somit überlegen, dass die Menge an Information, die in einer neuen Realisierung liegt, als nichtnegative Funktion der relativen Häufigkeiten beschrieben werden kann. Was fordern wir von einer solchen Funktion $s : [0, 1] \to [0, \infty]$? Zunächst soll $s(0) = \infty$ und $s(1) = 0$ gelten. Die Information zweier Ereignisse, die sich gegenseitig nicht beeinflussen, soll sich zudem aufaddieren, d.h. $s(ab) = s(a) + s(b)$. Eine stetige Funktion, die diese Eigenschaften erfüllt, ist die Logarithmus-Funktion. Zur Bestimmung der Streuung bei qualitativen Merkmalen kann das Konzept der Informationsentropie verwendet werden. Die Informationsentropie beschreibt die Gleichmäßigkeit der Häufigkeiten aufgrund der mittleren zu erwartenden Menge an Information (genannt Entropie) und ist ein Erwartungswert.

Definition 4.4: Streuung qualitativer Merkmale

Sei X ein qualitatives Merkmal mit $m = |M|$ Kategorien und seien $h_l, l = 1, \ldots, m$, die relativen Häufigkeiten der Kategorien bei einer gegebenen Stichprobe mit $\sum_{l=1}^{m} h_l = 1$. Es gelte $0 \cdot \ln 0 := 0$. Die Informationsentropie V der Verteilung ist dann definiert als Funktion $V : [0, 1]^m \to [0, 1]$ mit

$$(h_1, \ldots, h_m) \mapsto V(h_1, \ldots, h_m) := -\frac{1}{\ln(m)} \sum_{l=1}^{m} h_l \ln(h_l).$$

Der Wertebereich für V soll zwischen 0 und 1 liegen. Um das zu zeigen, müssen wir ein Minimierungsproblem lösen.

Satz und Definition 4.5: Allgemeine Minimierungsprobleme

Ein allgemeines Minimierungsproblem (MP) ist von der Form

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\Gamma} & \{f(\mathbf{x})\} \\ \text{unter} & g_i(\mathbf{x}) \leq 0 \quad \forall i = 1,\ldots,m, \\ & h_j(\mathbf{x}) = 0 \quad \forall j = 1,\ldots,l, \end{array} \tag{4.1}$$

mit $f : \Gamma \to \mathbb{R}$, $g_i : \Gamma \to \mathbb{R}$, $h_j : \Gamma \to \mathbb{R}$ und $\Gamma \subseteq \mathbb{R}^n$.

Oftmals lassen sich Minimierungsprobleme mit Nebenbedingungen (die (Un-)gleichungen mit g_i und h_j) nicht direkt lösen. Eine Möglichkeit besteht in der Umformulierung zu einem so genannten Lagrange-Problem.

Satz und Definition 4.6: Lagrange-Probleme

Gegeben sei ein Minimierungsproblem (MP). Dann heißt das Problem

$$\begin{array}{ll} \max\limits_{\mathbf{u}\in\mathbb{R}^m,\, \mathbf{v}\in\mathbb{R}^l} & \left\{\Theta(\mathbf{u},\mathbf{v}) := \inf\limits_{\mathbf{x}\in\Gamma}\{f(\mathbf{x}) + \sum\limits_{i=1}^{m} u_i g_i(\mathbf{x}) + \sum\limits_{j=1}^{l} v_j h_j(\mathbf{x})\}\right\} \\ \text{unter} & u_i \geq 0 \quad \forall i = 1,\ldots,m, \\ & \mathbf{v} \in \mathbb{R}^l, \end{array} \tag{4.2}$$

mit der Lagrange-Funktion $\Theta : \mathbb{R}^m \times \mathbb{R}^l \to \mathbb{R}$ das Lagrange-Duale Problem (DP) zu (MP).

In manchen Fällen gibt es Kandidaten für Optimalpunkte, d.h. Lösungen von (MP) und (DP). Dazu müssen sie folgende Bedingungen erfüllen.

Satz und Definition 4.7: KKT-Bedingungen

Ein Minimierungsproblem liege in der Formulierung (DP) gemäß (4.2) vor. Ein $\mathbf{x} \in \Gamma$ erfüllt die KKT-Bedingungen, falls f und $\mathbf{g} = (g_1,\ldots,g_m)$ differenzierbar sind und u_i, v_j existieren mit

$$\begin{aligned} \nabla f(\mathbf{x})^T + \mathbf{u}^T J_g(\mathbf{x}) + \mathbf{v}^T J_h(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{u}^T \mathbf{g}(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{g}(\mathbf{x}) &\leq \mathbf{0}, \\ \mathbf{h}(\mathbf{x}) &= \mathbf{0}, \\ \mathbf{u} &\geq \mathbf{0}. \end{aligned} \tag{4.3}$$

Sind f und diejenigen g_i mit $g_i(\mathbf{x}) = 0$ konvex bei $\mathbf{x}$ und sind für $v_j \neq 0$ die h_j affin linear, dann ist $\mathbf{x}$ ein Optimalpunkt für (MP) und $(\mathbf{u},\mathbf{v})$ einer für (DP).

Nun können wir uns den Wertebereich von V überlegen.

Satz 4.8

Die Funktion ist sinnvoll definiert, d.h. es gilt: $0 \leq V(h_1, \ldots, h_m) \leq 1$.

Beweis.

Es ist $h_l \ln(h_l) \leq 0$ für jedes l und damit ist V stets positiv. V ist eine durch die Nebenbedingung $\sum_{l=1}^{m} h_l = 1$ im Definitionsbereich eingeschränkte Funktion und nimmt das Maximum dort an, wo $h_l = \frac{1}{m}$ für alle $l \in \{1, \ldots, m\}$ gilt. Denn betrachten wir die dazugehörige zu minimierende Lagrange-Funktion

$$\Theta(v) := \frac{1}{\ln m} \sum_{l=1}^{m} h_l \ln h_l + v \left(\sum_{l=1}^{m} h_l - 1 \right),$$

und bilden die partiellen Ableitungen nach h_j, $\frac{\partial \Theta}{\partial h_j} = \frac{1}{\ln m}(\ln h_j + 1) + v$, bzw. v, $\frac{\partial \Theta}{\partial v} = \left(\sum_{l=1}^{m} h_l - 1 \right)$, so erhalten wir durch Nullsetzen der Gradienten

$$\begin{aligned} \frac{\partial L}{\partial h_j} = 0 \quad &\Leftrightarrow \quad h_j = e^{-u \ln m - 1} \text{ bzw. durch Einsetzen} \\ \frac{\partial \Theta}{\partial u} = 0 \quad &\Leftrightarrow \quad u = 1 - \frac{1}{\ln m}. \end{aligned}$$

Damit ergibt sich $h_j = \frac{1}{m}$. Weil die Hessematrix $H(h_1, \ldots, h_m)$ positiv definit ist, folgt, dass es sich um eine Minimalstelle handelt. Wir untersuchen die beiden Extremfälle. Sei zunächst $h_i = 1$ für ein $i \in \{1, \ldots, m\}$ und $h_j = 0$ für alle $j \neq i$. Dann ist

$$-\frac{1}{\ln(m)} \sum_{l=1}^{m} h_l \ln(h_l) = -\frac{1}{\ln(m)} 1 \cdot \ln(1) = 0.$$

Sei andererseits $h_l = \frac{1}{m}$ für alle $l \in \{1, \ldots, m\}$. Dann haben wir

$$-\frac{1}{\ln(m)} \sum_{l=1}^{m} h_l \ln(h_l) = -\frac{1}{\ln(m)} m \cdot \frac{1}{m} \ln\left(\frac{1}{m}\right) = -\frac{\ln(m^{-1})}{\ln(m)} = 1.$$

□

Beispiel 4.9: Titanic: Class und Age

Wir berechnen die Informationsentropien V_1, V_2 für die beiden qualitativen Merkmale

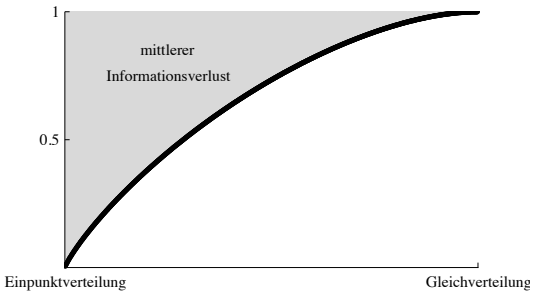
Class und Age. Es ist

$$V_2 = -\frac{1}{\log(4)}\left(\frac{885}{2201}\log\left(\frac{885}{2201}\right)+\ldots+\frac{706}{2201}\log\left(\frac{706}{2201}\right)\right) = 0.922$$
$$V_1 = -\frac{1}{\log(2)}\left(\frac{2092}{2201}\log\left(\frac{2092}{2201}\right)+\frac{109}{2201}\log\left(\frac{109}{2201}\right)\right) = 0.284$$

Für das Merkmal Klasse wird 92% der möglichen Information geliefert, es liegt viel Information in diesem Merkmal. Hingegen erhalten wir bei Age nur 28% Information, d.h. wir haben einen mittleren Informationsverlust von 72%. Es liegt wenig Information im diesem Merkmal.

```
informationsentropie<-function(X){
k<-prop.table(table(X))
inf<-sum(k*log(k))*(-1/log(length(k)))
print(inf)
}
informationsentropie(CarHire)
```

Bemerkung.
Die Informationsentropie darf nicht als linear interpretiert werden. Die folgende Abbildung verdeutlicht diesen Umstand durch die schematische Entwicklung der Informationsentropie von der Einpunkt- zur Gleichverteilung von relativen Häufigkeiten. Bei einer Einpunktverteilung enthalten die Daten keinerlei Information, während bei einer Gleichverteilung maximale Information in den Daten vorhanden ist.



4.2. Assoziationen

Die Häufigkeitstabellen einzelner qualitativer Merkmale können wir in einer Matrix, der so genannten Kontingenztabelle, zusammenbringen. Seien k qualitative Merkmale $X_1, \ldots, X_k$ gegeben. Die Kategorien des ersten Merkmals werden in die erste Spalte der Matrix eingetragen. Die Kategorien des zweiten Merkmals in der ersten Zeile. Mit dem dritten Merkmal findet eine neuerliche Unterteilung der Spalten statt. In die zweite Spalte werden nämlich für jede Kategorie des ersten Merkmals sämtliche Kategorien des dritten Merkmals eingetragen. Das vierte Merkmal unterteilt die erste Zeile mit dem zweiten Merkmal usw. Sei m_j die Anzahl der Kategorien des Merkmals X_j. Dann ergibt sich eine $\prod_{l=2j-1}^{k} m_l \times \prod_{l=2j}^{k} m_l$-Zellen Matrix mit $1 \leq j \leq k$, in deren Zellen die jeweils entsprechend der Kategorien vorhandenen Fälle einzutragen sind.

Beispiel 4.10

Wir erstellen eine Kontingenztabelle mit absoluten Häufigkeiten für die drei Merkmale Class, Age und Sex (in dieser Reihenfolge).

Kontingenztabelle

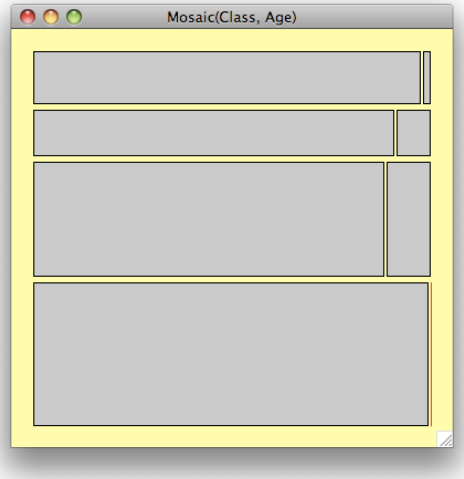
Class	Sex	Age		
		Adult	Child	Summe
Crew	Female	23	0	23
	Male	862	0	862
1st	Female	144	1	145
	Male	175	5	180
2nd	Female	93	13	106
	Male	168	11	179
3rd	Female	165	31	196
	Male	462	48	510
Summe		2092	109	2201

Die letzte Zeile und die letzte Spalte wurden hinzugefügt, um Zeilen- bzw. Spaltensummen zu erfassen. Genauso gut hätten wir die Kontingenztabelle mittels der relativen Häufigkeiten der einzelnen Fälle befüllen können.

Mosaicplot

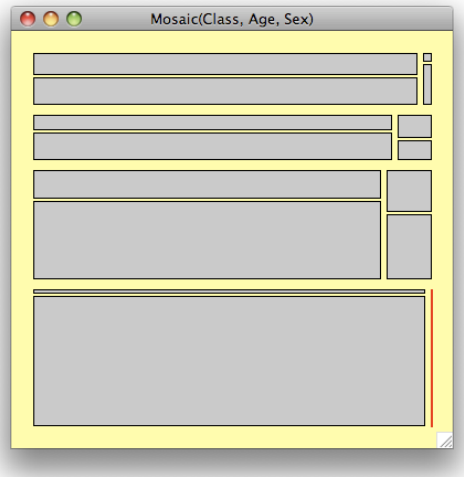
Der Prozess der Erstellung des Mosaicplots ist grundsätzlich gleich zum Aufbau einer Kontingenztabelle, wie wir es in Abschnitt 4.2 bei der Erstellung einer Kontingenztabelle gesehen haben.

Wir beginnen mit einem Spineplot für das erste Merkmal. Das zweite Merkmal wird hinzugefügt, indem die Flächen jeder Kategorie vertikal entsprechend der Proportionen der Kategorien des zweiten Merkmals unterteilt werden. Im Beispiel auf der rechten Seite beginnen wir mit einem Spineplot für das Merkmal Class und fügen das Merkmal Age mit seinen zwei Kategorien hinzu. Die Fläche oben rechts repräsentiert hier den Anteil der Kinder der ersten Klasse, die Fläche oben links den Anteil der Erwachsenen der ersten Klasse. Unten rechts ist durch den roten Strich zu erkennen, dass keine Kinder Mitglied der Crew der Titanic waren.



```
mosaicplot(Class,Age)
```

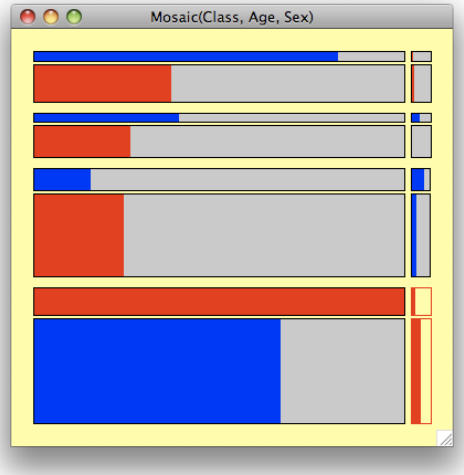
Das Merkmal Sex gibt das Geschlecht jedes Merkmalsträgers an. Erweitern wir den Mosaicplot um das Merkmal, werden die Flächen wiederum horizontal unterteilt. Die oberen Flächen repräsentieren weibliche Personen, die unteren Flächen männliche. Es zeigt sich, dass es in der ersten Klasse lediglich ein Mädchen unter sechs Kindern gab. Lässt sich nun die Frage beantworten, ob auch auf der Titanic das Prinzip „Frauen und Kinder zuerst galt“ oder ob überproportional viele Personen der ersten Klasse überlebt haben? An dieser Stelle soll auf die interaktiven Möglichkeiten der Exploration der Daten verwiesen werden! Die Vorgehensweise des Hinzufügens von Merkmalen kann nun beliebig mit weiteren qualitativen Merkmalen fortgesetzt werden.



Anhand des Mosaicplots können wir Aussagen über die Abhängigkeit der einzelnen Merkmale treffen. Um das zu untersuchen, müssen sämtliche Reihenfolgen der betrachteten Merkmale gebildet und davon jeweils ein Mosaicplot erstellt werden. Wir betrachten die Räume zwischen den Flächen. Lassen sich diese in jedem Mosaicplot ohne „übermäßiges Abknicken“ durchlaufen, d.h. es liegt eine gleichmäßige Partitionierung vor, kann von der Unabhängigkeit der Merkmale ausgegangen werden (mutual independence). Andernfalls muss eine Abhängigkeit angenommen werden, die weiter untersucht werden kann.

Bei drei Merkmalen können dann die partial independence, die conditional independence, die no three-way interaction und die three-way interaction erkannt werden. Bei der partial independence ist ein Merkmal von den beiden anderen unabhängig, was sich anhand der Gleichmäßigkeit der beiden Plots zeigt, in denen das bestimmte Merkmal mittig angeordnet ist. Die conditional independence bedeutet, dass zwei Merkmale unter gegebenem dritten unabhängig voneinander sind. Das zeigt sich anhand einer gleichmäßigen Partitionierung innerhalb der Kategorien des gegebenen ersten Merkmals. Allerdings ist die Partitionierung in den beiden entsprechenden Mosaicplots nicht gleichförmig. Die beiden letzten Formen der Unabhängigkeit sind sowohl aus Sicht der Interpretation als auch anhand der Mosaicplots schwer zu erkennen. In unserem Beispiel interpretieren wir anhand der sechs Mosaicplots, dass bei gegebenem Merkmal Class die beiden anderen Merkmale nahezu unabhängig sind, also eine conditional independence vorliegt. Wir stellen die Hypothese auf, dass sich ein Modell aus den drei Merkmalen, und so genannten Interaktionstermen zwischen Class und Age bzw. Class und Sex zusammensetzt.

Es besteht die Möglichkeit, einen Mosaicplot nicht in der Intention wie bisher, sondern unter Annahme des Unabhängigkeitsmodells die erwarteten Werte darzustellen. Ist eine beobachtete Häufigkeit einer „Zelle“ größer als die erwartete, wird die Proportion des Überschusses blau dargestellt. Ist dagegen die beobachtete Häufigkeit kleiner als die erwartete, wird die Proportion des Fehlers rot gezeichnet. In unserem Beispiel werden in der Zelle oben links deutlich mehr erwachsene Frauen in der ersten Klasse erwartet als tatsächlich auf dem Schiff waren. Auch leere Zellen, wie wir sie bei Kindern in der Crew haben, werden nun entsprechend der erwarteten Werte dargestellt. Dieser Mosaicplot ist durchgängig gleichmäßig partitioniert. Wären die beobachteten gleich den erwarteten Werten, ergäbe sich genau dieser Mosaicplot.



Die Darstellung in einer Kontingenztabelle bietet die Möglichkeit herauszufinden, ob es Assoziationen zwischen verschiedenen Merkmalen gibt. Ein derartiges Assoziationsmaß sollte auf das Intervall $[-1, 1]$ beschränkt sein, wobei der Wert 0 die Unabhängigkeit beider Variablen, -1 einen perfekt negativen und 1 einen perfekt positiven Zusammenhang charakterisieren sollte. Werte zwischen 0 und ± 1 deuten dann auf einen mehr oder weniger starken, jedoch keinen perfekten Zusammenhang hin. Die Richtung des Zusammenhangs (negativ, positiv) ist nur bei mindestens ordinal skalierten Variablen sinnvoll interpretierbar. Einige Assoziationsmaße für nominal skalierte Variablen verwenden daher nur das Intervall $[0, 1]$ und werden auch als richtungslose Assoziationsmaße bezeichnet. Wir betrachten zunächst zwei qualitative Merkmale und folgende Kontingenztafel:

		Merkmal X_2			
		$\tilde{c}_1$	$\dots$	$\tilde{c}_{m_2}$	
Merkmal X_1	c_1	n_{11}	$\dots$	n_{1m_2}	$n_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	c_{m_1}	$n_{m_1 1}$	$\dots$	$n_{m_1 m_2}$	$n_{m_1 \cdot}$
		$n_{\cdot 1}$	$\dots$	$n_{\cdot m_2}$	n

Die Grundlage eines Assoziationsmaßes für zwei qualitative Merkmale stellt die χ^2-Größe dar. Seien in einer Kontingenztabelle bestehend aus zwei qualitativen Merkmalen X_1 und X_2 für $i = 1, \dots, m_1$ und $j = 1, \dots, m_2$ n_{ij} die absoluten Häufigkeiten für die Anzahl Fälle in Kategorie c_i des Merkmals X_1 und in Kategorie $\tilde{c}_j$ des Merkmals X_2 bzw. h_{ij} die relativen Häufigkeiten für die Proportion in Kategorie c_i des Merkmals X_1 und in Kategorie $\tilde{c}_j$ des Merkmals X_2. Die Zeilensumme der i-ten Zeile wird mit $n_{i\cdot} = n_{i1} + \dots + n_{im_2}$ für alle $i = 1, \dots, m_1$ und die Spaltensumme der j-ten Spalte wird mit $n_{\cdot j} = n_{1j} + \dots + n_{m_1 j}$ für alle $j = 1, \dots, m_2$ bezeichnet. Zeilen- und Spaltensummen werden als Randsummen bezeichnet. Entsprechendes gilt für die relativen Häufigkeiten und die Zeilen- und Spaltensummen, sie heißen Randhäufigkeiten.

Ein wichtiger Begriff ist die bedingte Häufigkeitsverteilung. Wir bilden die bedingte Häufigkeit für eine Merkmalsausprägung eines Merkmals unter der Bedingung, dass für ein anderes Merkmal eine feste Merkmalsausprägung eintritt. Dann heißt

$$h_{c_i|\tilde{c}_j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{h_{ij}}{h_{\cdot j}} \tag{4.4}$$

für $n_{\cdot j} > 0$ und $i = 1, \ldots, m_1$ bedingte Häufigkeit von $X_1 = c_i$ unter der Bedingung $X_2 = \tilde{c}_j$. Weiter heißt

$$h_{\tilde{c}_j|c_i} = \frac{n_{ij}}{n_{i\cdot}} = \frac{h_{ij}}{h_{i\cdot}} \tag{4.5}$$

für $n_{i\cdot} > 0$ und $j = 1, \ldots, m_2$ bedingte Häufigkeit von $X_2 = \tilde{c}_j$ unter der Bedingung $X_1 = c_i$. Somit kann eine bedingte Häufigkeitsverteilung für ein Merkmal unter der Bedingung einer festgelegten Merkmalsausprägung eines anderen Merkmals bestimmt werden.

Beispiel 4.11

Die bedingte Häufigkeit von CarHire=No unter der Bedingung Main.Purpose=2 ist $\frac{2}{3}$. Die bedingte Häufigkeit von Main.Purpose=2 unter der Bedingung CarHire=No ist $\frac{2}{22}$. Die bedingte Häufigkeitsverteilung für CarHire unter der Bedingung Main.Purpose=2 lautet $h_{\text{No}|2} = \frac{2}{3}, h_{\text{Yes}|2} = \frac{1}{3}$. Die bedingten relativen Häufigkeiten summieren sich wieder zu 1.

Es gilt

$$\begin{aligned} 1 &= \sum_{i=1}^{m_1} h_{c_i|\tilde{c}_j} = \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{\cdot j}}{n_{\cdot j}} \\ &= \sum_{j=1}^{m_2} h_{\tilde{c}_j|c_i} = \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{i\cdot}}{n_{i\cdot}}. \end{aligned}$$

Wir definieren nun die χ^2-Größe, die uns die Grundlage für verschiedene Assoziationsmaße liefert. Hierzu nutzen wir die Randhäufigkeiten. Nehmen wir an, dass die beiden Merkmale empirisch unabhängig voneinander sind, so lassen sich die bedingten Häufigkeiten dafür, dass $X_1 = c_i$ und $X_2 = \tilde{c}_j$ gilt, multiplizieren und wir erhalten dann eine erwartete Häufigkeit (absolut)

$$\nu_{ij} = n \cdot \frac{n_{i\cdot} n_{\cdot j}}{n^2} = \frac{n_{i\cdot} n_{\cdot j}}{n} \tag{4.6}$$

für $X_1 = c_i$ und $X_2 = \tilde{c}_j$. Hieraus lässt sich eine Testgröße bestimmen, indem die normierten quadrierten Abstände der tatsächlichen Häufigkeiten zu den erwarteten Häufigkeiten aufsummiert werden.

Definition 4.12: χ^2-Größe

Die χ^2-Größe zweier qualitativer Merkmale X_1 und X_2 mit m_1 bzw. m_2 Kategorien

und n Realisierungen lautet

$$\chi^2 := \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{(n_{ij}-\nu_{ij})^2}{\nu_{ij}}, \tag{4.7}$$

wobei ν_{ij} die erwarteten Häufigkeiten gemäß Gleichung (4.6) sind.

Bemerkung.
Wir schließen den Fall einer Randhäufigkeit von Null aus, da dann die jeweilige Kategorie nicht berücksichtigt werden müsste und sie weggelassen wird.

Bei empirischer Unabhängigkeit der beiden Merkmale stimmen die beobachteten mit den erwarteten Häufigkeiten überein. Dann ist der χ^2-Wert Null. Auf der anderen Seite ist der χ^2-Wert nach oben beschränkt. Allgemein gilt zunächst

$$\begin{aligned}
\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{(n_{ij}-\nu_{ij})^2}{\nu_{ij}} &= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \frac{n_{ij}^2 - 2n_{ij}\nu_{ij} + \nu_{ij}^2}{\nu_{ij}} \\
&= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \left(\frac{n_{ij}^2}{\nu_{ij}} - 2n_{ij} + \nu_{ij}\right) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \left(\frac{n_{ij}^2}{\nu_{ij}}\right) - n \\
&= n\left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \left(\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}\right) - 1\right).
\end{aligned} \tag{4.8}$$

Um den χ^2-Wert nach oben abzuschätzen, müssen wir den Fall bestimmen, bei welchem die Merkmale vollkommen abhängig sind. Zwei Merkmale sind vollkommen abhängig, wenn in einer Kontingenztabelle

- für $m_1 > m_2$ in jeder Zeile die Häufigkeiten in genau einem Feld konzentriert sind, $h_{\tilde{c}_j|c_i} = \frac{n_{ij}}{n_{i\cdot}}$ ist 1 für genau ein j ansonsten 0,
- für $m_1 = m_2$ in jeder Zeile und Spalte die Häufigkeiten in genau einem Feld konzentriert sind, $h_{\tilde{c}_j|c_i} = h_{c_i|\tilde{c}_j} = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{ij}}{n_{\cdot j}}$ ist 1 für genau ein j bzw. i ansonsten 0, bzw.
- für $m_1 < m_2$ in jeder Spalte die Häufigkeiten in genau einem Feld konzentriert sind, $h_{c_i|\tilde{c}_j} = \frac{n_{ij}}{n_{\cdot j}}$ ist 1 für genau ein i ansonsten 0.

Die Gleichung (4.8) vereinfacht sich zu

$$\chi^2 = n \cdot (\min\{m_1, m_2\} - 1).$$

Insgesamt haben wir gezeigt, dass

$$0 \leq \chi^2 \leq n \cdot (\min\{m_1, m_2\} - 1).$$

gilt und damit ist auch zu sehen, dass in Abhängigkeit von der Anzahl der Realisierungen n der χ^2-Wert unbeschränkt ansteigt. Das ist jedoch im Hinblick auf ein Assoziationsmaß problematisch, da stets zuerst die obere Schranke bestimmt werden muss. Deswegen wird der χ^2 selbst nicht als Maß verwendet, sondern daraus werden Maße konstruiert. Wir definieren

Definition 4.13: Kontingenzkoeffizient nach Pearson

Der Kontingenzkoeffizient C nach Pearson ist definiert als

$$C := \sqrt{\frac{\chi^2}{n + \chi^2}}. \tag{4.9}$$

Setzen wir die untere bzw. obere Schranke für den χ^2-Wert ein, so ist der Kontingenzkoeffizient C beschränkt durch

$$0 \leq C \leq \sqrt{\frac{n \cdot (\min\{m_1, m_2\} - 1)}{n + n \cdot (\min\{m_1, m_2\} - 1)}} = \sqrt{\frac{\min\{m_1, m_2\} - 1}{\min\{m_1, m_2\}}} < 1.$$

Ein noch vorhandenes Problem besteht darin, dass der Koeffizient von der Anzahl der Kategorien abhängt. Auch das Problem können wir durch eine Normierung beseitigen.

Definition 4.14: Korrigierter Kontingenzkoeffizient nach Pearson

Der korrigierte Kontingenzkoeffizient nach Pearson ist definiert als

$$C_* := C\sqrt{\frac{\min\{m_1, m_2\}}{\min\{m_1, m_2\} - 1}}. \tag{4.10}$$

Es ist offensichtlich, dass $0 \leq C_* \leq 1$ gilt. Dabei liegt völlige empirische Unabhängigkeit für $C_* = 0$ und völlige empirische Abhängigkeit für $C_* = 1$ vor. Der korrigierte Kontingenkoeffizient C_* ist ungerichtet.

Beispiel 4.15

		Sex		
		Female	Male	
Survived	No	126 (*318.2*)	1364 (*1171.8*)	1490
	Yes	344 (*151.8*)	367 (*559.2*)	711
		470	1731	2201

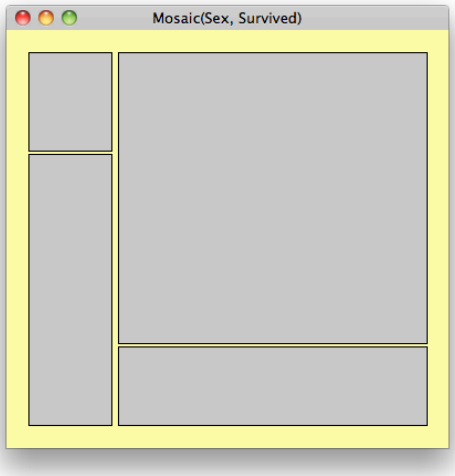
Für die beiden qualitativen Merkmale Sex und Survived ergibt sich die aufgeführte Kontingenztabelle aus den Beobachtungswerten, in Klammern stehen die erwarteten Werte. Aus $\chi^2 = 456.9$ ergeben sich die Kontingenzkoeffizienten $C = 0.415$ und $C_* = 0.586$. Das lässt auf eine vorhandene empirische Abhängigkeit schließen.

Auch in R gibt es die Möglichkeit, den Pearsonschen-Kontingenzkoeffizienten zu bestimmen. Dazu benötigen wir die Library vcd und den Befehl assocstats.

```
library(vcd)
assocstats(ftable(Sex, Survived))
```

```
                    X^2    df   P(> X^2)
Likelihood Ratio   434.47  1    0
Pearson            456.87  1    0
Phi-Coefficient   : 0.456
Contingency Coeff.: 0.415
Cramers V         : 0.456
```

Wir können $\chi = 456.87$ (Pearson) und $C = 0.415$ (Contingency Coeff.) ablesen und damit $C_* = 0.586$ (siehe Beispiel 4.15) bestimmen. Dies können wir auch graphisch erkennen, indem wir uns den Mosaicplot der beiden Merkmale ansehen. Unabhängigkeit würde bedeuten, dass die Fläche des Quadrats so in Rechtecksflächen aufgeteilt würde, dass jede Rechtecksfläche das Produkt der dazugehörigen Randhäufigkeiten darstellt. Es ergäben sich durchgezogene Linien innerhalb des Quadrats, da sich in jeder Seitenlänge die entsprechende Randhäufigkeit wiederspiegelte. Je weiter ein Mosaicplot davon abweicht, desto weniger ist diese Unabhängigkeitsannahme gegeben.



Das Beispiel zeigt deutlich die Abweichung von der Unabhängigkeitsannahme.

Bemerkung.
(I) Es sei noch einmal betont, dass eine Assoziation nichts über kausale Zusammenhänge zwischen Merkmalen aussagt.
(II) Bedingte Häufigkeiten und damit auch Kontingenzkoeffizienten lassen sich genauso für mehr als zwei Merkmale bestimmen. Für den korrigierten Kontingenzkoeffizienten wird entsprechend die minimale Kategorienanzahl aller Merkmale bestimmt.

Beispiel 4.16

Wir betrachten Daten zum Unglück der Titanic im Jahre 1912. Von 2201 Personen überlebten 711 den Untergang des Schiffs. Wir haben die qualitativen Merkmale

Class, Age, Sex und Survived zur Verfügung. Stellen wir die drei ersten Merkmale (etwa in dieser Reihenfolge) in einer Kontingenztabelle dar, so erhalten wir $\chi^2 = 523.740$, $C = 0.438$ und $C_* = 0.620$. Das ergibt einen relativ hohen Wert, der auf eine ausgeprägte empirische Abhängigkeit schließen lässt.

Teil II.

Zusammenhänge

5. Hauptkomponentenanalyse

Warum?

Werden viele Eigenschaften auf Basis der Wertemenge $\mathbb{R}$ erhoben, so kann versucht werden, die Information in den Daten, die als Summe aller einzelnen Varianzen (Totalvariation) angesetzt wird, so durch neue nicht erhobene Eigenschaften (latente Variablen) darzustellen, dass darunter latente Variablen sind, die nur einen kleinen Beitrag zur Totalvariation liefern und so vernachlässigbar sind. Das Ergebnis dieser Reduktion von Variablen muss zuletzt noch zusammen mit Experten untersucht werden. Während die ursprünglichen Eigenschaften aus der Erhebung heraus interpretierbar sind, müssen die verbliebenen latenten Variablen noch durch Beschreibung und Benennung als Eigenschaften interpretiert werden.

Sei $X \in \mathbb{R}^{n,k}$ eine Datenmatrix bestehend aus kardinalen Merkmalen $X_1, \ldots, X_k$. Eine orthogonale Transformation der Datenmatrix ändert gemäß (3.27) an der Totalvariation nichts. Werden durch eine orthogonale Transformation neue Variablen (so genannte latente Merkmale) $Y_1, \ldots, Y_k$ derart erzeugt, dass wenige von ihnen einen großen Anteil an der Totalvariation haben, wird die Totalvariation durch diese weitestgehend erklärt. Die verbleibenden Merkmale können so möglicherweise begründet weggelassen werden. Mit der Hauptkomponentenanalyse kann eine Variablenreduktion durchgeführt werden. Dabei soll auf einer Variablen ein möglichst großer Anteil an der Totalvariation der beobachteten Merkmale gesammelt werden. Werden die Variablen nach der Varianz absteigend geordnet, so erklären die ersten $p < k$ Variablen den größten Teil der Totalvariation. Hier ist auch der Unterschied zum allgemeinen Ansatz der so genannten Faktorenanalyse zu sehen, da lediglich die Varianzen und nicht die Kovarianzen betrachtet werden. Wir müssen folgende Fragen untersuchen:

- Welche orthogonale Transformation $Q \in \mathbb{R}^{k,k}$, $Y = XQ$ ermöglicht eine im beschriebenen Sinne bestmögliche Verteilung der Varianzen?
- Auf wie viele Variablen $1 \leq p < k$ kann reduziert werden?
- Wieviel der Varianz eines Merkmals X_i ist in der Variablen Y_j enthalten und welcher Anteil der Varianz von Y_j wird durch Varianzanteile aller X_i erklärt?

Die Hauptkomponentenanalyse bildet die Grundlage der Hauptkomponentenmethode. Zunächst jedoch betrachten wir das zentrale Element der Hauptkomponentenanalyse, die Hauptachsentransformation.

5.1. Hauptachsentransformation

Ist eine Datenmatrix X zentriert (wie etwa bei kleinen standardisierten Daten), d.h. $X = HX$, so gilt

$$S_X = \frac{1}{n-1} X^T H X = \frac{1}{n-1} X^T H^T H X = \frac{1}{n-1} (HX)^T HX = \frac{1}{n-1} X^T X. \tag{5.1}$$

Sei $X = HX$ zentriert. Die Hauptachsentransformation hat zum Ziel, die Spaltenvektoren $\mathbf{X}_1, \ldots, \mathbf{X}_k$ von X der zentrierten Datenmatrix so orthogonal auf Spaltenvektoren $\mathbf{Y}_1, \ldots, \mathbf{Y_k} \in \mathbb{R}^n$ zu transformieren,

$$Y = XQ, \quad Q \in \mathbb{R}^{k,k}, Q^T Q = I^{k,k},$$

dass $\mathbf{Y_j}^T \mathbf{Y_j} > \mathbf{Y_{j+1}}^T \mathbf{Y_{j+1}}$ gilt für alle $j = 1, \ldots, k-1$. Denn mit $X = HX$ folgt $Y = XQ = HXQ$, Y ist zentriert und so $(n-1)S_Y = Y^T Y$. $\mathbf{Y_1}_j^T \mathbf{Y_1}_j = (n-1) s^2_{\mathbf{Y_1}}$ ist proportional zur Varianz von $\mathbf{Y_j}$.

Seien $\mathbf{Q_1}, \ldots, \mathbf{Q_k} \in \mathbb{R}^k$ die Spalten von Q mit $\mathbf{Q_j}^T \mathbf{Q_j} = 1$ und $\mathbf{Q_j}^T \mathbf{Q_l} = 0$ für $j \neq l$. Wir betrachten zunächst Y_1 und suchen ein $\mathbf{Q_1}$ so, dass $\frac{1}{n-1} \mathbf{Q_1}^T X^T X \mathbf{Q_1}$ maximal wird,

$$\begin{aligned} &\max \left\{ \frac{1}{n-1} \mathbf{Y_1}^T \mathbf{Y_1} = \frac{1}{n-1} \mathbf{Q_1}^T X^T X \mathbf{Q_1} = \mathbf{Q_1}^T S_X \mathbf{Q_1} \right\} \\ &\text{unter } \mathbf{Q_1}^T \mathbf{Q_1} = 1 \end{aligned} \tag{5.2}$$

In Satz und Definition 4.7 haben wir eine Möglichkeit gesehen, Optimierungsprobleme mit Nebenbedingungen zu lösen. Diese funktioniert hier jedoch nicht, da die Nebenbedingung nicht affin-linear in $\mathbf{Q_1}$ ist. Auf ähnliche Weise jedoch lässt sich das Problem lösen:

Satz 5.1: Lagrange-Methode

Seien $f, h_j : \Gamma \to \mathbb{R}$, $j = 1, \ldots, l$ stetig differenzierbare Funktionen. Sei **x** ein lokales Extremum von f unter den Nebenbedingungen $h_j(\mathbf{x}) = 0$. Seien $\{h_1(\mathbf{x}), \ldots, h_l(\mathbf{x})\}$ linear unabhängig. Dann existieren $v_j \in \mathbb{R}$ mit

$$\nabla f(\mathbf{x}) = -\sum_{j=1}^{l} v_j \nabla h_j(\mathbf{x}).$$

Wir setzen die Lagrangefunktion

$$\Theta(\mathbf{v}, \mathbf{x}) = f(\mathbf{x}) + \sum_{j=1}^{l} v_j h_j(\mathbf{x}).$$

an und bestimmen Kandidaten für Extrema durch Nullsetzen der ersten Gleichung der KKT-Bedingungen gemäß (4.3), d.h. Ermitteln der stationären Punkte der Lagrange-Funktion. Das ist jedoch nur ein notwendiges Kriterium. Ist f zweimal stetig differenzierbar und $(\mathbf{v}, \mathbf{x})$ ein solcher stationärer Punkt, und ist $\Theta_{\mathbf{v}}(\mathbf{x})$ für festes **v** konvex (konkav), so stellt **x** ein globales Minimum (Maximum) des Problems mit Nebenbedingungen dar.

Definition 5.2: Konvexität und Konkavität

Sei $f : \Gamma \to \mathbb{R}$ zweimal stetig differenzierbar. f heißt konvex (konkav) auf Γ, falls alle Eigenwerte der Hessematrix $H_f(\mathbf{x})$ für alle $\mathbf{x} \in \Gamma$ größer (kleiner) oder gleich 0 sind.

Nun versuchen wir, unser Problem zu lösen.

Satz 5.3

Ist $\mathbf{Q_1}$ ein normierter Eigenvektor zum größten Eigenwert λ_1 von S_X, so löst $\mathbf{Q_1}$ das Problem (5.2). Dann gilt $\frac{1}{n-1}\mathbf{Y_1}^T\mathbf{Y_1} = \lambda_1$.

Beweis.

Die Lagrangefunktion des Problems lautet

$$\Theta(v, \mathbf{Q_1}) = \mathbf{Q_1}^T S_X \mathbf{Q_1} + v(\mathbf{Q_1}^T\mathbf{Q_1} - 1).$$

Die stationären Punkte von Θ erhalten wir über

$$\nabla\Theta(v, \mathbf{Q_1}) = \begin{pmatrix} \mathbf{Q_1}^T\mathbf{Q_1} - 1 \\ 2S_X\mathbf{Q_1} + 2v\mathbf{Q_1} \end{pmatrix} \stackrel{!}{=} \mathbf{0} \Leftrightarrow S_X\mathbf{Q_1} = -v\mathbf{Q_1},\ \mathbf{Q_1}^T\mathbf{Q_1} = 1, \mathbf{Q_1} \neq \mathbf{0}.$$

Damit muss $-v$ Eigenwert von S_X zum (normierten) Eigenvektor $\mathbf{Q_1}$ sein. Da $S_X = \frac{1}{n-1}X^TX$ positiv semidefinit ist, gilt $v \leq 0$.

Die Hessematrix der Lagrangefunktion in Abhängigkeit von $\mathbf{Q_1}$ bei festem v lautet

$$H_{\Theta_v(\mathbf{Q_1})} = 2(S_X + vI^{k,k}).$$

Sei $P \in \mathbb{R}^{k,k}$ eine S_X diagonalisierende Matrix, d.h. $P^TS_XP = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Sei weiter o.B.d.A $\lambda_j \geq \lambda_{j+1}$ für alle $j = 1, \ldots, k-1$. Dann gilt mit $-v := \lambda_1 = \max\{\lambda_1, \ldots, \lambda_k\}$

$$S_X + vI^{k,k} = P\Lambda P^T - \lambda_1 P I^{k,k} P^T = P(\Lambda - \lambda_1 I^{k,k})P^T.$$

Die Matrix $S_X + vI^{k,k}$ ist ähnlich zur Matrix $\Lambda - \lambda_1 I^{k,k}$. Letztere besitzt auf der Diagonalen eine Null und sonst nur nicht-positive Einträge und ist damit negativ semidefinit. Also ist $\Theta_{\mathbf{v}}(\mathbf{Q_1})$ konkav, es liegt somit ein globales Maximum vor. Es gilt

$$\frac{1}{n-1}\mathbf{Y_1}^T\mathbf{Y_1} = \frac{1}{n-1}\mathbf{Q_1}^TX^TX\mathbf{Q_1} = \mathbf{Q_1}^TS_X\mathbf{Q_1} = -\mathbf{Q_1}^Tv\mathbf{Q_1} = \lambda_1.$$

□

Wir müssen nun die weiteren $k-1$ Spalten von Q bestimmen. Dazu betrachten wir für

$1 < j \leq k$ folgende Optimierungsprobleme:

$$\begin{array}{ll} \max & \left\{ \frac{1}{n-1} \mathbf{Y_j}^T \mathbf{Y_j} = \frac{1}{n-1} \mathbf{Q_j}^T X^T X \mathbf{Q_j} = \mathbf{Q_j}^T S_X \mathbf{Q_j} \right\} \\ \text{unter} & \mathbf{Q_j}^T \mathbf{Q_1} = 0 \\ & \mathbf{Q_j}^T \mathbf{Q_2} = 0 \\ & \vdots \\ & \mathbf{Q_j}^T \mathbf{Q_{j-1}} = 0 \\ & \mathbf{Q_j}^T \mathbf{Q_j} = 1 \end{array}$$

Um sie lösen zu können überlegen wir uns zunächst, dass

$$\mathbf{Q_j}^T (S_X - \lambda_l I^{k,k}) \mathbf{Q_l} = \mathbf{Q_j}^T S_X \mathbf{Q_l} - \mathbf{Q_j}^T \lambda_l \mathbf{Q_l} = \mathbf{Q_j}^T S_X \mathbf{Q_l} = \mathbf{Q_l}^T S_X \mathbf{Q_j} = 0 \tag{5.3}$$

für alle $l = 1, \ldots, j-1$ gilt.

Satz 5.4

Ist $\mathbf{Q_j}$ ein normierter Eigenvektor zum Eigenwert λ_j von S_X, so löst $\mathbf{Q_j}$ das Problem (5.3) bei gegebenem $\mathbf{Q_1}, \ldots, \mathbf{Q_{j-1}}$. Dann gilt $\frac{1}{n-1} \mathbf{Y_j}^T \mathbf{Y_j} = \lambda_j$.

Beweis.

Die Lagrangefunktion des Problems lautet

$$\Theta(\mathbf{v}, \mathbf{Q_j}) = \mathbf{Q_j}^T S_X \mathbf{Q_j} + \sum_{l=1}^{j-1} v_l \mathbf{Q_j}^T \mathbf{Q_l} + v_j (\mathbf{Q_j}^T \mathbf{Q_j} - 1).$$

Die stationären Punkte von Θ erhalten wir über

$$\nabla \Theta(\mathbf{v}, \mathbf{Q_j}) = \begin{pmatrix} \mathbf{Q_j}^T \mathbf{Q_1} \\ \vdots \\ \mathbf{Q_j}^T \mathbf{Q_{j-1}} \\ \mathbf{Q_j}^T \mathbf{Q_j} - 1 \\ 2 S_X \mathbf{Q_j} + \sum_{l=1}^{j-1} v_l \mathbf{Q_l} + 2 v_j \mathbf{Q_j} \end{pmatrix} \stackrel{!}{=} \mathbf{0}.$$

Wir multiplizieren die letzte Zeile des Gradienten mit $\mathbf{Q_l}$ für $l = 1, \ldots, j-1$:

$$\mathbf{Q_l}^T (2 S_X \mathbf{Q_j} + \sum_{l=1}^{j-1} v_l \mathbf{Q_l} + 2 v_j \mathbf{Q_j}) = 2 \underbrace{\mathbf{Q_l}^T S_X \mathbf{Q_j}}_{=0} + \underbrace{\sum_{a=1}^{j-1} v_a \mathbf{Q_l}^T \mathbf{Q_a}}_{=v_l} + 2 v_j \underbrace{\mathbf{Q_l}^T \mathbf{Q_j}}_{=0}.$$

Der Ausdruck ist Null genau dann wenn $v_l = 0$ ist. Es bleibt damit

$$S_X \mathbf{Q_j} + v_j \mathbf{Q_j} \stackrel{!}{=} \mathbf{0}$$

übrig. Dies gilt genau dann wenn $-v_j$ Eigenwert von S_X zum Eigenvektor $\mathbf{Q_j}$ ist. Wegen

$$\frac{1}{n-1}\mathbf{Y_j}^T\mathbf{Y_j} = \frac{1}{n-1}\mathbf{Q_j}^T X^T X\mathbf{Q_j} = \mathbf{Q_j}^T S_X\mathbf{Q_j} = -\mathbf{Q_j}^T v_j\mathbf{Q_j} = -v_j$$

wird die Varianz am größten, wenn $-v_j = \lambda_j$ (dem größten verbliebenen Eigenwert) entspricht.

□

Die Sätze 5.3 und 5.4 erzeugen eine Matrix Q, welche die empirische Varianz-Kovarianz-Matrix S_X diagonalisiert, wobei Λ die Diagonalmatrix bestehend aus den Eigenwerten von S_X sei:

$$\begin{aligned} S_Y &= \frac{1}{n-1}Y^TY = \frac{1}{n-1}(XQ)^TXQ = \frac{1}{n-1}Q^TX^TXQ = Q^TS_XQ = \Lambda, \\ R_Y &= D_Y^{-\frac{1}{2}}S_YD_Y^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}}\Lambda\Lambda^{-\frac{1}{2}} = I^{k,k}. \end{aligned}$$

Die Spalten $\mathbf{Y_j}$ von Y sind orthogonal und unkorreliert. Der Vorgang der Diagonalisierung der empirischen Varianz-Kovarianz-Matrix wird Hauptachsentransformation genannt, die $\mathbf{Y_j}$ heißen Hauptachsen.

Beispiel 5.5: I

Gegeben sei die Datenmatrix

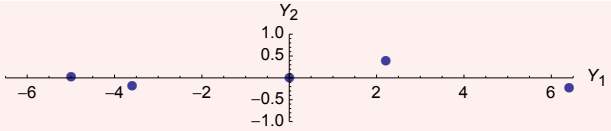
$$\tilde{X} = \begin{pmatrix} 22 & 5 \\ 25 & 10 \\ 21 & 4 \\ 28 & 13 \\ 24 & 8 \end{pmatrix} \in \mathbb{R}^{5,2} \text{ mit } \bar{\mathbf{x}} = (24, 8) \text{ und } S_{\tilde{X}} = \begin{pmatrix} 7.5 & 10 \\ 10 & 13.5 \end{pmatrix} \in \mathbb{R}^{2,2}.$$

Wir erhalten die zentrierten Daten

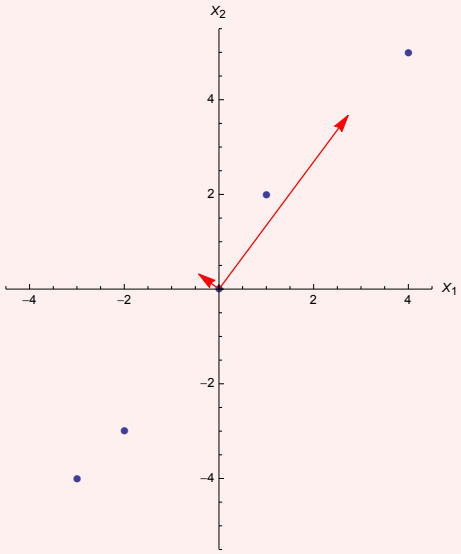
$$X = \begin{pmatrix} -2 & -2 \\ 1 & 2 \\ -3 & -4 \\ 4 & 5 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{5,2} \in \mathbb{R}^{5,2}, \quad S_X = S_{\tilde{X}}, R_X = \begin{pmatrix} 1.000 & 0.994 \\ 0.994 & 1.000 \end{pmatrix} \in \mathbb{R}^{2,2}.$$

Die Eigenwerte von X sind $\lambda_1 = 20.94$ und $\lambda_2 = 0.06$ und die dazugehörigen normierten Eigenvektoren ergeben sich zu $\mathbf{Q_1} = (0.597, 0.802)^T$ und $\mathbf{Q_2} = (-0.802, 0.597)^T$. Wir erhalten Y durch

$$Y = XQ = \begin{pmatrix} -3.601 & -0.186 \\ 2.202 & 0.392 \\ -5.000 & 0.019 \\ 6.399 & -0.225 \\ 0 & 0 \end{pmatrix}.$$



Zeichen wir die gewichteten Eigenvektoren in einen Scatterplot der zentrierten Daten,



erkennen wir die Bedeutung der ersten Komponente. Man kann anhand der Abbildung sehen, dass die erste Koordinate im wesentlichen genügt, um die Daten zu erklären. Eine Reduktion der beiden Variablen auf eine erscheint sinnvoll.

```
x=c(22,25,21,28,24)
y=c(5,10,4,13,8)
X=cbind(x-mean(x),y-mean(y))
cov(X)
cor(X)
p=eigen(cov(X))
p$vectors
Y=X%*%p$vectors
X%*%p$vectors
plot(Y,pch=19)
```

Wegen der Orthogonalität von Q ist $X = YQ^T$. Seien nun $F := Y\Lambda^{-\frac{1}{2}}$ und $L := Q\Lambda^{\frac{1}{2}}$, $L = (l_{ij})$. Die Spalten $\mathbf{F_j}$ von F sind normiert und werden Hauptkomponenten genannt. Es gilt

$$X = YQ^T = F\Lambda^{\frac{1}{2}}Q^T = F\Lambda^{\frac{1}{2}}\Lambda^{-\frac{1}{2}}L^T = FL^T. \tag{5.4}$$

Weiter ist

$$\begin{aligned} F^TF &= \Lambda^{-\frac{1}{2}}Y^TY\Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}}Q^TX^TXQ\Lambda^{-\frac{1}{2}} = (n-1)I^{k,k}, \\ L^TL &= \Lambda^{\frac{1}{2}}Q^TQ\Lambda^{\frac{1}{2}} = \Lambda \qquad (5.5) \\ LL^T &= Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q^T = Q\Lambda Q^T = S_X. \qquad (5.6) \end{aligned}$$

```
F=t(t(Y)/sqrt(p$values))
t(F)%*%F
```

5.2. Hauptkomponentenmethode

Wir gehen nun der Frage nach, auf wie viele Hauptkomponenten reduziert werden soll. Allgemein benötigen wir Kriterien, mit denen eine günstige Auswahl von p aus k Hauptkomponenten erfolgen kann. An dieser Stelle betrachten wir zwei solcher Kriterien: Das Kaiser-Kriterium und den Scree-Test.

Die Totalvariation gemäß (3.25) für X ist

$$s_t^2 = \text{sp}(S_X) = \text{sp}(Q\Lambda Q^T) = \text{sp}(\Lambda Q^T Q) = \text{sp}(\Lambda) = \sum_{i=1}^{k} \lambda_i,$$

wenn $\lambda_j,\ j = 1, \ldots, k$, die Eigenwerte von S_X sind. Dies wird als Informationsmaß bzgl. der Streuung der Daten interpretiert. Durch die Auswahl von p Hauptkomponenten wird $\frac{\lambda_1+\ldots+\lambda_p}{\sum_{i=1}^{k}\lambda_i}$ der Totalvariation erklärt. Der empirische Mittelwert

$$\bar{\lambda} = \frac{1}{k}\sum_{i=1}^{k} \lambda_i$$

liefert einen durchschnittlichen Varianzbeitrag für eine Hauptkomponente. Das Kaiserkriterium besagt, dass jede Hauptkomponente $\mathbf{Y_j}$, deren Varianzbeitrag λ_j größer oder gleich $\bar{\lambda}$ ist, einen überdurchschnittlichen Beitrag zur Totalvariation liefert und somit wichtig ist. Diese Grenze wird manchmal durch einen Faktor (oft: 0.7) nach unten korrigiert.

Beim Scree-Test wird das so genannte Ellbogenkriterium angewendet. Dabei werden die der Größe nach geordneten Eigenwerte $\lambda_1 \geq \ldots \geq \lambda_k$ in einem Lineplot (Screeplot, Scatterplot mit verbundenen Punkten) gegen ihren Index aufgetragen und der Eigenwert, bei dem ein deutlicher Knick dahingehend erkennbar ist, dass die Eigenwerte vor dem Knick relativ groß und die Eigenwerte nach dem Knick gleichbleibend klein sind, wird als erster weggelassen.

Beispiel 5.6: II

Im Beispiel 5.5 wählen wir eine Hauptkomponente aus: Es ist $s_t^2 = 21$, $\bar{\lambda} = 10.5$. Mit $\lambda_1 = 20.94$ entfallen auf diese Hauptkomponente 99.7% der Totalvariation.

Warum führen wir überhaupt Hauptkomponenten ein und begnügen uns nicht mit Hauptachsen? Dazu überlegen wir, wieviel der Varianz eines Merkmals in einer Hauptkomponente enthalten ist bzw. umgekehrt wie viel Varianz eines Merkmals durch gewählte Hauptkomponenten erklärt wird. Die Merkmale $\mathbf{X_i}$ lassen sich als Linearkombination

$$\mathbf{X_i} = Y\mathbf{Q_i}.^T = \sum_{j=1}^{k} q_{ij}\mathbf{Y_j} = F\mathbf{L_i}.^T = \sum_{j=1}^{k} l_{ij}\mathbf{F_j}$$

darstellen. Die l_{ij} heißen Faktorladungen und man sagt, dass die Hauptkomponente $\mathbf{F_j}$ mit l_{ij} auf das Merkmal $\mathbf{X_i}$ lädt. Die l_{ij} besitzen nun eine wichtige Eigenschaft. Mit $S_X = LL^T$ gemäß (5.6) sind die Varianzen der Merkmale $\mathbf{X_i}$ durch

$$s_{X_i}^2 = \sum_{j=1}^{k} l_{ij}^2$$

gegeben. Werden p Hauptkomponenten ausgewählt, so entsteht ein Fehler $\mathbf{E_i}$ durch

$$\mathbf{X_i} = \underbrace{\sum_{j=1}^{p} l_{ij}\mathbf{F_j}}_{\mathbf{X_i^{(p)}}} + \underbrace{\sum_{j=p+1}^{k} l_{ij}\mathbf{F_j}}_{\mathbf{E_i}}.$$

Damit wird nicht mehr die gesamte Varianz von $\mathbf{X_i}$ erfasst, sondern lediglich

$$h_i^2 := \sum_{j=1}^{p} l_{ij}^2 \leq s_{X_i}^2.$$

Die Zahl h_i^2 wird als Kommunalität des Merkmals $\mathbf{X_i}$ bezeichnet und gibt den Anteil der vom i-ten Merkmal $\mathbf{X_i}$ auf die ersten p Hauptkomponenten übertragenen Varianz-Information an.

Beispiel 5.7: III

Im Beispiel 5.6 wählten wir eine Hauptkomponente aus. Mit dieser wird 99.5% der Varianz des ersten bzw. 99.8% der Varianz des zweiten Merkmals erklärt.

```
L=t(t(p$vectors)*sqrt(p$values))
L%*%t(L)
t(L)%*%L
L[1,1]%*%L[1,1]
L[2,1]%*%L[2,1]
L[1,1]*L[1,1]/(L[1,]%*%L[1,])
L[2,1]*L[2,1]/(L[2,]%*%L[2,])
```

5.3. Spezialfall: Kleine standardisierte Daten

Anstelle der zentrierten Daten werden zur Durchführung der Hauptkomponentenanalyse manchmal auch die kleinen standardisierten Daten

$$Z = \frac{1}{\sqrt{n-1}} HXD_X^{-\frac{1}{2}},$$

$$Z = (z_{ij}) \in \mathbb{R}^{n,k},\ i = 1,\dots,n,\ j = 1,\dots,k, \quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j\sqrt{n-1}}, \tag{5.7}$$

mit

$$\begin{aligned} S_Z &= \frac{1}{n-1} R_X, \\ R_Z &= R_X = Z^T Z, \end{aligned}$$

benutzt. Die Hauptachsentransformation wird für $R_X = Z^T Z$ durchgeführt und entspricht damit einer Diagonalisierung der empirischen Varianz-Kovarianz-Matrix der standardisierten Daten W. Für $Y = WQ$ folgt damit

$$S_Y = \frac{1}{n-1} Q^T W^T W Q = Q^T R_X Q = \Lambda,$$

wobei Λ hier die Diagonalmatrix bestehend aus den Eigenwerten von R_X ist. Der Vorteil dabei ist, dass Merkmale mit überproportional starker Varianz im Vergleich zu anderen Merkmalen bei der Diagonalisierung von S_X einen starken Einfluss auf die Transformation hat. Sind alle Varianzen gleich, entfällt dies. Andererseits führt das zu anderen Ergebnissen, was bei der Interpretation berücksichtigt werden muss.

Beispiel 5.8: I

Gegeben sei wiederum die Datenmatrix

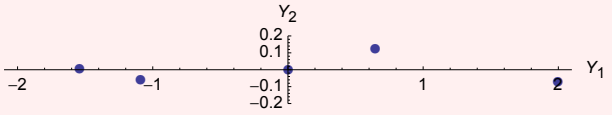
$$\tilde{X} = \begin{pmatrix} 22 & 5 \\ 25 & 10 \\ 21 & 4 \\ 28 & 13 \\ 24 & 8 \end{pmatrix} \in \mathbb{R}^{5,2} \text{ mit } \bar{\mathbf{x}} = (24, 8) \text{ und } S_{\tilde{X}} = \begin{pmatrix} 7.5 & 10 \\ 10 & 13.5 \end{pmatrix} \in \mathbb{R}^{2,2}.$$

Wir erhalten die kleinen standardisierten Daten

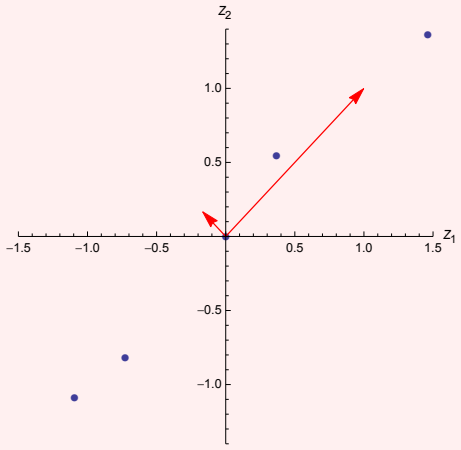
$$Z = \begin{pmatrix} -0.365 & -0.408 \\ 0.183 & 0.272 \\ -0.548 & -0.544 \\ 0.730 & 0.680 \\ 0.000 & 0.000 \end{pmatrix} \in \mathbb{R}^{5,2}, \quad 4 \cdot S_Z = R_X = Z^T Z = \begin{pmatrix} 1.000 & 0.994 \\ 0.994 & 1.000 \end{pmatrix} \in \mathbb{R}^{2,2}.$$

Die Eigenwerte von R_X sind $\lambda_1 = 1.994$ und $\lambda_2 = 0.006$ und die normierten Eigenvektoren dazu lauten $\mathbf{Q_1} = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})^T$ und $\mathbf{Q_2} = (\frac{-\sqrt{2}}{2}, \frac{\sqrt{2}}{2})^T$. Damit errechnet sich Y zu

$$Y = WQ = \begin{pmatrix} -1.094 & 0.061 \\ 0.643 & -0.127 \\ -1.544 & -0.005 \\ 1.995 & 0.071 \\ 0.000 & 0.000 \end{pmatrix}.$$



Zeichen wir die gewichteten Eigenvektoren in einen Scatterplot der standardisierten Daten,



erkennen wir wiederum die Bedeutung der ersten Komponente. Die unterschiedliche Skalierung der Achsen und die anderen Eigenvektoren im Vergleich zu Beispiel 5.5 zeigen den Einfluss unterschiedlicher Varianz in den Merkmalen.

Die Darstellung $Z = F\Lambda^{\frac{1}{2}}Q^T$ heißt Singulärwertzerlegung von Z, da dabei F eine orthogonale Matrix ist:

$$F^TF = \Lambda^{-\frac{1}{2}}Q^TZ^TZQ\Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}}Q^TR_XQ\Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}}\Lambda\Lambda^{-\frac{1}{2}} = I^{k,k}.$$

Damit bekommen wir

$$R_X = Z^TZ = LF^TFL^T = LL^T = T\Lambda T^T \text{ und } F = ZQ\Lambda^{-\frac{1}{2}} = ZQ\Lambda^{-1}Q^TQ\Lambda^{\frac{1}{2}} = ZR_X^{-1}L.$$

Satz und Definition 5.9: Varianzzerlegung

Bei der Hauptachsentransformation gilt $k - \lambda_1 - \ldots - \lambda_k = 0$. Sei für $p \leq k$ Faktoren $\tilde{R}_X := \tilde{Z}^T\tilde{Z}$ mit $\tilde{Z} = (\mathbf{F}_1 \ldots \mathbf{F}_p)(\mathbf{L}_1 \ldots \mathbf{L}_p)^T$, wobei $\mathbf{F_j}$ und $\mathbf{L_j}$ die Spalten der Matrizen F bzw. L sind. Dann ist die empirische Residuenkovarianzmatrix $\tilde{V} := R_X - \tilde{R}_X$ durch $\tilde{V} = R_X - \mathbf{L_1}\mathbf{L_1}^T - \ldots - \mathbf{L_p}\mathbf{L_p}^T$ bestimmt.

Beweis.

Zunächst ist die Spur von R_X gegeben durch

$$k = \text{sp}(R_X) = \text{sp}(Q^T\Lambda Q) = \text{sp}(\Lambda Q^TQ) = \text{sp}(\Lambda).$$

Also ist $k = \lambda_1 + \ldots + \lambda_k$. Weiter betrachten wir

$$Z = FL^T = (\mathbf{F_1}, \ldots, \mathbf{F_k})(\mathbf{L_1}, \ldots, \mathbf{L_k})^T = \mathbf{F_1}\mathbf{L_1}^T + \ldots + \mathbf{F_k}\mathbf{L_k}^T.$$

Werden $p \leq k$ Faktoren berechnet, so ergibt sich als Residuum (Fehler) $Z - \sum_{j=1}^{p} \mathbf{F_j}\mathbf{L_j}^T$.

Die zu erklärende empirische Kovarianzmatrix $R_X = Z^T Z$ wird durch p Faktoren angenähert durch

$$\begin{aligned} \tilde{R}_X := \tilde{Z}^T \tilde{Z} &= (\mathbf{L}_1 \dots \mathbf{L}_p)(\mathbf{F}_1 \dots \mathbf{F}_p)^T \cdot (\mathbf{F}_1 \dots \mathbf{F}_p)(\mathbf{L}_1 \dots \mathbf{L}_p)^T \\ &= (\mathbf{L}_1 \dots \mathbf{L}_p) \cdot (\mathbf{L}_1 \dots \mathbf{L}_p)^T \\ &= \sum_{j=1}^{p} (\mathbf{L_j}\mathbf{L_j}^T). \end{aligned} \tag{5.8}$$

Somit gilt $\tilde{V} := R_X - \tilde{R}_X = R_X - \sum_{j=1}^{p} (\mathbf{L_j}\mathbf{L_j}^T)$.

□

Mit der Zerlegung

$$Z = YQ^T = FL^T, \quad R_X = LL^T \tag{5.9}$$

haben wird das volle Modell beschrieben, es entsteht kein Fehler bei der Schätzung von R_X und somit ist $\tilde{V} = 0$. Wählen wir $p < k$ Hauptkomponenten aus, bezeichnen wir das mit $H^{(p)}, F^{(p)} \in \mathbb{R}^{n,p}$ bzw. $T^{(p)}, L^{(p)} \in \mathbb{R}^{k,p}$. Es ist dann mit $E := Z - H^{(p)}{T^{(p)}}^T = Z - F^{(p)}{L^{(p)}}^T$ und $\tilde{V} = L^{(p)}{L^{(p)}}^T$

$$Z = H^{(p)}{T^{(p)}}^T + E = F^{(p)}{L^{(p)}}^T + E \tag{5.10}$$

mit der „Grundgleichung“

$$R_X = Z^T Z = L^{(p)}{L^{(p)}}^T + \tilde{V} \tag{5.11}$$

und

$$E = Z - F^{(p)}{L^{(p)}}^T \tag{5.12}$$

$$\begin{aligned} \tilde{V} &= L^{(p)}{F^{(p)}}^T E + E^T F^{(p)}{L^{(p)}}^T + E^T E \\ &= L^{(p)}{F^{(p)}}^T (Z - F^{(p)}{L^{(p)}}^T) + (Z - F^{(p)}{L^{(p)}}^T)^T F^{(p)}{L^{(p)}}^T + E^T E \\ &= L^{(p)}{F^{(p)}}^T (FL^T - F^{(p)}{L^{(p)}}^T) + (LF^T - L^{(p)}{F^{(p)}}^T) F^{(p)}{L^{(p)}}^T + E^T E \\ &= \underbrace{L^{(p)}{F^{(p)}}^T (FL^T - F^{(p)}{L^{(p)}}^T)}_{0} + \underbrace{(LF^T - L^{(p)}{F^{(p)}}^T) F^{(p)}{L^{(p)}}^T}_{0} + E^T E \\ &= E^T E. \end{aligned} \tag{5.13}$$

Bemerkung.
(1) $\tilde{V} = E^T E$ ist im allgemeinen keine Diagonalmatrix.

Beispiel 5.10: II

Wählen wir in Fortsetzung des vorherigen Beispiels die erste Hauptkomponente aus,

so erhalten wir

$$F^{(1)} = \begin{pmatrix} -0.387 \\ 0.228 \\ -0.547 \\ 0.706 \\ 0.000 \end{pmatrix}, \quad L^{(1)} = \begin{pmatrix} 0.998 \\ 0.998 \end{pmatrix}.$$

Für die Residuen bekommen wir

$$E = \begin{pmatrix} 0.022 & -0.022 \\ -0.045 & 0.045 \\ -0.002 & 0.002 \\ 0.025 & -0.025 \\ 0 & 0 \end{pmatrix}, \quad \tilde{V} = \begin{pmatrix} 0.003 & -0.003 \\ -0.003 & 0.003 \end{pmatrix}.$$

Es wird 99.7% der Gesamtvarianz erklärt.

```
Z%*%p$vectors[,1]/sqrt(p$values[1])
L=p$vectors[,1]*sqrt(p$values[1])
L%*%t(L)
cor(Z)-L%*%t(L)
```

Im letzten Beispiel ist offensichtlich, dass es genügt, sich auf die erste Hauptkomponente zu beschränken, da damit die Gesamtvarianz nahezu erklärt werden kann.

Beispiel 5.11: Bilderkennung

Die Hauptkomponentenmethode kann in der Bilderkennung benutzt werden. Dabei werden Bilder (hier: 20 Stück, $170 \cdot 240$ Pixel) gleicher Dimension mit a Zeilen und b Spalten betrachtet. Die Farbinformation liege als 8-bit Grauwert vor. Dann kann ein Bild als Vektor des $\mathbb{R}^k$ mit $k = 170 \cdot 240$ aufgefasst werden. Liegen nun n Bilder vor ergibt sich eine Datenmatrix $X \in \mathbb{R}^{n,k}$ mit $n < k$.

Bei der Hauptkomponentenmethode benutzt man hierbei lediglich zentrierte, aber keine standardisierten Daten. Durch Verwendung von $p = 6$ Hauptkomponenten, was einer Reduktion der Datenmenge um 70% bedeutet, werden die Bilder wie folgt dargestellt.

Dabei wird die Matrix $F^{(p)} \in \mathbb{R}^{20,6}$ und die Matrix $L^{(p)} \in \mathbb{R}^{40800,6}$ gespeichert. Ein neues Bild wird zentriert und durch seine sechs Hauptkomponenten dargestellt. So wird das Bild von Fourier

durch

$$\tilde{h} = (-10451.3, -894.245, 1181.67, 974.06, 1842.38, -1173.06)^T$$

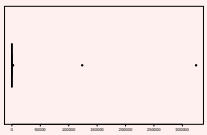
repräsentiert. Der „originale“ Fourier liegt als

$$h = (-11172.6, -342.18, 4978.22, 4462.49, 2097.13, -2510.3)^T$$

vor. Unter Verwendung der euklidischen Norm lässt sich die Repräsentation finden, zu der das neue Bild in seiner Repräsentation am nächsten ist. Für das Beispiel wird Fourier richtig „erkannt“. Je heterogener die Bilder sind, (Kopfgröße, Winkel, Kopfbedeckung,. . .) desto schlechter sind die Ergebnisse.

Beispiel 5.12

Wir betrachten die Wisconsin-Brustkrebsdaten. Bei den Merkmalen fällt auf, dass sich die minimale und maximale Varianz stark unterscheiden: Die kleinste empirische Varianz besitzt das Merkmal fracd.sd, die Standardabweichung des Merkmals, das die fraktale Dimension der Zellkerne beschreibt. Ein größerer Wert in diesem Merkmal deutet auf eine weniger reguläre Umrandung hin[a]. Die Varianz ist $7 \cdot 10^{-6}$. Die größte Varianz weist das Merkmal area.mv, die durchschnittliche Fläche der betrachteten Zellkerne, mit dem Wert 324167.4 auf. Ein Boxplot aller Varianzen zeigt zwei starke Ausreißer. Die Hauptachsentransformation bzgl. der zentrierten Daten wird die Unterschiede durch eine unterschiedlich starke Gewichtung berücksichtigen.

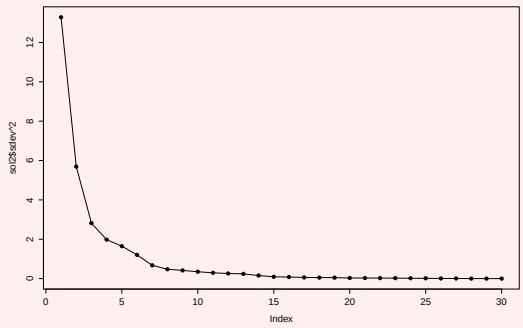


Die Totalvarianz ist $s_t^2 = 451896.6$. Durch eine Hauptkomponente wird 98.2% der Gesamtvarianz erklärt ($\lambda_1 = 443782.6$), mit zwei Hauptkomponenten 99.8%. Sowohl mit

dem Scree-Test als auch nach dem Kaiserkriterium wird eine Hauptkomponente gewählt. Hier zeigt sich der Einfluss der beiden Merkmale mit großer Varianz.

Wir führen die Hauptkomponentenanalyse mit den standardisierten Daten durch. Mit der Totalvarianz $s_t^2 = 30$ erhalten wir folgende Situation:

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
Eigenwert	13.28	5.69	2.82	1.98	1.65	1.21	0.68
Erkl. Var.	44.27%	63.24%	72.64%	79.24%	84.73%	88,76%	91.01%



Der Scree-Test zeigt kein eindeutiges Bild. Drei oder sechs Hauptkomponenten könnten hier angezeigt sein. Durch das Kaiserkriterium werden sechs Hauptkomponenten empfohlen, da die ersten sechs Eigenwerte größer als Eins sind.

[a] für eine Beschreibung siehe http://dollar.biz.uiowa.edu/~street/research/spie93.ps

6. Graphische Zusammenhangsanalyse

6.1. Biplots

Warum?

Die Reduktion von Eigenschaften auf latente Variablen auf Basis einer Hauptkomponentenanalyse kann visuell dargestellt werden. Das setzt zunächst voraus, dass es eine Reduktion auf höchstens drei latente Variablen gibt. Dann lassen sich die latenten Variablen in Form von Vektoren und die zu den Daten gehörenden Koordinaten aller Objekte als Punkte visualisieren und dementsprechend interpretieren. Doch nicht immer liegen die erhobenen Daten so vor, dass eine Hauptkomponentenanalyse möglich ist. Dennoch kann die eben angesprochene Visualisierung erzeugt werden. Dazu bedarf es jedoch einer speziellen Form von Daten: für jedes Objekt der Untersuchung muss die Nähe zu jedem anderen Objekt der Untersuchung (Distanz) bekannt sein. Auf Basis der Distanzen können die Objekte wiederum positioniert werden, d.h. es können Koordinaten in einem maximal dreidimensionalen Koordinatensystem bestimmt werden, welche die gegebenen Distanzen möglichst gut repräsentieren.

Mit Hilfe von Biplots werden sowohl Merkmalsträger als auch Merkmale in einer Graphik dargestellt. Ist $Y = XQ$ eine Hauptachsentransformation wie im letzten Kapitel, so gilt zunächst für Zeilenvektoren $\mathbf{y}_i, \mathbf{y}_l$ von Y und $\mathbf{x}_i, \mathbf{x}_l$ von X

$$\begin{aligned} ||\mathbf{y}_i - \mathbf{y}_l||^2 &= (\mathbf{y}_i - \mathbf{y}_l)(\mathbf{y}_i - \mathbf{y}_l)^T \\ &= (\mathbf{x}_i - \mathbf{x}_l)QQ^T(\mathbf{x}_i - \mathbf{x}_l)^T \\ &= (\mathbf{x}_i - \mathbf{x}_l)(\mathbf{x}_i - \mathbf{x}_l)^T. \end{aligned} \tag{6.1}$$

Die euklidischen Distanzen bleiben erhalten. Die Frage ist nun, wie gut sich diese Distanzen allein durch die ersten p Hauptkomponenten repräsentieren lassen. Wegen

$$S_Y = \frac{1}{n-1} Y^T Y = \Lambda$$

gilt $\sigma^2_{Y_j} = \lambda_j$ d.h.

$$\sum_{i=1}^{n} y_{ij}^2 = (n-1)\lambda_j.$$

Bei der komponentenweisen Berechnung des euklidischen Abstands d_{il} zweier Merkmalsträger ist $y_{ij} - y_{lj}$ zu bestimmen. Dies lässt sich nach oben abschätzen. Dazu betrachten wir das Maximierungsproblem

$$\begin{array}{ll} \max & \{(y_{ij} - y_{lj})^2\} \\ \text{unter} & y_{ij}^2 + y_{lj}^2 = (n-1)\lambda_j,\ j = p+1, \ldots, k, \\ & \sum_{j=p+1}^{k} (y_{ij} - y_{lj})^2 \leq d_{il}^2. \end{array}$$

Zur Lösung des Problems sind mehrere Fälle zu unterscheiden. Der Fall $y_{ij} = -y_{lj} = \pm\sqrt{\frac{n-1}{2}\lambda_j}$, falls $2(n-1)\lambda_j \leq d_{il}^2$ liefert eine worst-case-Abschätzung für die Werte. Dabei zeigt sich die Abhängigkeit von λ_j. Der Beitrag der letzten $k-p$ Hauptkomponenten zur euklidischen Distanz lässt sich hiermit durch

$$\sum_{j=p+1}^{k} (y_{ij} - y_{lj})^2 \leq 2(n-1) \sum_{j=p+1}^{k} \lambda_j$$

grob abschätzen. Wir wollen die Idee der Abhängigkeit von λ_j nutzen, verwenden jedoch anstelle der empirischen Varianz die empirische Standardabweichung s. In Beispiel 1.1 haben wir gesehen, dass knapp 70% der Werte der Standardnormalverteilung im Intervall $[-1, 1]$ liegen. Dies entspricht genau dem Intervall $[\bar{x} - s, \bar{x} + s]$. Als Richtschnur können wir damit das Intervall $[-\sqrt{\lambda_j}, \sqrt{\lambda_j}]$ angeben, in dem die meisten Werte der dazugehörigen Hauptkomponente liegen. Damit schätzen wir den komponentenweisen Beitrag zur euklidischen Distanz durch

$$(y_{ij} - y_{lj})^2 \approx (2\sqrt{\lambda_j})^2 = 4\lambda_j$$

ab. Die ersten Hauptkomponenten liefern den größten Varianzbeitrag und deswegen wird der euklidische Abstand durch diese Hauptkomponenten im Sinne der Abschätzung gut erfasst. In der Praxis werden oft die ersten beiden Hauptkomponenten in einem Scatterplot gezeichnet. Die euklidischen Abstände der Objekte sollen die tatsächlichen euklidischen Abstände widerspiegeln.

In einen Scatterplot der ersten $p \in \{2, 3\}$ Hauptkomponenten tragen wir zusätzlich die zeilenweisen Komponenten der Matrix L als Vektoren ein. Denn mit

$$R_X = D_X^{-\frac{1}{2}} S_X D_X^{-\frac{1}{2}} = D_X^{-\frac{1}{2}} L L^T D_X^{-\frac{1}{2}}$$

ist der Korrelationskoeffizient ρ_{X_i,X_l} gegeben durch

$$\rho_{X_i,X_l} = \underbrace{(D_X^{-\frac{1}{2}} L)_i}_{i-\text{te Zeile}} \underbrace{(D_X^{-\frac{1}{2}} L)_l^T}_{l-\text{te Zeile}} = \frac{\mathbf{l}_i \mathbf{l}_l^T}{s_{X_i} s_{X_l}}.$$

Der Korrelationskoeffizient zweier Merkmale entspricht geometrisch dem Winkel zwischen den beiden Merkmalen. Dieser wird durch das mit den Standardabweichungen skalierte Skalarprodukt der Zeilen von L repräsentiert. Wegen

$$\mathbf{l}_i \mathbf{l}_l^T = \sum_{j=1}^{k} l_{ij} l_{lj} = \sum_{j=1}^{k} q_{ij}\sqrt{\lambda_j} \cdot q_{lj}\sqrt{\lambda_j} \leq \sum_{j=1}^{k} \lambda_j$$

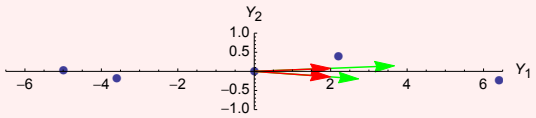
liefern die ersten Komponenten den größten Beitrag zum Skalarprodukt. Das Reduzieren auf p Hauptkomponenten lässt nun die näherungsweise Darstellung in einem p-dimensionalen Koordinatensystem zu. Dabei werden die n Merkmalsträger durch die p-dimensionalen Zeilenvektoren von $Y^{(p)}$ als Punkte und die p Merkmale durch p-dimensionale Zeilenvektoren von ${L^{(p)}}^T$ als Vektoren repräsentiert.

Die Winkel zwischen den Vektoren und den Koordinatenachsen entsprechen ebenfalls den Korrelationen zwischen den gegebenen Merkmalen und den Hauptkomponenten:

$$\begin{aligned}\rho_{Y_j,X_i} &= \frac{(\mathbf{Y_j})^T\mathbf{X_i}}{||\mathbf{Y_j}|| \cdot ||\mathbf{X_i}||} = \frac{(X\mathbf{Q_j})^T\mathbf{X_i}}{||X\mathbf{Q_j}|| \cdot ||\mathbf{X_i}||} \\ &= \frac{\mathbf{Q_j}^T(S_X)_i}{\sqrt{\lambda_j} \cdot s_{X_i}} = \frac{\mathbf{L_j}^T L\mathbf{l}_i^T}{\lambda_j \cdot s_{X_i}} \\ &\overset{(5.5)}{=} \frac{l_{ij}}{s_{X_i}}.\end{aligned} \tag{6.2}$$

Beispiel 6.1

Für unser Beispiel 5.8 ergibt sich folgender Biplot.



Da es sich um das volle Modell handelt, stimmen die Winkel exakt überein. Rot sind die skalierten Vektoren zu sehen. Zur besseren Sichtbarkeit wurden die beiden Längen verdoppelt. In grün dagegen die nicht skalierten. je länger der Vektor, desto größer ist die Varianz des dazugehörigen Merkmals. Die beiden Merkmale sind mit der Hauptkomponente Y_1 hoch korreliert, mit der Hauptkomponente Y_2 kaum.

Spezialfall: Kleine standardisierte Daten
Mit der Darstellung

$$ZQ = Y \text{ und } R^{-1} = (Q\Lambda Q^T)^{-1} = Q\Lambda^{-1}Q^T$$

gemäß (5.4) gilt für Zeilenvektoren $\mathbf{f}_i$ von F mit $\mathbf{f}_i = \mathbf{z}_i T\Lambda^{-\frac{1}{2}}$

$$\begin{aligned}||\mathbf{f}_i - \mathbf{f}_j||^2 &= (\mathbf{f}_i - \mathbf{f}_j)(\mathbf{f}_i - \mathbf{f}_j)^T \\ &= (\mathbf{z}_i - \mathbf{z}_j)Q\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}Q^T(\mathbf{z}_i - \mathbf{z}_j)^T \\ &= (\mathbf{z}_i - \mathbf{z}_j)Q\Lambda^{-1}Q^T(\mathbf{z}_i - \mathbf{z}_j)^T \\ &= (\mathbf{z}_i - \mathbf{z}_j)R^{-1}(\mathbf{z}_i - \mathbf{z}_j)^T.\end{aligned} \tag{6.3}$$

Der euklidische Abstand der Zeilen von F entspricht einer über die Korrelationsmatrix erzeugten gewichteten euklidischen Distanz (Mahalanobis-Distanz) der ursprünglichen Zeilen von Z.

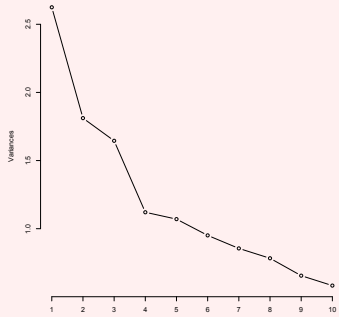
Beschränken wir uns für eine graphische Darstellung auf ein bis drei Dimensionen, d.h. $Z = F^{(p)}L^{(p)^T}$ mit $p \in \{1, 2, 3\}$, ist (6.3) zwar wie oben ausgeführt nicht mehr gültig, die grundsätzliche Richtung aber ist aufgrund der erklärenden Variabilität durch die Kommunalitäten richtig. Mit $R = Z^TZ = LL^T$ ist

$$\rho_{X_i,X_j} = \mathbf{l}_i\mathbf{l}_j^T$$

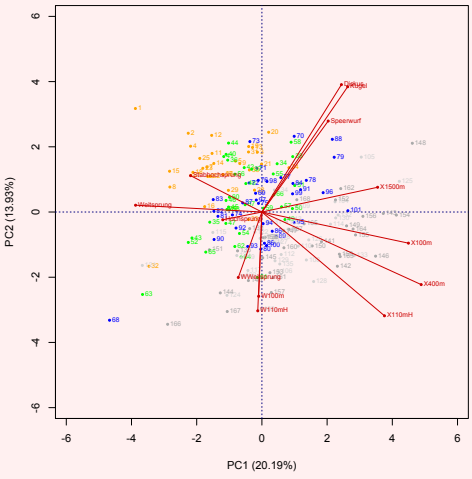
für jeden Eintrag der Korrelationsmatrix, wobei $\mathbf{l}_i$ wiederum die i-te Zeile von L ist.

Beispiel 6.2

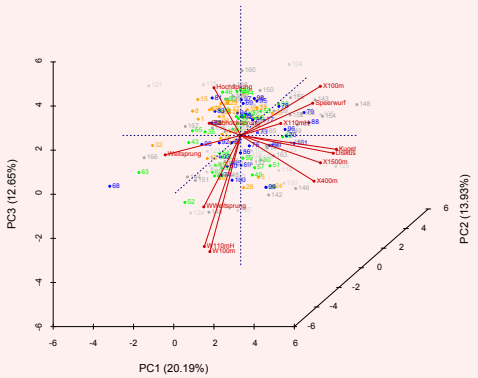
Wir untersuchen Zehnkampfdaten von 168 Spitzensportlern aus dem Jahr 2013 und versuchen, die Sportler in einem zwei- bzw. dreidimensionalen Raum zu positionieren.



Der Screeplot zeigt den deutlichsten Knick beim vierten Eigenwert. Nach diesem Kriterium sollten vier Hauptkomponenten (55% der Varianz) gewählt werden. Mit dem Kaiserkriterium wären es sogar fünf (64% der Varianz) Hauptkomponenten. Weiter ist zu erkennen, dass auch die weiteren nächsten Hauptkomponenten deutliche Beiträge zur Varianz liefern. Zwei Hauptkomponenten liefern 34% der Varianz. Damit erzeugen wir einen Biplot.



Die Tendenz, dass die den Wind beschreibenden Merkmale, die Laufdisziplinen und die technischen Disziplinen in ähnliche Richtungen zeigen, ist erkennbar. Wird aber eine Hauptkomponenten hinzugefügt, zeigen sich deutlichere Unterschiede innerhalb der Gruppen.



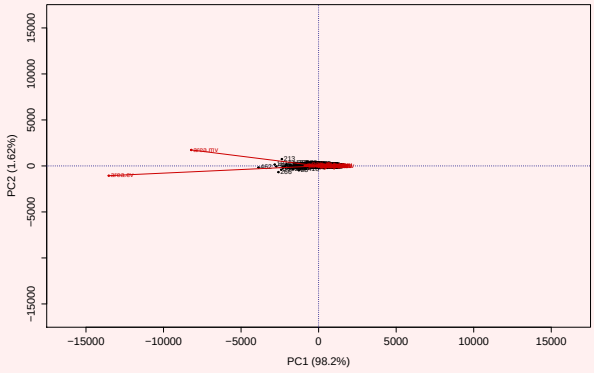
Lediglich Diskuswurf und Kugelstoßen bleiben sehr nahe beieinander. Weiter interpretieren wir, dass gerade im Stabhochsprung die besten Sportler gute Ergebnisse erzielen, was durch eine erkennbare Korrelation von 0.47 bestätigt wird. Eine noch höhere (negative) Korrelation gibt es zwischen den Punkten und den 110m Hürden (-0.63).

```
data=read.table("Zehnkampf2.txt",sep="\t",head=TRUE)
attach(data)
plot(X100m,W100m,cex=0.7,pch=19)
cor(X100m,W100m)
library(lattice)
splom(cbind(data[10:22]))
cor(Kugel,Diskus)
cor(X100m,X110mH)
cor(cbind(data[10:22]))

p=prcomp(cbind(data[10:22]),cor=TRUE,retx=TRUE,
   center=TRUE,scale=TRUE)
print(p)
plot(p,type="lines")
sum(p$sdev[1:2]^2)/sum(p$sdev^2)
sum(p$sdev[1:3]^2)/sum(p$sdev^2)
sum(p$sdev[1:4]^2)/sum(p$sdev^2)
sum(p$sdev[1:5]^2)/sum(p$sdev^2)
biplot(p)
library(bpca)
plot(bpca(cbind(data[10:22])),var.col='red',
   var.factor=.6,var.cex=.6,obj.names=FALSE,
   obj.cex=.,obj.col=c('gold','lightgrey','brown3',
   'springgreen1','springgreen2','springgreen3',
   'springgreen4','slateblue','slateblue1',
   'slateblue2')[(unclass(Punkte)-7500)%/%140+1])
```

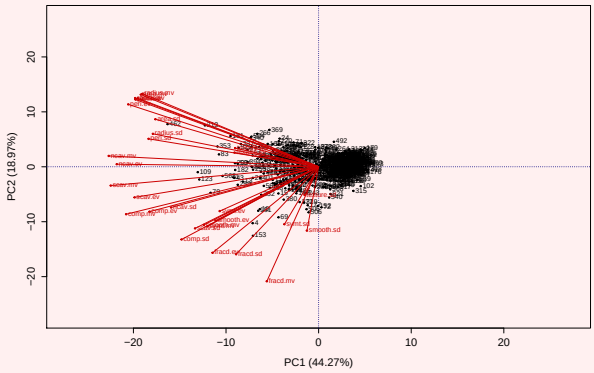
```
plot(bpca(cbind(data[10:22]),method='hj',lambda.end=3),
   rgl.use=TRUE,var.col='red',var.factor=.6,var.cex=.6,
   obj.names=FALSE,obj.cex=.7,obj.col=c('gold',
   'lightgrey','brown3','springgreen1','springgreen2',
   'springgreen3','springgreen4','slateblue','slateblue1',
   'slateblue2')[(unclass(Punkte)-7500)%/%140+1],
   simple.axes=FALSE,box=TRUE)
cor(cbind(Stabhochsprung,X110mH,Punkte))
```

Beispiel 6.3

Bei unserem Brustkrebs-Beispiel 5.12 haben wir zunächst die zentrierten Daten betrachtet und eine bzw. zwei Hauptkomponenten extrahiert. Der dazugehörige Biplot zeigt diese beiden Hauptkomponenten, die anderen sind aufgrund ihrer kurzen Länge nicht zu sehen.



Für die kleinen standardisierten Daten haben wir sechs Hauptkomponenten als sinnvoll erachtet. Mit zwei Hauptkomponenten haben wir 63.24% der Totalvariation erklärt. Wir betrachten den Biplot für die beiden ersten Hauptkomponenten.



6.2. Nichtmetrische multidimensionale Skalierung

Bei der multidimensionalen Skalierung (MDS) wird versucht, n Merkmalsträger so in einem niedrig-dimensionalen Raum zu positionieren, dass gegebene Informationen über ihren Abstand zueinander erhalten bleiben. Dabei genügt es grundsätzlich sogar, nur Abstandsinformationen zur Verfügung zu haben. Die Hauptkomponentenanalyse liefert p Hauptkomponenten. Die Positionierung der Objekte entsprechend dieser Hauptkomponenten erhält einen großen Teil der Abstandsinformation. Für gegebene Datenmatrizen mit kardinalskalierten Merkmalen ermöglicht die Hauptkomponentenanalyse eine solche Positionierung. Wir betrachten nun auch den Fall, dass lediglich Abstandsinformationen bzw. allgemeiner Aussagen bezüglich der Ähnlichkeit zweier Merkmalsträger vorliegen.

Abstände auf Basis von Datenmatrizen quantitativer Merkmale

Gegeben sei eine Datenmatrix $X \in \mathbb{R}^{n,k}$ mit n Objekten und k Merkmalen. Eine Möglichkeit Distanzen zu bestimmen ist bei rein quantitativen Merkmalen die Minkowski-Metrik für $p, q \in \mathbb{N} \setminus \{0\}$

$$d_{p,q} : \mathbb{R}^k \times \mathbb{R}^k \to [0, \infty),\ (\mathbf{x}, \mathbf{y}) \mapsto d_{p,q}(\mathbf{x}, \mathbf{y}) := \left(\sum_{j=1}^{k} |x_j - y_j|^p \right)^{\frac{1}{q}},$$

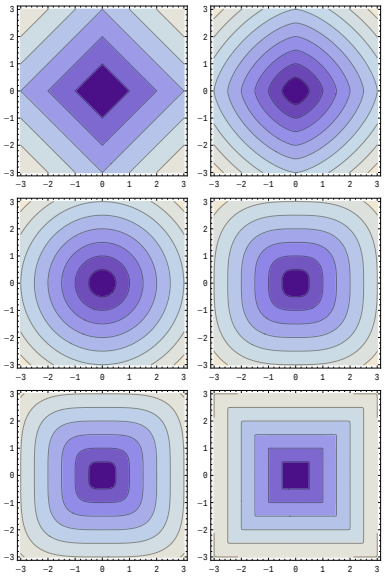
Wichtige Spezialfälle sind:

p	q	Bezeichnung
1	1	City-Block-Distanz
2	2	Euklidische Distanz
2	1	Quadrierte Euklidische Distanz

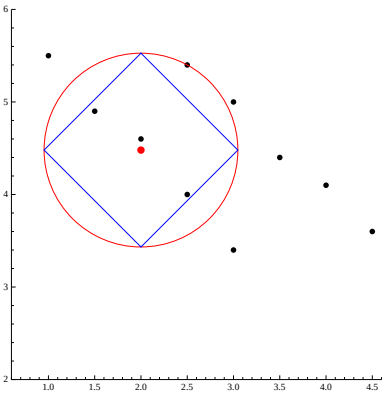
Ein weiteres Distanzmaß ist die so genannte Chebychev-Distanz

$$d : \mathbb{R}^k \times \mathbb{R}^k,\ (\mathbf{x}, \mathbf{y}) \mapsto d(\mathbf{x}, \mathbf{y}) := \max_{j \in \{1,\ldots,k\}} \{|x_j - y_j|\}.$$

In der nebenstehenden Darstellung sind sechs Minkowski-Metriken im Vergleich zu sehen. Es werden die Werte $p = q \in \{1, 3/2, 2, 3, 4, 100\}$ zur Abstandsbestimmung vom Ursprung benutzt. Die Annäherung an die Chebychev-Distanz ist mit steigendem p deutlich zu erkennen.



Die nebenstehende Situation zeigt jedoch, dass diese Distanzmaße die Situation bei einer Analyse nicht immer zufriedenstellend beschreiben. Werden Abstände zu dem roten Punkt betrachtet, suggerieren die schwarzen Punkte zwei verschiedene Bereiche, es scheint eine gedankliche Trennlinie zwischen den zehn Objekten möglich. Der rote Punkt könnte der zentrale Punkt für den linken Bereich sein. Sowohl die Euklidische Distanz als auch die City-Block-Distanz können diesen Sachverhalt nicht ganz abbilden, da es Punkte im rechten Bereich gibt, deren Abstand zum roten Punkt kleiner ist als der Abstand verschiedener Punkte im linken Bereich. Beide Bereiche im Beispiel zeichnet aus, dass die entsprechenden Merkmale eine hohe negative Korrelation besitzen. In einem ersten Schritt könnte das Koordinatensystem gedreht und das Zentrum auf den roten Punkt gelegt werden. Ein Distanzmaß müsste in einem zweiten Schritt Abstände entlang der zu den Punkten parallelen Achse weniger gewichten als entlang der dazu orthogonalen Achse.



Sei $\mathbf{x} = (x_1, x_2)^T$ ein Punkt bzgl. des kartesischen Koordinatensystems. Einer Drehung des Koordinatensystems entspricht ein Basiswechsel mit Basisvektoren $\mathbf{q}_1, \mathbf{q}_2$, $\mathbf{x} = Q\mathbf{y}$, mit $Q = (q_1 \ q_2)$ und $\mathbf{y} = (y_1, y_2)^T$. $\mathbf{y}$ ist die Darstellung von $\mathbf{x}$ in der neuen Basis. Der negativ-lineare Zusammenhang im Beispiel wird durch die Varianz-Kovarianzmatrix S beschrieben. Die Drehung des Koordinatensystems bewirkt, dass die Kovarianz zu Null wird. Das entspricht einer Diagonalisierung der Varianz-Kovarianzmatrix und q_1, q_2 sind die Eigenvektoren, λ_1, λ_2 die Eigenwerte der Matrix. Es gilt dann

$$\Lambda := \text{diag}(\lambda_1, \lambda_2) = Q^T S Q.$$

Sei $\tilde{w} = y - \bar{y}$ mit $\bar{y} = Q^{-1}\bar{x}$. Alle so erzeugten $\tilde{w}$ sind zentriert, das entspricht der angedachten Verschiebung. Die Streuung der gewonnenen $\tilde{w}$ ist jedoch für jede Komponente noch verschieden und ist durch die Varianzen im Koordinatensystem q_1, q_2 bestimmt. Die bisher betrachteten Distanzmaße legen aber eine gleiche Gewichtung jeder Koordinate zugrunde, was etwa anschaulich durch die konzentrischen Kreise bei der Euklidischen Distanz wird. Eine Division durch die Standardabweichung s ergibt für jede Komponente

eine Varianz von Eins und entspricht genau der Wurzel des inversen Eigenwerts (diese müssen als von Null verschieden angenommen werden). Mit $w := \frac{y-\bar{y}}{s}$ ist die quadrierte euklidische Distanz in der neuen Basis durch

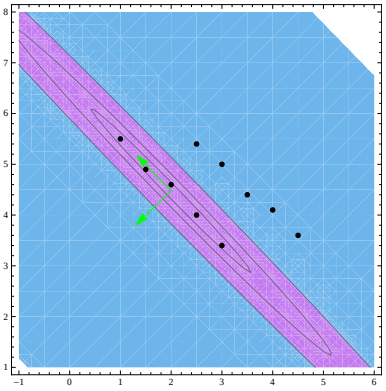
$$\begin{aligned} w^T w &= (y-\bar{y})^T \text{diag}\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}\right)(y-\bar{y}) \\ &= (y-\bar{y})^T \Lambda^{-1}(y-\bar{y}) \\ &= (y-\bar{y})^T Q^T S^{-1} Q(y-\bar{y}) \\ &= (x-\bar{x})^T S^{-1}(x-\bar{x}) \end{aligned} \tag{6.4}$$

gegeben.

Durch

$$d(x,y) := \sqrt{(x-y)^T S^{-1}(x-y)}$$

wird die so genannte Mahalanobis-Distanz bestimmt. Geometrisch wird um das Zentrum ein Ellipsoid gelegt, dessen Radien sich entlang der Eigenvektoren und in der Länge an den inversen Eigenwerten orientieren. Die gewünschte Eigenschaft der Abstandsbestimmung wird damit erreicht.



Monotone Übereinstimmung der Abstände

Gegeben sei eine Distanzmatrix $D = (d_{ij}) \in \mathbb{R}^{n,n}$. Wir suchen eine Repräsentation der n Objekte im p-dimensionalen Raum $\mathbb{R}^p$, dass für die Distanzen δ_{ij} der Objekte der Zusammenhang

$$\delta_{ij} = f(d_{ij}),\ f : \mathbb{R} \to \mathbb{R} \text{ monoton wachsend,}$$

gilt. Es soll also die Monotoniebedingung

$$d_{ij} \leq d_{kl} \Rightarrow f(d_{ij}) = \delta_{ij} \leq \delta_{kl} = f(d_{kl})\ \forall i,j,k,l \in \{1,\ldots,n\} \tag{6.5}$$

gelten. Das folgende Verfahren von Kruskal (aus [2]) liefert das grundsätzliche Vorgehen aller nicht-metrischen MDS-Verfahren.

1) Für eine Startkonfiguration $X^0 \in \mathbb{R}^{n,p}$, $p < n-1$, werden die Distanzen $\delta_{ij}(X^0)$ gemäß einer vorgegebenen Metrik bestimmt. Nun erfolgt eine Iteration (Laufindex q).

2) Um die Monotoniebedingung zu erfüllen, wird eine monotone Regression der Distanzen $\delta_{ij}(X^q)$ auf die Rangordnung der Distanzen d_{ij} durchgeführt. Daraus resultieren neue Distanzen δ_{ij}^q, die nun (6.5) erfüllen.

3) Ein Gütekriterium (Stress) wird angewandt, um die „Nähe“ der Distanzen δ_{ij}^q zu Distanzen $\delta_{ij}(X^q)$ der Konfiguration zu überprüfen.

4) Mittels einer Abstiegsmethode wird die Konfiguration X^q so zu X^{q+1} verschoben, dass der Stress nach Möglichkeit verringert wird.

5) Solange der Stress abnimmt und kein Abbruchkriterium erfüllt ist, wiederhole die Schritte ab 2).

Startkonfiguration

Im Verfahren wird eine Abstiegsmethode benutzt. Da dies ein lokales Verfahren ist, muss eine geeignete Wahl einer Startkonfiguration besonderes Augenmerk bekommen. Eine rein zufällige Wahl erscheint somit nicht angebracht. In der Praxis haben sich ein paar Varianten durchgesetzt, von denen wir das folgende betrachten. Wir beginnen mit den geordneten Abständen der Objekte mit Hilfe der Ordnungsstatistik $T : \mathbb{R}^{\frac{n(n-1)}{2}} \to \mathbb{R}^{\frac{n(n-1)}{2}}$,

$$(d_{12}, d_{13}, \ldots, d_{n-1,n})^T \mapsto \left(d_{(1)}, d_{(2)}, \ldots, d_{\left(\frac{n(n-1)}{2}\right)}\right)^T := T\left(d_{12}, d_{13}, \ldots, d_{n-1,n}\right)$$

und setzen $\delta_{ij} := k$ für $d_{ij} = d_{(k)}$ und $i, j \in \{1, \ldots, n\}$. Es sei $\Delta = (\delta_{ij})$ die Matrix der so gewonnen Rangdistanzen. Gesucht werde eine Ausgangsmatrix $X \in \mathbb{R}^{n,p}$, welche die gegebenen Distanzen möglichst gut wiedergibt. Es gilt mit $\mathbf{c}$:=diag(XX^T)

$$\Delta = \mathbf{c}\mathbf{1}^T + \mathbf{1}\mathbf{c}^T - 2XX^T.$$

Durch eine zeilen- und spaltenweise Zentrierung $H\Delta H$ mit der Zentrierungsmatrix $H = I^{n,n} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ gemäß Abschnitt 3.2 folgt

$$H\Delta H = -2HXX^TH = -2(HX)(HX^T).$$

Mit $Z := (HX)^T$ folgt mit Hilfe der Hauptachsentransformation dann $Z^TZ = LL^T = -\frac{1}{2}H\Delta H$ und mit $X := L$ entsprechend $HLL^TH = -\frac{1}{2}HH\Delta HH = -\frac{1}{2}H\Delta H$. Wir verfahren nach folgendem Algorithmus (aus [6]):

1) Setze $A = (a_{ij})$, $a_{ij} := -\frac{1}{2}\delta_{ij}^2$ und $H = I^{n,n} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.
2) Berechne $B := HAH$.
3) Bestimme die p größten Eigenwerte $\lambda_1 > \ldots > \lambda_p$ von B, sowie die zugehörigen normierten Eigenvektoren $\mathbf{v}_1, \ldots, \mathbf{v}_p$.
4) Bilde $X^0 := (x_{ij}) = (\sqrt{\lambda_1}\mathbf{v}_1, \ldots, \sqrt{\lambda_p}\mathbf{v}_p)$.

Die Koordinaten der gesuchten Konfiguration sind die Zeilen von X^0. Zur Bestimmung einer Startkonfiguration nutzt R standardmäßig den Befehl

```
cmdscale(d,k=2)
```

Monotone Regression

Auf Basis der Ordnungsstatistik der gegebenen Distanzen d_{ij} werden die Distanzen $\delta_{ij}(X^q)$ der Konfiguration betrachtet. Wird die Monotoniebedingung (6.5) aufeinanderfolgender Distanzen nicht verletzt, bleibt die Distanz zunächst erhalten. Wird sie verletzt, werden diese zu einem Block zusammengefasst und ihr arithmetisches Mittel gebildet. Verstößt dieser Wert bzgl. der direkt vorherigen oder der direkt nachfolgenden Distanz gegen die Monotoniebedingung, wird die entsprechende Distanz zum Block hinzugefügt und erneut das arithmetische Mittel des Blocks gebildet. Solange in eine Richtung gegen die Monotoniebedingung verstoßen wird, wird der Vorgang in diese Richtung fortgesetzt. Die Distanzen werden dann durch das arithmetische Mittel ersetzt. Es resultieren die Distanzen δ_{ij}^q im q-ten Iterationsschritt.

Gütekriterium Stress

Die Distanzen $\delta_{ij}(X^q)$ hängen von einer Konfiguration $X \in \mathbb{R}^{n,p}$ ab. Ziel ist es, eine solche Konfiguration zu finden, welche die Monotoniebedingung (6.5) möglichst gut erfüllt und so ein vorgegebenes Gütemaß minimiert wird. Ein mögliches Gütemaß haben Kruskal und Shepard entwickelt und es wird als Stress S bezeichnet:

$$S(X)^2 := \frac{\sum\limits_{i<j} (\delta_{ij}(X^q) - \delta_{ij}^q)^2}{\sum\limits_{i<j} \delta_{ij}(X^q)^2} \tag{6.6}$$

Ein Wert von $S < 0.05$ wird als gut bezeichnet. Dennoch muss berücksichtigt werden, dass der Stress von der gewählten Metrik, der Dimension p und der Anzahl Objekte abhängt.

Erzeugen einer neuen Objektkonfiguration

Sei $\kappa_{ij} := \frac{\delta_{ij}^q(X^q) - \delta_{ij}^q}{\delta_{ij}(X^q)} = 1 - \frac{\delta_{ij}^q}{\delta_{ij}(X^q)}$ die relative Änderung der Distanz aus der Konfiguration. Damit können wir die „verbesserten“ Positionen der Punkte der Konfiguration bestimmen. Den Korrekturvektor von Punkt i bzgl. Punkt j erhalten wir aus der Multiplikation des Vektors von i nach j mit dem Korrekturfaktor κ_{ij}. Beim Übergang der Konfiguration X^q zu X^{q+1} wird der Punkt i auf der Achse $l \in \{1, \ldots, p\}$ von x_{il}^q nach x_{il}^{q+1} durch $\kappa_{ij}(x_{jl} - x_{il})$ bewegt. Für alle Punkte j entlang der Achse l erhalten wir in Summe

$$\sum_{j \neq i} \kappa_{ij}(x_{jl} - x_{il}).$$

Die neue Koordinate x_{il}^{q+1} bekommen wir mit dieser Überlegung durch

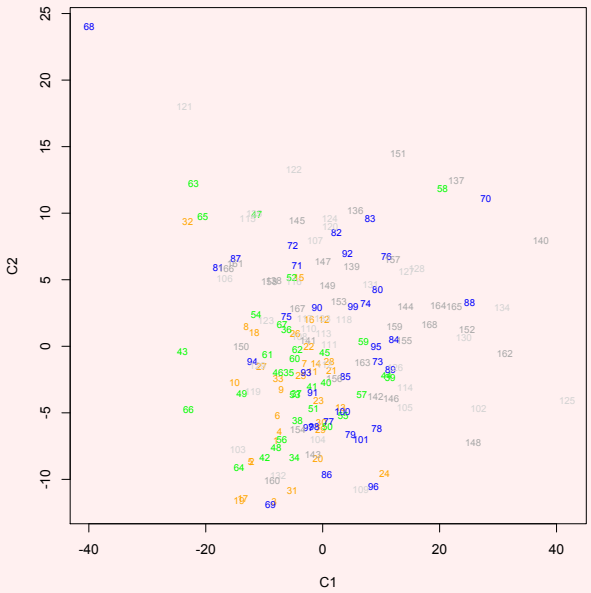
$$x_{il}^{q+1} = x_{il}^q + \alpha^q \sum_{j \neq i} \left(1 - \frac{\delta_{ij}^q}{\delta_{ij}^q(X^q)}\right)(x_{jl} - x_{il}).$$

Dabei ist $\alpha^q > 0$ ein Korrekturfaktor, um den Betrag des Korrekturvektors anzupassen. Oft wird $\alpha^0 = 0.2$ gewählt und zunehmend verringert.

Beispiel 6.4

Die Zehnkämpfer sollen zweidimensional positioniert werden. Die Ausgabe in R zeigt den Stress in Prozent im jeweiligen Iterationsschritt:

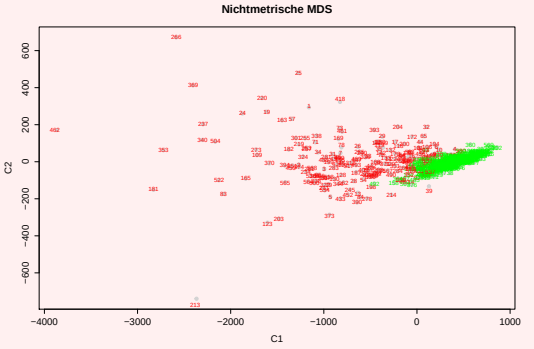
```
initial  value 7.169175
iter   5 value 6.245310
iter  10 value 6.135196
iter  10 value 6.133430
iter  10 value 6.132463
final  value 6.132463
converged
```

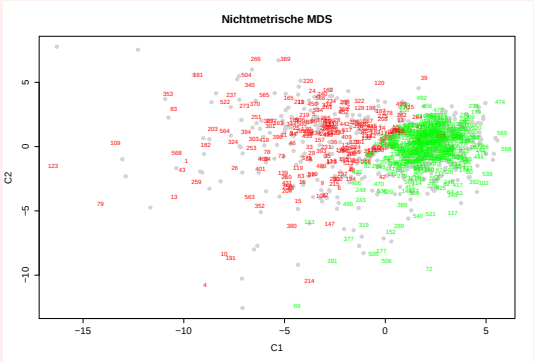



```
library(MASS)
d=dist(cbind(data[10:22]))
fit=isoMDS(d,k=2,p=2)
fit
x=fit$points[,1]
y=fit$points[,2]

plot(x,y,xlab="C1",ylab="C2",main="",type="n")
text(x,y,labels=row.names(cbind(data[10:22])),
   cex=.7,col=c('orange','green','blue','lightgrey',
   'darkgrey')[as.numeric(unclass(row.names(
   cbind(data[10:22]))))%/%34+1])
cmdscale(d,k=2)
```

Beispiel 6.5

Wir wollen wiederum die Brustkrebs-Daten untersuchen und die einzelnen Personen im $\mathbb{R}^2$ positionieren. Dabei unterscheiden wir wiederum zwischen den ursprünglichen und den kleinen standardisierten Daten. Hinzu kommt hier eine Farbkomponente, die zwischen gesunden und kranken Personen unterscheidet. Es zeigt sich, dass es eine gute Trennung dieser beiden Gruppen gibt, was wir später noch weiter untersuchen werden.





In beiden Fällen ist die Startkonfiguration genau das Ergebnis des Hauptkomponentenanalyse mit zwei Hauptkomponenten. Bei den standardisierten Daten fallen die Verschiebungen durch die nichtmetrische MDS stärker auf und sind auch etwas stärker im Vergleich. Dennoch ändert sich in beiden Fällen die Anfangslösung nur wenig.

Teil III.

Abhängigkeiten

7. Assoziationsregeln

Warum?

Oft soll das Verhalten oder eine Eigenschaft von Untersuchungsobjekten vorhergesagt werden. Dabei soll sich die Vorhersage auf andere zu erhebende Eigenschaften stützen. Werden sämtliche Eigenschaften in Form von Kategorien angesetzt und ist es möglich, auf Daten zu allen Kategorien zurückzugreifen, so kann darauf basierend (supervised) ein Regelwerk erzeugt werden, das die Vorhersage bei einem neuen Untersuchungsobjekt ohne Kenntnis der vorherzusagenden Eigenschaft ermöglicht.

Assoziationsregeln sind Implikationen, die Abhängigkeiten zwischen Merkmalswerten beschreiben und nach gewissen Gütekriterien beurteilt werden sollen. So kann etwa bei den Titanic-Daten danach gefragt werden, ob aus der Zugehörigkeit zur ersten Klasse und der Gruppe der Frauen mit einer gewissen Güte auf das Überleben geschlossen werden kann.

7.1. Modellierung von Assoziationsregeln

Sei $I = \{i_1, \ldots, i_k\}$ eine endliche Menge so genannter Items. Jede Teilmenge $X \subseteq I$ von I wird Itemset genannt. Aus der Menge $\mathcal{P}(I)$ aller Itemsets werde eines beobachtet. Diese Beobachtung wird als Transaktion T bezeichnet. Die Menge aller Transaktionen heißt Transaktionsdatenbank und wird mit $\mathcal{D}$ bezeichnet.

Die Menge der Items besteht im Titanic-Datensatz aus $4 + 2 + 2 + 2 = 10$ Elementen,

$$I = \{\text{First, Second, Third, Crew, Male, Female, Adult, Child, Yes, No}\}.$$

Ein mögliches beobachtetes Itemset ist

$$T_1 = \{\text{First, Female, Adult, Yes}\}.$$

An dieser Stelle sind zunächst auch auf Basis der Datenstruktur sinnlose Itemsets wie $X = \{\text{First, Second, Female, Child}\}$ möglich. Die Transaktionsdatenbank besteht aus 2201 Transaktionen.

Eine Implikation besteht aus einer Prämisse und einer Konklusion und ist von der Form

WENN Prämisse DANN Konklusion.

Im Beispiel der Titanic-Daten handelt es sich dabei um die Implikation

WENN First und Female DANN Yes.

Wie oft ist diese Kombination im Datensatz aufgetreten? Wir untersuchen das Itemset $X = \{\text{First, Female, Yes}\}$ und sehen zunächst $X \subseteq T_1$. Insgesamt gibt es 141 Transaktionen,

in denen das Itemset auftaucht. Das entspricht einer relativen Häufigkeit von 0.064. Diese relative Häufigkeit wird der Support von X genannt,

$$\text{supp}_{\mathcal{D}}(X) := \frac{|\{T \in \mathcal{D};\ X \subseteq T\}|}{|\mathcal{D}|}.$$

Wir fassen die bisherigen Überlegungen zusammen. Eine Beobachtung kann als Realisierung einer Zufallsvariablen angesehen werden. Wir betrachten einen Wahrscheinlichkeitsraum

$$(\mathcal{P}(I), \mathcal{P}(\mathcal{P}(I)), \mathbb{P})$$

mit einem unbekannten Wahrscheinlichkeitsmaß $\mathbb{P}$. Einem Ereignis A entspricht dabei die Menge aller Itemsets, die ein vorgegebenes Itemset als Teilmenge haben. A tritt ein, wenn für die Beobachtung T gilt $T \in A$. Da wir das Wahrscheinlichkeitsmaß nicht kennen, müssen wir es durch n-malige Realisierung des Zufallsexperiments schätzen, d.h. durch Abzählen festlegen.

Gibt es zehn Items, so ist eine unvorstellbare Zahl von $2^{1024} \sim 10^{308}$ verschiedenen Ereignissen vorhanden. Oft kann in der Anwendung auf die Angabe aller Elemente des Ereignisses verzichtet werden. In unserem Beispiel $(127 \to 3)$ ist eine sinnvolle und den Sachverhalt beschreibende Darstellung des Ereignisses A gegeben durch

$$\begin{aligned} A &= \{Z \subseteq I;\ X \subseteq Z\} \\ &= \{\{\text{First,Female,Yes}\}, \{\text{First,Child,Female,Yes}\}, \{\text{First,Adult,Female,Yes}\}, \ldots\}, \end{aligned}$$

jedes Element enthält das Itemset X. Dem Durchschnitt zweier Ereignisse entspricht die Vereinigung der dazugehörigen Itemsets, denn mit den Itemsets X, Y ist

$$\underbrace{\{Z \subseteq I;\ X \subseteq Z\}}_{A} \cap \underbrace{\{Z \subseteq I;\ Y \subseteq Z\}}_{B} = \{Z \subseteq I;\ X \cup Y \subseteq Z\} \tag{7.1}$$

Gilt nun $X \cap Y = \emptyset$ mit $X, Y \neq \emptyset$, so wird die Implikation $X \to Y$ Assoziationsregel genannt. Die Wahrscheinlichkeit $\mathbb{P}(A \cap B)$ lässt sich als die Wahrscheinlichkeit für das Vorkommen aller Bestandteile der Assoziationsregel interpretieren. Die Anzahl Transaktionen, welche die Items von X und die Items von Y enthalten, lässt sich bei einer konkreten Realisierung einer Transaktionsdatenbank $\mathcal{D}$ gemäß (7.1) über die Vereinigung der beiden Itemsets bestimmen. Der Support der Assoziationsregel kann somit durch

$$\text{supp}_{\mathcal{D}}(X \to Y) := \text{supp}_{\mathcal{D}}(X \cup Y) = \frac{|\{T \in \mathcal{D};\ X \cup Y \subseteq T\}|}{|\mathcal{D}|}.$$

beschrieben werden. Je größer der Wert, desto öfter kommen die Items in dieser Kombination in der Transaktionsdatenbank vor. Dies können wir als ein erstes Gütemaß für die gebildete Assoziationsregel verwenden, indem wir sagen, eine Assoziationsregel muss einen Mindestgrad an Support haben, um in Betracht gezogen zu werden.

Ist der Support als Messlatte niedrig angesetzt, werden viele Regeln akzeptiert, auch wenn sie möglicherweise nur zufälliger Natur sind. Im umgekehrten Fall könnten relevante Regeln nicht mehr erkannt werden. So werden Verbraucher zwar nicht sehr häufig Großgeräte anschaffen (geringer Support), dennoch könnte es Prämissen geben, die zu einer erhöhten Kaufmotivation führen. Solche Assoziationen gilt es herauszufinden.

An diesen Punkt anknüpfend lässt sich ein weiteres Kriterium für Assoziationsmaße entwickeln. Wenn sich die relative Häufigkeit der Konklusion Y unter der Bedingung der Prämisse X ungleich anders als ohne Prämisse verhält, so ist ein Bezug zwischen Prämisse und Konklusion durchaus möglich. In der Sprache der Wahrscheinlichkeit geht es um die a-priori Wahrscheinlichkeit $\mathbb{P}(B)$ mit $B = \{Z \subseteq I;\ Y \subseteq Z\}$ im Vergleich zur bedingten Wahrscheinlichkeit $\mathbb{P}(B|A)$ mit $A = \{Z \subseteq I;\ X \subseteq Z\}$.

Satz und Definition 7.1: Bedingte Wahrscheinlichkeit

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein beliebiger Wahrscheinlichkeitsraum. Sind $A, B \in \mathcal{F}$ beliebige Ereignisse mit $\mathbb{P}(B) > 0$, so heißt

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

die bedingte Wahrscheinlichkeit von A unter B. Eine Erweiterung auf n Ereignisse ist folgendermaßen möglich: Zunächst formen wir etwas um zu

$$\mathbb{P}(A|B) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B).$$

Dann betrachten wir drei Ereignisse $A_1, A_2, A_3 \in \mathcal{F}$ mit $\mathbb{P}(A_1 \cap A_2) > 0$ und erhalten die bedingte Wahrscheinlichkeit

$$\mathbb{P}(A_3|A_1 \cap A_2) \cdot \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_3 \cap (A_1 \cap A_2)).$$

Iterativ setzen wir mit $A_1, A_2, \ldots, A_n \in \mathcal{F}$ und $\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) > 0$ fort zu

$$\begin{aligned}
\mathbb{P}(A_1 \cap \ldots \cap A_n) &= \mathbb{P}(A_n \cap (A_1 \cap \ldots \cap A_{n-1})) \\
&= \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_1 \cap \ldots \cap A_{n-1}) \\
&= \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_{n-1}|A_1 \cap \ldots \cap A_{n-2}) \\
&\quad \cdot \mathbb{P}(A_1 \cap \ldots \cap A_{n-2}) \\
&= \mathbb{P}(A_n|A_1 \cap \ldots \cap A_{n-1}) \cdot \mathbb{P}(A_{n-1}|A_1 \cap \ldots \cap A_{n-2}) \\
&\quad \cdot \ldots \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1)
\end{aligned}$$

Dies wird als Multiplikationssatz bezeichnet. Sei nun $(A_n)_{n\in\mathbb{N}}$ mit $A_i \in \mathcal{F}$ für alle $i = 1, \ldots, n$ eine Folge paarweise disjunkter Ereignisse in $\mathcal{F}$ mit $\bigcup_{i=1}^{\infty} A_i = \Omega$ und $\mathbb{P}(A_i) > 0$, so heißt $(A_n)_{n\in\mathbb{N}}$ eine messbare Zerlegung von Ω. Für ein Ereignis $A \in \mathcal{F}$ gilt dann

$$\begin{aligned}
\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) &= \mathbb{P}\left(A \cap \left(\bigcup_{i=1}^{\infty} A_i\right)\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{\infty} (A \cap A_i)\right) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap A_i) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \frac{\mathbb{P}(A \cap A_i)}{\mathbb{P}(A_i)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i)
\end{aligned}$$

Dies ist der Satz der totalen Wahrscheinlichkeit.

Wir betrachten für $\mathbb{P}(A) > 0$ die bedingte Wahrscheinlichkeit

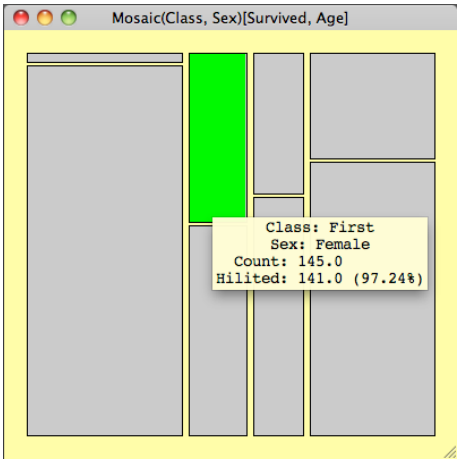
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B) \cdot \mathbb{P}(A|B)}{\mathbb{P}(A)}.$$

Bei einer gegebenen Transaktionsdatenbank lassen sich wiederum die relativen Häufigkeiten in Analogie zu den beiden angesprochenen Wahrscheinlichkeiten bestimmen. $\mathbb{P}(B)$ wird über den Support $\text{supp}_{\mathcal{D}}(Y)$ bestimmt, während der bedingten Wahrscheinlichkeit die als Konfidenz bezeichnete relative Häufigkeit

$$\text{conf}_{\mathcal{D}}(X \to Y) := \frac{\text{supp}_{\mathcal{D}}(X \cup Y)}{\text{supp}_{\mathcal{D}}(X)} = \frac{\frac{|\{T \in \mathcal{D};\ X \cup Y \subseteq T\}|}{|\mathcal{D}|}}{\frac{|\{T \in \mathcal{D};\ X \subseteq T\}|}{|\mathcal{D}|}} = \frac{|\{T \in \mathcal{D};\ X \cup Y \subseteq T\}|}{|\{T \in \mathcal{D};\ X \subseteq T\}|}$$

entspricht. Sollte dabei das Itemset X nicht auftreten ist für die Assoziationsregel keine Konfidenz definiert und die Regel darf auf Basis der vorliegenden Transaktionsdatenbank nicht in Betracht gezogen werden.

Im Titanic-Beispiel betrachten wir die Itemsets $X = \{\text{First, Female}\}$ und $Y = \{\text{Yes}\}$. Support und Konfidenz sind über Barcharts und Mosaicplots leicht bestimmbar und wir erhalten für die Untersuchung der Assoziationsregel $X \to Y$



$$\begin{aligned}
\text{supp}_{\mathcal{D}}(X) &= \frac{145}{2201} = 0.066,\\
\text{supp}_{\mathcal{D}}(Y) &= \frac{711}{2201} = 0.323,\\
\text{supp}_{\mathcal{D}}(X \cup Y) &= \frac{141}{2201} = 0.064,\\
\text{conf}_{\mathcal{D}}(X \to Y) &= \frac{141}{145} = 0.972.
\end{aligned}$$

Wegen $\text{supp}_{\mathcal{D}}(X \cup Y) \leq \text{supp}_{\mathcal{D}}(X)$ gilt $1 \geq \text{conf}_{\mathcal{D}}(X \to Y) \geq 0$. Die Konfidenz als weiteres Gütemaß muss wie auch der Support stets hinterfragt werden. Es genügt nicht von einer hohen Konfidenz auf eine Relevanz einer Regel zu schließen. Um dies zu sehen, benötigen wir das Konzept der Unabhängigkeit zweier Itemsets. Zwei Itemsets X und Y werden unabhängig genannt, wenn

$$\text{supp}_{\mathcal{D}}(X \cup Y) = \text{supp}_{\mathcal{D}}(X) \cdot \text{supp}_{\mathcal{D}}(Y)$$

gilt. Dies überlegen wir uns aus der Definition der stochastischen Unabhängigkeit der dazugehörigen Ereignisse.

Satz und Definition 7.2: Stochastische Unabhängigkeit von Ereignissen

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein beliebiger Wahrscheinlichkeitsraum. Die Ereignisse $(A_i)_{i \in I},\ A_i \in \mathcal{F}$ heißen stochastisch unabhängig, wenn für jede nichtleere endliche Teilmenge $J \subseteq I$

gilt:

$$\mathbb{P}\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} \mathbb{P}(A_j).$$

Daraus folgt unmittelbar für zwei stochastisch unabhängige Ereignisse A und B, dass

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

Sind zwei Itemsets unabhängig, so folgt für die Konfidenz der Regel $X \to Y$

$$\mathsf{conf}_{\mathcal{D}}(X\to Y) = \frac{\mathsf{supp}_{\mathcal{D}}(X\cup Y)}{\mathsf{supp}_{\mathcal{D}}(X)} = \frac{\mathsf{supp}_{\mathcal{D}}(X)\cdot\mathsf{supp}_{\mathcal{D}}(Y)}{\mathsf{supp}_{\mathcal{D}}(X)} = \mathsf{supp}_{\mathcal{D}}(Y).$$

Der Wert hängt damit nur noch von der relativen Häufigkeit des Vorkommens von Y ab. Um das zu berücksichtigen, können wir die Konfidenz $\mathsf{conf}_{\mathcal{D}}(X\to Y)$ und den Support $\mathsf{supp}_{\mathcal{D}}(Y)$ ins Verhältnis setzen, was wir als Lift bezeichnen. Für $\mathsf{supp}_{\mathcal{D}}(Y) > 0$ sei der Lift der Assoziationsregel gegeben durch

$$\mathsf{lift}_{\mathcal{D}}(X\to Y) := \frac{\mathsf{conf}_{\mathcal{D}}(X\to Y)}{\mathsf{supp}_{\mathcal{D}}(Y)} = \frac{\mathsf{supp}_{\mathcal{D}}(X\cup Y)}{\mathsf{supp}_{\mathcal{D}}(X)\cdot\mathsf{supp}_{\mathcal{D}}(Y)}.$$

Die Unabhängigkeit von X und Y ist äquivalent zu einem Lift von 1. Ist der Lift größer als 1, so heißt dies, dass bei Vorliegen der Prämisse die relative Häufigkeit der Konklusion steigt. Ein kleiner Support der Konklusion kann zu sehr großen Lift-Werten führen. Dies muss stets berücksichtigt werden.

Der Lift im Titanic-Beispiel beträgt $\mathsf{lift}_{\mathcal{D}}(X\to Y) = \frac{141}{145}\cdot\frac{2201}{711} = 3.010$.

Ist der Lift kleiner als 1, so führt das Vorliegen der Prämisse zu einem Sinken der relativen Häufigkeit der Konklusion. Eine solche Regel wird nicht akzeptiert. Weiter ist der Lift symmetrisch, $\mathsf{lift}_{\mathcal{D}}(X\to Y) = \mathsf{lift}_{\mathcal{D}}(Y\to X)$, die Richtung der Assoziationsregel ist nicht erkennbar.

Alle drei Gütekriterien gemeinsam können die Schwächen der einzelnen Maße überwinden. Es gibt weitere Maße, die die Schwächen der drei vorgestellten Maße auszugleichen versuchen. So wird die Konviktion definiert durch

$$\mathsf{conv}_{\mathcal{D}}(X\to Y) := \frac{\mathsf{supp}_{\mathcal{D}}(X)\cdot\mathsf{supp}_{\mathcal{D}}(\neg Y)}{\mathsf{supp}_{\mathcal{D}}(X\cup\neg Y)},$$

mit $\mathsf{conv}_{\mathcal{D}}(X\to Y) := \infty$ für $\mathsf{supp}_{\mathcal{D}}(X\cup\neg Y) = 0$. Die Idee stammt von der Überlegung, dass aus logischer Sicht

$$X\to Y \equiv \neg(X\wedge\neg Y)$$

ist. Wir messen damit den Unterschied der Wahrscheinlichkeiten des zu $X\wedge\neg Y$ gehörenden Ereignisses zum Fall der Unabhängigkeit der beiden dazugehörigen Einzelereignisse. Wegen

$$\begin{aligned}\mathsf{conf}_{\mathcal{D}}(X\to\neg Y) &= \frac{\mathsf{supp}_{\mathcal{D}}(X\cup\neg Y)}{\mathsf{supp}_{\mathcal{D}}(X)} = \frac{\mathsf{supp}_{\mathcal{D}}(X) - \mathsf{supp}_{\mathcal{D}}(X\cup Y)}{\mathsf{supp}_{\mathcal{D}}(X)}\\ &= 1 - \mathsf{conf}_{\mathcal{D}}(X\to Y)\end{aligned}$$

lässt sich die Konviktion auch schreiben als

$$\text{conv}_{\mathcal{D}}(X \to Y) = \frac{1 - \text{supp}_{\mathcal{D}}(Y)}{1 - \text{conf}_{\mathcal{D}}(X \to Y)},$$

wobei $\text{conv}_{\mathcal{D}}(X \to Y) := \infty$ für $\text{conf}_{\mathcal{D}}(X \to Y) = 1$. Die Konviktion entspricht dem Inversen des Lift $\text{lift}_{\mathcal{D}}(X \to \neg Y)$, stark abweichende Werte von 1 weisen auf eine Assoziation hin. Eine hohe Konviktion interpretieren wir als Sinken der relativen Häufigkeit der negierten Konklusion bei Vorliegen der Prämisse.

Für die Regel $\{\text{First, Female}\} \to \{\text{Yes}\}$ im Titanic-Beispiel erhalten wir

$$\text{conv}_{\mathcal{D}}(X \to Y) = \frac{1490 \cdot 145}{2201 \cdot 4} = 24.540.$$

7.2. Support-Konfidenz-Ansatz zur Regelerzeugung

Sei $I = \{i_1, \ldots, i_k\}$ die Menge der Items. Zunächst stellen wir zwei Vorüberlegungen an:

- Die Menge $\mathcal{P}(I)$ aller Itemsets wird schnell sehr groß, sie enthält 2^k Elemente.
- Die Menge $\mathcal{R}$ aller Assoziationsregeln, die aus I generiert werden kann, wird sehr groß. Für jede Assoziationsregel $X \to Y$ gilt $X, Y \neq \emptyset$ und $X \cap Y = \emptyset$. Sowohl die Prämisse als auch die Konklusion einer jeden Regel in $\mathcal{R}$ besitzt wenigstens ein und höchstens $k - 1$ Items. Es gibt $\binom{k}{p}$ p-elementige Teilmengen von I. Jede solche Teilmenge besitzt wenigstens zwei Elemente. Aus jeder solchen Menge lassen sich für die Prämisse zwischen einem und $p - 1$ Items auswählen, der Rest steckt in der Konklusion. Insgesamt gibt es dann

$$\begin{aligned}
|\mathcal{R}| &= \sum_{p=2}^{k} \left(\binom{k}{p} \cdot \sum_{j=1}^{p-1} \binom{p}{j} \right) \\
&= \sum_{p=2}^{k} \left(\binom{k}{p} \cdot (2^p - 2) \right) \\
&= \sum_{p=0}^{k} \left(\binom{k}{p} \cdot (2^p - 2) \right) + 1 \\
&= \sum_{p=0}^{k} \binom{k}{p} 2^p \cdot 1^{k-p} - 2^{k+1} + 1 \\
&= 3^k - 2^{k+1} + 1
\end{aligned}$$

  mögliche Regeln.

Damit wird es praktisch unmöglich, alle Regeln aufzuzählen bzw. den für alle genannten Gütekriterien notwendigen Support zu bestimmen. Wir möchten nach Vorgabe eines minimalen Support-Wertes, genannt minsup, die Menge aller Itemsets effizient durchsuchen.

Dazu seien Itemsets X, Y gegeben mit $\emptyset \neq Y \subseteq X \subseteq I$. Dann gilt wegen $Y \cap X \setminus Y = \emptyset$

$$\begin{aligned}\text{supp}_{\mathcal{D}}(X) &= \text{supp}_{\mathcal{D}}(Y \cup X \setminus Y) \\ &= \frac{|\{T \in \mathcal{D};\ Y \cup X \setminus Y \subseteq T\}|}{|\mathcal{D}|} \\ &\leq \frac{|\{T \in \mathcal{D};\ Y \subseteq T\}|}{|\mathcal{D}|} \\ &= \text{supp}_{\mathcal{D}}(Y)\end{aligned}$$

Diese Eigenschaft wird als downward closure property bezeichnet[1]. Ist also ein Itemset gefunden, welches den minimal geforderten Support hat, haben sämtliche Teilmengen ebenfalls diesen Minimal-Support. Enthält eine Transaktion ein Itemset X, so auch sämtliche Teilmengen $Y \subseteq X$. Sei $\mathcal{F} := \{F \subseteq I;\ \text{supp}_{\mathcal{D}}(F) \geq \text{minsup}\} \subseteq \mathcal{P}(I)$ die Menge aller Itemsets mit Minimalsupport minsup. Für jedes Itemset $F \in \mathcal{F}$ werden alle nichtleeren echten Teilmengen $Z \subset F$ betrachtet, entsprechend Assoziationsregeln $X \to Z \setminus X$ erzeugt und hinsichtlich der Konfidenz untersucht. Ist die Konfidenz über einem Schwellenwert, so wird die Regel aufgenommen. Wird der Schwellenwert nicht erreicht, folgt für $\emptyset \neq Y \subseteq X$ wegen

$$\text{conf}_{\mathcal{D}}(Y \to F \setminus Y) = \frac{\text{supp}_{\mathcal{D}}(F)}{\text{supp}_{\mathcal{D}}(Y)} \leq \frac{\text{supp}_{\mathcal{D}}(F)}{\text{supp}_{\mathcal{D}}(X)} = \text{conf}_{\mathcal{D}}(X \to F \setminus X),$$

dass auch der Schwellenwert durch die Regel $Y \to F \setminus Y$ nicht erreicht wird. Alle akzeptierten Regeln werden zur Menge $\mathcal{R}' \subseteq \mathcal{R}$ zusammengefasst.

Beim so genannten apriori-Algorithmus werden genau die beiden Schritte durchgeführt: Zunächst wird die Menge $\mathcal{F}$ erzeugt, dann darauf aufbauend die Regelmenge $\mathcal{R}'$. Das Vorgehen nutzt die downward-closure-Eigenschaft des Support entsprechend aus. Voraussetzung ist die lexikographische Sortierung der Items in I.

Bei unserem Titanic-Beispiel legen wir minsup $= 0.1$ und minconf $= 0.7$ fest und beginnen mit dem Support für alle einelementigen Itemsets:

$$\begin{aligned}\text{supp}_{\mathcal{D}}(\{\text{First}\}) &= 0.148, & \text{supp}_{\mathcal{D}}(\{\text{Second}\}) &= 0.130 \\ \text{supp}_{\mathcal{D}}(\{\text{Third}\}) &= 0.321, & \text{supp}_{\mathcal{D}}(\{\text{Crew}\}) &= 0.402 \\ \text{supp}_{\mathcal{D}}(\{\text{Male}\}) &= 0.787 & \text{supp}_{\mathcal{D}}(\{\text{Female}\}) &= 0.214 \\ \text{supp}_{\mathcal{D}}(\{\text{Adult}\}) &= 0.951, & \text{supp}_{\mathcal{D}}(\{\text{Child}\}) &= 0.050 \\ \text{supp}_{\mathcal{D}}(\{\text{Yes}\}) &= 0.323, & \text{supp}_{\mathcal{D}}(\{\text{No}\}) &= 0.677\end{aligned}$$

Die Festlegung von minsup führt zum Ausschluss des Items Child für die weitere Betrachtung. Nun erzeugen wir aus den einelementigen Itemsets zweielementige durch Vereinigungsbildung und bestimmen deren Support. Wir erhalten zwar zunächst 36 solcher Itemsets, allerdings können bereits einige nicht zu gebrauchende wie etwa $\{\text{First, Second}\}$ weggelassen werden. Weiter werden diejenigen entfernt, deren Support kleiner als min-

[1] [8]

sup ist. Es bleiben 15 Itemsets übrig.

$$\begin{aligned}
\text{supp}_{\mathcal{D}}(\{\text{First, Adult}\}) &= 0.145, & \text{supp}_{\mathcal{D}}(\{\text{Second, Adult}\}) &= 0.119\\
\text{supp}_{\mathcal{D}}(\{\text{Third, Male}\}) &= 0.232, & \text{supp}_{\mathcal{D}}(\{\text{Third, Adult}\}) &= 0.285\\
\text{supp}_{\mathcal{D}}(\{\text{Third, No}\}) &= 0.240 & \text{supp}_{\mathcal{D}}(\{\text{Crew, Male}\}) &= 0.392\\
\text{supp}_{\mathcal{D}}(\{\text{Crew, Adult}\}) &= 0.402, & \text{supp}_{\mathcal{D}}(\{\text{Crew, No}\}) &= 0.306\\
\text{supp}_{\mathcal{D}}(\{\text{Male, Adult}\}) &= 0.757, & \text{supp}_{\mathcal{D}}(\{\text{Male, Yes}\}) &= 0.167\\
\text{supp}_{\mathcal{D}}(\{\text{Male, No}\}) &= 0.620, & \text{supp}_{\mathcal{D}}(\{\text{Female, Adult}\}) &= 0.193\\
\text{supp}_{\mathcal{D}}(\{\text{Female, Yes}\}) &= 0.156, & \text{supp}_{\mathcal{D}}(\{\text{Adult,Yes}\}) &= 0.297\\
\text{supp}_{\mathcal{D}}(\{\text{Adult,No}\}) &= 0.653.
\end{aligned}$$

Aus den verbleibenden zweielementigen Itemsets werden dreielementige Itemsets gebildet, indem solche Itemsets vereinigt werden, die genau zwei Items gemeinsam haben und deren sämtliche erzeugbaren zweielementigen Teilmengen wiederum zu den verbleibenden Itemsets gehört. Wir erhalten neun solche dreielementigen Itemsets.

$$\begin{aligned}
\text{supp}_{\mathcal{D}}(\{\text{Third, Male, Adult}\}) &= 0.210, & \text{supp}_{\mathcal{D}}(\{\text{Third, Male, No}\}) &= 0.192\\
\text{supp}_{\mathcal{D}}(\{\text{Third, Adult, No}\}) &= 0.216, & \text{supp}_{\mathcal{D}}(\{\text{Crew, Male, Adult}\}) &= 0.392\\
\text{supp}_{\mathcal{D}}(\{\text{Crew, Male, No}\}) &= 0.304, & \text{supp}_{\mathcal{D}}(\{\text{Crew, Adult, No}\}) &= 0.306\\
\text{supp}_{\mathcal{D}}(\{\text{Male, Adult, Yes}\}) &= 0.154, & \text{supp}_{\mathcal{D}}(\{\text{Male, Adult, No}\}) &= 0.604\\
\text{supp}_{\mathcal{D}}(\{\text{Female, Adult, Yes}\}) &= 0.144.
\end{aligned}$$

Die vierelementigen Itemsets werden analog bestimmt.

$$\begin{aligned}
\text{supp}_{\mathcal{D}}(\{\text{Third, Male, Adult, No}\}) &= 0.176,\\
\text{supp}_{\mathcal{D}}(\{\text{Crew, Male, Adult, No}\}) &= 0.304
\end{aligned}$$

Aus den zwei-, drei- und vierelementigen Itemsets lassen sich Regeln erzeugen. Ohne weitere Vorgaben können wir $15 \cdot 1 \cdot 2 + 9 \cdot 3 \cdot 2 + 2 \cdot 7 \cdot 2 = 112$ Regeln bilden[2]. Interessieren wir uns für Regeln, die als Konklusion ausschließlich die Merkmalswerte des Merkmals Survived enthalten, also Folgerungen bzgl. des Überlebens liefern, bleiben noch 16 Regeln

[2] Dies folgt aus der Anzahl p-elementiger Itemsets mal der Anzahl 2-elementiger Partitionen dieser p-elementigen Itemsets (Stirlingzahl 2. Ordnung) mal 2 (Prämisse und Konklusion tauschen) für jedes $p \geq 2$.

übrig.

Regel	Konfidenz	Lift	Konviktion
{Third} → {No}	0.748	1.105	1.282
{Crew} → {No}	0.760	1.123	1.347
{Male} → {Yes}	0.212	0.656	0.859
{Male} → {No}	0.788	1.002	1.524
{Female} → {Yes}	0.732	2.266	2.526
{Adult} → {Yes}	0.313	0.968	0.985
{Adult} → {No}	0.687	1.015	1.032
{Third, Male} → {No}	0.827	1.222	1.868
{Third, Adult} → {No}	0.759	1.121	1.340
{Crew, Male} → {No}	0.777	1.148	1.449
{Crew, Adult} → {No}	0.760	1.123	1.347
{Male, Adult} → {Yes}	0.203	0.628	0.849
{Male, Adult} → {No}	0.797	1.178	1.593
{Female, Adult} → {Yes}	0.744	2.302	2.644
{Third, Male, Adult} → {No}	0.838	1.237	1.991
{Crew, Male, Adult} → {No}	0.777	1.148	1.449

Gemäß der Festlegung von minconf bleiben die Regeln 1, 2, 4, 5, 8, 9, 10, 11, 13, 14, 15 und 16 übrig. Höhere Werte für Lift und Konviktion (je ≥ 2) führen zum Ausschluss aller Regeln bis auf die Regeln 5 und 14. Die beiden Assoziationsregeln betreffen Überlebende, die Frauen waren. Auf Basis der Berechnungen können wir zumindest den ersten Teil der Frage „Frauen und Kinder zuerst?“ zunächst positiv beantworten. Wegen der Vorgabe von minsup haben wir die Kinder aber gar nicht mehr in Betracht ziehen können.

7.3. Erzeugung von Assoziationsregeln

Wie im letzten Abschnitt offensichtlich geworden ist, können nicht sämtliche Assoziationsregeln von Hand bestimmt werden. Insbesondere ist eine flexible Anpassung der Schranken wie etwa minsup und minconf notwendig. Mit der Wahl von minsup $= 0.1$ beim Titanic-Datensatz werden Regeln, welche das Item Child beinhalten, nicht erfasst. Gerade aber die Frage nach der Überlebenschance von Kindern auf der Titanic ist eine der am häufigsten gestellten Fragen. Deswegen nutzen wir nun R, um Assoziationsregeln zu generieren und zu visualisieren.

Zur Berechnung von Assoziationsregeln benötigen wir das Paket „arules“. Hier ist der R-Code für das Beispiel des letzten Abschnitts. Die Titanic-Daten liegen als Datenmatrix im data.frame-Format in der Variablen titanic.raw vor.

```
library(arules)
rules=apriori(titanic.raw,
        parameter=list(maxlen=4,supp=0.1,conf=0.7),
        appearance=list(rhs=c("Survived=No",
        "Survived=Yes"),default="lhs"))
quality(rules)=round(quality(rules),digits=3)
rules.sorted=sort(rules,by="lift")
inspect(rules.sorted)
m <- interestMeasure(rules,
        c("support", "confidence", "lift","conviction"),
```

```
        transactions = titanic.raw)
m
```

inspect liefert eine Ausgabe der nach lift sortierten Regeln.

```
1  {Sex=Female,
    Age=Adult}   => {Survived=Yes}    0.144    0.744 2.302
2  {Sex=Female} => {Survived=Yes}    0.156    0.732 2.266
3  {Class=3rd,
    Sex=Male,
    Age=Adult}   => {Survived=No}     0.176    0.838 1.237
4  {Class=3rd,
    Sex=Male}    => {Survived=No}     0.192    0.827 1.222
5  {Sex=Male,
    Age=Adult}   => {Survived=No}     0.604    0.797 1.178
6  {Sex=Male}   => {Survived=No}     0.620    0.788 1.164
7  {Class=Crew,
    Sex=Male}    => {Survived=No}     0.304    0.777 1.148
8  {Class=Crew,
    Sex=Male,
    Age=Adult}   => {Survived=No}     0.304    0.777 1.148
9  {Class=Crew} => {Survived=No}     0.306    0.760 1.123
10 {Class=Crew,
    Age=Adult}   => {Survived=No}     0.306    0.760 1.123
11 {Class=3rd,
    Age=Adult}   => {Survived=No}     0.216    0.759 1.121
12 {Class=3rd}  => {Survived=No}     0.240    0.748 1.105
```

Um weitere Gütekriterien anzusehen, benutzen wir interestMeasure. So erhalten wir beispielsweise zusätzlich die Werte für Konviktion.

```
   support confidence  lift conviction
1    0.144      0.744 2.302   2.643761
2    0.156      0.732 2.266   2.525984
3    0.176      0.838 1.237   1.991078
4    0.192      0.827 1.222   1.868443
5    0.604      0.797 1.178   1.593249
6    0.620      0.788 1.164   1.523698
7    0.304      0.777 1.148   1.449196
8    0.304      0.777 1.148   1.449196
9    0.306      0.760 1.123   1.346839
10   0.306      0.760 1.123   1.346839
11   0.216      0.759 1.121   1.339942
12   0.240      0.748 1.105   1.282051
```

Nun möchten wir aber auch Assoziationsregeln die Kinder betreffend erzeugen. Dazu müssen wir den Wert von minsup verringern. Wir versuchen minsup $= 0.005$ und interpretieren wieder die Ausgabe:

```
   support confidence  lift conviction
1    0.011      1.000 3.096        Inf
2    0.006      1.000 3.096        Inf
3    0.064      0.972 3.010  24.181300
```

```
4     0.064       0.972  3.010   24.181300
5     0.042       0.877  2.716    5.504867
6     0.009       0.870  2.692    5.206309
7     0.009       0.870  2.692    5.206309
8     0.036       0.860  2.663    4.836114
9     0.144       0.744  2.302    2.643761
10    0.156       0.732  2.266    2.525984
11    0.070       0.917  1.354    3.888523
12    0.070       0.860  1.271    2.309767
13    0.176       0.838  1.237    1.991078
14    0.192       0.827  1.222    1.868443
15    0.604       0.797  1.178    1.593249
16    0.620       0.788  1.164    1.523698
17    0.304       0.777  1.148    1.449196
18    0.304       0.777  1.148    1.449196
19    0.306       0.760  1.123    1.346839
20    0.306       0.760  1.123    1.346839
21    0.216       0.759  1.121    1.339942
22    0.240       0.748  1.105    1.282051
23    0.016       0.729  1.077    1.192324
```

Die Regeln 1 und 2 haben eine Konviktion von unendlich, d.h. die Konfidenz muss bei 1 sein. Die Regeln 3 und 4 stechen ebenfalls mit hohen Werten für die Konviktion hervor, 5 bis 11 sollten ebenfalls noch einmal näher betrachtet werden.

```
    lhs                     rhs              support confidence lift
1   {Class=2nd,
     Age=Child}          => {Survived=Yes}    0.011      1.000 3.096
2   {Class=2nd,
     Sex=Female,
     Age=Child}          => {Survived=Yes}    0.006      1.000 3.096
3   {Class=1st,
     Sex=Female}         => {Survived=Yes}    0.064      0.972 3.010
4   {Class=1st,
     Sex=Female,
     Age=Adult}          => {Survived=Yes}    0.064      0.972 3.010
5   {Class=2nd,
     Sex=Female}         => {Survived=Yes}    0.042      0.877 2.716
6   {Class=Crew,
     Sex=Female}         => {Survived=Yes}    0.009      0.870 2.692
7   {Class=Crew,
     Sex=Female,
     Age=Adult}          => {Survived=Yes}    0.009      0.870 2.692
8   {Class=2nd,
     Sex=Female,
     Age=Adult}          => {Survived=Yes}    0.036      0.860 2.663
9   {Sex=Female,
     Age=Adult}          => {Survived=Yes}    0.144      0.744 2.302
10  {Sex=Female}        => {Survived=Yes}    0.156      0.732 2.266
11  {Class=2nd,
     Sex=Male,
```

```
     Age=Adult}   => {Survived=No}     0.070     0.917  1.354
12 {Class=2nd,
     Sex=Male}    => {Survived=No}     0.070     0.860  1.271
13 {Class=3rd,
     Sex=Male,
     Age=Adult}   => {Survived=No}     0.176     0.838  1.237
14 {Class=3rd,
     Sex=Male}    => {Survived=No}     0.192     0.827  1.222
15 {Sex=Male,
     Age=Adult}   => {Survived=No}     0.604     0.797  1.178
16 {Sex=Male}     => {Survived=No}     0.620     0.788  1.164
17 {Class=Crew,
     Sex=Male}    => {Survived=No}     0.304     0.777  1.148
18 {Class=Crew,
     Sex=Male,
     Age=Adult}   => {Survived=No}     0.304     0.777  1.148
19 {Class=Crew}   => {Survived=No}     0.306     0.760  1.123
20 {Class=Crew,
     Age=Adult}   => {Survived=No}     0.306     0.760  1.123
21 {Class=3rd,
     Age=Adult}   => {Survived=No}     0.216     0.759  1.121
22 {Class=3rd}    => {Survived=No}     0.240     0.748  1.105
23 {Class=3rd,
     Sex=Male,
     Age=Child}   => {Survived=No}     0.016     0.729  1.077
```

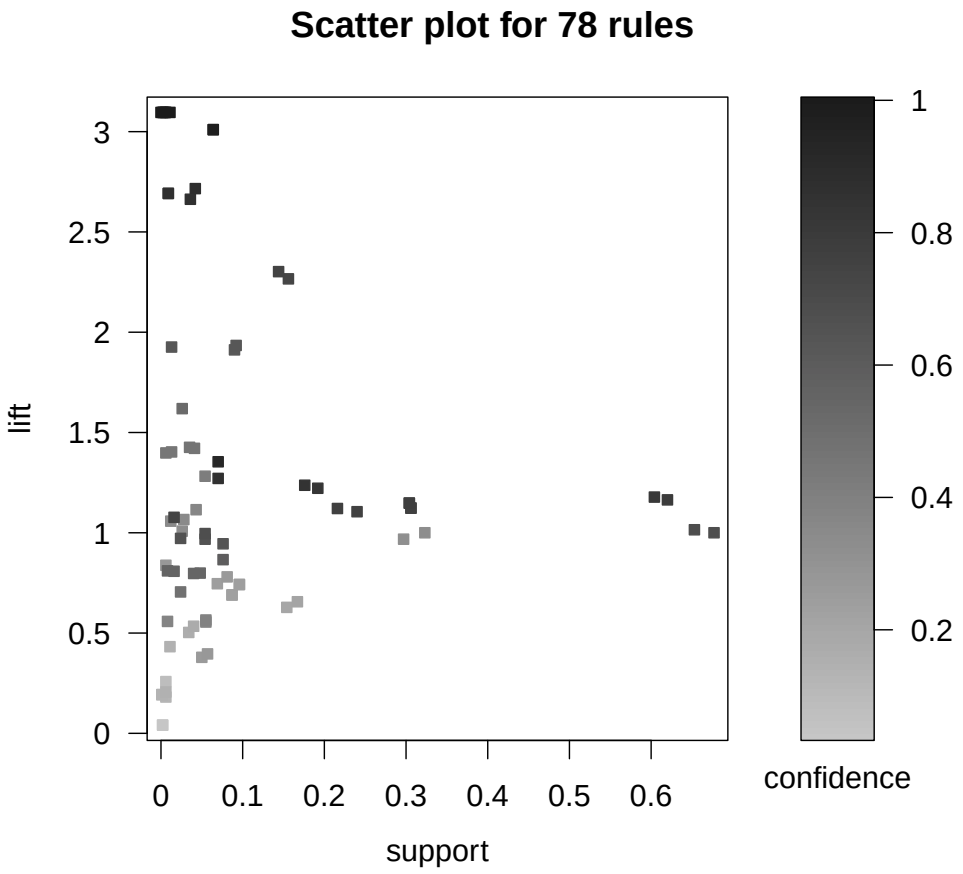
Es fällt bei Betrachten der Regeln 1 und 2 auf, dass die Regel $\{\text{Second, Male, Child}\} \rightarrow \{\text{Yes}\}$ fehlt. Dies liegt daran, dass der Support knapp unter 0.005 liegt. Konfidenz und Konviktion sind identisch zu den beiden Regeln 1 und 2. Ebenso wenig gibt es Regeln für Kinder der ersten und zweiten Klasse. Hier haben alle überlebt, jedoch gab es nur sechs Kinder in den beiden Klassen. In der Crew gab es keine Kinder. Die Regeln 2 bis 10 betreffen überlebende Frauen. Die zu betrachtenden Assoziationsregeln für Kinder und Frauen der dritten Klasse werden aufgrund ihrer geringen Konfidenz nicht berücksichtigt. Die Regeln verallgemeinern sich entsprechend ihrer aufsteigenden Numerierung. Somit kann die Frage „Frauen und Kinder zuerst?“ positiv beantwortet werden für Frauen und Kinder, die nicht der dritten Klasse angehört haben. Für Frauen und Kindern aller anderen Klassen war gemäß der erhaltenen Assoziationsregeln die Chance zu Überleben deutlich erhöht. Die Regel 2 liefert gegenüber der Regel 1 keine neue Information, sie ist redundant, da ihr Lift nicht größer ist als der der Regel 1 und ihre Prämisse eine Teilmenge der Prämisse von Regel 1 darstellt. Solche Regeln lassen sich folgendermaßen beseitigen:

```
subset.matrix=is.subset(rules.sorted,rules.sorted)
subset.matrix[lower.tri(subset.matrix,diag=T)]=NA
redundant=colSums(subset.matrix,na.rm=T)>=1
which(redundant)
rules.pruned=rules.sorted[!redundant]
inspect(rules.pruned)
```

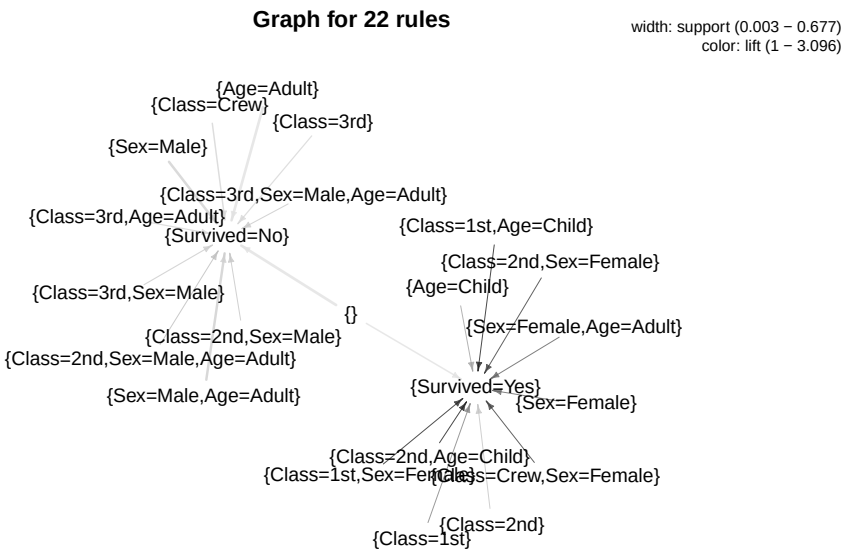
Die Regeln 2, 4, 7 und 8 sind redundant, Zuletzt sehen wir uns die Regel 11 an. Es ist die einzige der ausgewählten Regeln, die ein Nicht-Überleben folgern. Dies ergibt sich für erwachsene Männer der zweiten Klasse. Nur 14 von 168 Männern der zweiten Klasse ha-

ben tatsächlich überlebt. Die Assoziationsregel kann sicher akzeptiert werden.

Informationen über Assoziationsregeln $X \to Y$ lassen sich visualisieren. Eine Möglichkeit besteht in einem Scatterplot, wobei die Regeln als Objekte und Werte ihrer Gütekriterien dargestellt werden. Eine Möglichkeit besteht in der Darstellung von Support gegen Lift. Bei einem hohen Support von $X \cup Y$ ist auch der Support von X und von Y entsprechend hoch, mindestens so hoch wie bei der Vereinigungsmenge. Damit wird der Lift nicht sehr stark von eins abweichen können. Im Gegensatz dazu sind bei einem kleinen Support von $X \cup Y$ viel größere Schwankungen um die eins für den Lift-Wert möglich. Dies zeigt sich an folgendem Plot für alle 78 möglichen Assoziationsregeln im Titanic-Datensatz.



Weiter ist zu erkennen, dass die einzelnen Objekte (die Assoziationsregeln) farblich hinsichtlich der Konfidenz gekennzeichnet sind. Dies liefert zusätzliche Information. So lässt sich aus der Konfidenz einer Regel und deren Support etwas über den Support der Konklusion aussagen. Die Assoziationsregeln lassen sich darstellen, indem die Prämissen und Konklusionen über gerichtete Kanten verbunden sind. Je nach Länge und farblicher Darstellung der Kanten werden die Regeln interpretiert. Je länger die Kanten sind, desto höher ist der Support, je dunkler desto größer der Lift der jeweiligen Regel.



Bemerkung.
Mit den bisherigen Assoziationsregeln können wir noch keine Regeln der Form

$$\{\text{(First oder Second oder Third) und Female}\} \to \{\text{Yes}\}$$

bilden, da wir die oder-Verknüpfung nicht addressieren können.

Simpsonsches Paradoxon

Bei näherer Betrachtung fallen folgende vier in gewissem Sinn zusammengehörige Regeln auf:

Assoziationsregel	Support	Konfidenz
$\{\text{Crew, Female}\} \to \{\text{Yes}\}$	0.009	0.870
$\{\text{passenger, Female}\} \to \{\text{Yes}\}$	0.147	0.725
$\{\text{Crew, Male}\} \to \{\text{Yes}\}$	0.087	0.223
$\{\text{passenger, Male}\} \to \{\text{Yes}\}$	0.154	0.203

Die Zahlen zeigen: Die Chance zu überleben war bei Passagieren wohl geringer als bei Mitgliedern der Crew.

Doch stimmt das wirklich? Betrachten wir den Anteil Überlebender unter allen Crew-Mitgliedern, so ist dieser mit 0.24 geringer als der Anteil Überlebender aller Passagiere mit 0.38.

Satz und Definition 7.3: Rechnen mit bedingten Wahrscheinlichkeiten

Sei $(\Omega, \mathcal{F}, \mathbb{P})$ ein beliebiger Wahrscheinlichkeitsraum. Gegeben seien drei Ereignisse $A, B, C \in \mathcal{F}$ und deren Gegenereignisse. Mit Hilfe bedingter Wahrscheinlichkeiten gilt

dann:

$$\begin{aligned}
\mathbb{P}(A|B) &= \mathbb{P}(A \cap \Omega|B) = \mathbb{P}(A \cap (C \cup \neg C)|B) \\
&= \mathbb{P}((A \cap C) \cup (A \cap \neg C)|B) = \mathbb{P}(A \cap C|B) + \mathbb{P}(A \cap \neg C|B) \\
&= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B)} + \frac{\mathbb{P}(A \cap B \cap \neg C)}{\mathbb{P}(B)} \\
&= \mathbb{P}(A|B \cap C) \cdot \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(B)} + \mathbb{P}(A|B \cap \neg C) \cdot \frac{\mathbb{P}(B \cap \neg C)}{\mathbb{P}(B)} \\
&= \mathbb{P}(A|B \cap C) \cdot \mathbb{P}(C|B) + \mathbb{P}(A|B \cap \neg C) \cdot \mathbb{P}(\neg C|B)
\end{aligned}$$

Analog folgt $\mathbb{P}(A|\neg B) = \mathbb{P}(A|\neg B \cap C) \cdot \mathbb{P}(C|\neg B) + \mathbb{P}(A|\neg B \cap \neg C) \cdot \mathbb{P}(\neg C|\neg B)$. Die bedingten Wahrscheinlichkeiten für das Ereignis A unter $(\neg)B$ sind gewichtete Summen von Wahrscheinlichkeiten für A unter $(\neg)B$ und C bzw. $(\neg)B$ und $\neg C$.

Im Titanic Beispiel ist entsprechend $A = \{\text{Yes}\}$, $B = \{\text{passenger}\}$ und $C = \{\text{Female}\}$. Wir erhalten

$$\begin{aligned}
\mathbb{P}(A|B) &= \frac{324}{447} \cdot \frac{447}{1316} + \frac{175}{869} \cdot \frac{869}{1316} = \frac{499}{1316} \\
&= 0.72 \cdot 0.34 + 0.20 \cdot 0.66 = 0.38, \\
\mathbb{P}(A|\neg B) &= \frac{20}{23} \cdot \frac{23}{885} + \frac{192}{862} \cdot \frac{862}{885} = \frac{212}{885} \\
&= 0.87 \cdot 0.03 + 0.22 \cdot 0.97 = 0.24.
\end{aligned}$$

Die Gewichte führen zur umgekehrten Aussage. Die Ursache liegt darin, dass die Frauen bzgl. der Crew weit unterrepräsentiert sind. Unter den Crew-Mitgliedern gab es weniger als 3% Frauen.

8. Regressionsanalyse

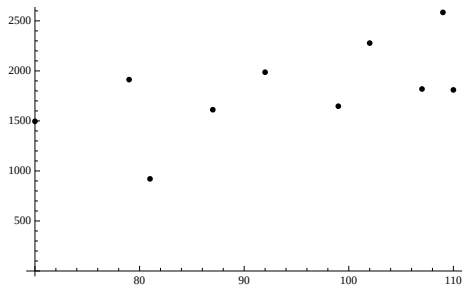
Warum?

Es sollen von Untersuchungsobjekten Werte einer Eigenschaft vorhergesagt werden. Dabei soll sich die Vorhersage auf andere zu erhebende Eigenschaften stützen, deren Wertemenge jeweils als die Menge $\mathbb{R}$ angesetzt wird. Ist es möglich, auf Daten zu allen Eigenschaften zurückzugreifen, so kann darauf basierend (supervised) ein Modell erzeugt werden, das die Vorhersage bei einem neuen Untersuchungsobjekt ohne Kenntnis der vorherzusagenden Eigenschaft ermöglicht.

Soll ein funktionaler Zusammenhang zwischen einer kardinal skalierten abhängigen und einer oder mehreren anderen Variablen beliebiger Skala modelliert werden, wird die Regressionsanalyse eingesetzt. Wir betrachten die lineare Regressionsanalyse. Es seien mit Y die abhängige und mit $X_1, \ldots, X_J$ die unabhängigen Variablen bezeichnet. Ein funktionaler Zusammenhang bildet einen Kausalzusammenhang zwischen Y und $X_1, \ldots, X_J$ ab.

8.1. Exploration und Modellformulierung

Bei jeder Modellerstellung sollen sachlogische Überlegungen eine zentrale Rolle einnehmen. Diese Überlegungen können durch interaktive statistische Graphiken unterstützt werden. So kann in unserem Rechenbeispiel ein kausaler Zusammenhang zwischen dem Absatz als abhängige Variable (Y) und den unabhängigen Variablen Vertreterbesuche (X_1), Preis (X_2) und Ausgaben (X_3) angesetzt werden. Ein Scatterplot zwischen der abhängigen und einer unabhängigen Variablen kann einen möglichen funktionalen Zusammenhang suggerieren. In unserem Rechenbeispiel (vgl. [5]) wird die Absatzmenge gegen die Anzahl der Vetreterbesuche geplottet und es lässt sich eine gewisse lineare Tendenz erkennen.



Ein lineares Modell kann durch

$$f : \mathbb{R}^J \to \mathbb{R},\ (X_1, \ldots, X_J) \mapsto \hat{Y} := f(X) := b_0 + \sum_{j=1}^{J} b_j X_j \tag{8.1}$$

für gewisse noch zu bestimmende $b_0, \ldots, b_J \in \mathbb{R}$ angesetzt werden und wird Regressionsmodell genannt. Das Dach über dem Y deute die Schätzung der Y-Werte durch das Modell an. Denn es besteht im Regelfall kein exakt linearer Zusammenhang zwischen den Variablen. Im einfachsten Fall hängt die abhängige Variable von genau einer unabhängigen Variablen ab. So kann das Modell

$$\text{„MENGE = } b_0 + b_1 \cdot \text{VERTRETERBESUCHE“},$$

$$\hat{Y} = b_0 + b_1 X_1$$

angesetzt werden. Es stellt sich die Frage, wie die Koeffizienten b_0 und b_1 geschätzt werden können. Dazu wird ein Kriterium benötigt, anhand dessen eine Schätzung vorgenommen werden kann.

8.2. Schätzung der Regressionsfunktion

Das Regressionsmodell ist im vorliegenden Fall geometrisch gesehen eine Gerade, die nicht alle Punkte durchläuft. Die Gerade kann so festgelegt werden, dass die Abstände zu den gegebenen Punkten möglichst klein werden. Die Differenz des i-ten Schätzwertes $\hat{y}_i = b_0 + b_1 x_i$ zum gemessenen Wert y_i kann positiv oder negativ werden. Durch Aufsummieren können sich die Schätzfehler gegenseitig auslöschen, was berücksichtigt werden muss. Eine Möglichkeit besteht darin, nur positive Differenzwerte zu erzeugen, etwa durch Betragsbildung oder Quadrieren der Fehler. Auf diese Weise lassen sich die folgenden beiden Kriterien formulieren.

Least Absolute Deviation (LAD, L_1-Problem)

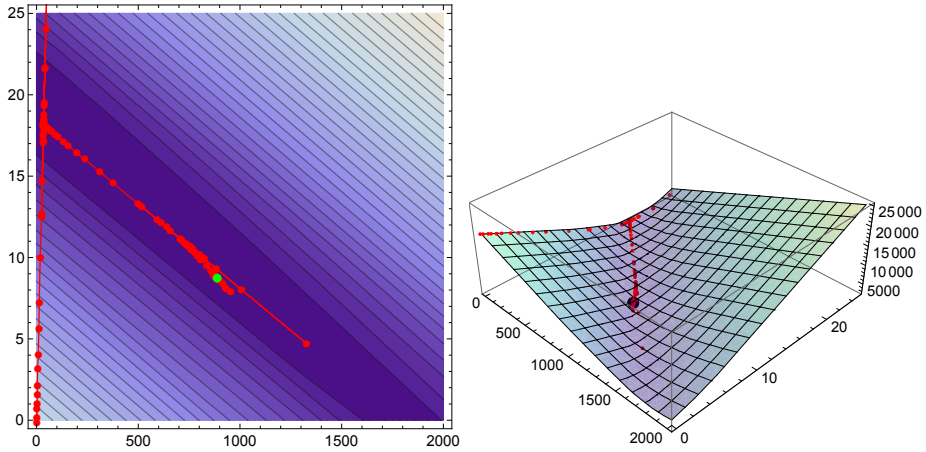
Eine intuitive Wahl ist es, die Beträge der Fehler zu verwenden. Es kann demnach versucht werden, die Summe der absoluten Schätzfehler,

$$S_{\text{LAD}} : \mathbb{R}^2 \to \mathbb{R},\ (b_0, b_1) \mapsto S_{\text{LAD}}(b_0, b_1) := \sum_{i=1}^{n} |y_i - (b_0 + b_1 x_i)|$$

zu minimieren, d.h. es ergibt sich das Optimierungsproblem

$$\min_{b_0, b_1 \in \mathbb{R}} \{S_{\text{LAD}}(b_0, b_1)\}.$$

Die Betragsstriche ermöglichen keine partiellen Ableitungen nach den beiden Parametern b_0 und b_1 und somit verhindern sie eine analytische Lösung des Problems. Es muss ein (numerisches) globales Optimierungsverfahren verwendet werden. Es ist nicht auszuschließen, dass es keine eindeutige Optimallösung des Problems gibt. Die Zielfunktion ist in den beiden folgenden Abbildungen illustriert.



Ein ableitungsfreies Verfahren wie das Nelder-Mead-Verfahren liefert für das Rechenbeispiel als Ergebnis $b_0 = 884.92$ und $b_1 = 8.73$ bei einem L_1-Fehler von $S_{\text{LAD}} = 2729.35$. Für den Wert $x_1 = 109$ lässt sich ein Schätzwert $\hat{y}_1 = 1836.49$ schätzen. Der gemessene Wert liegt bei 2585, es ergibt sich ein absoluter Fehler von 748.51.

```
x=c(109,107,99,70,81,102,110,92,87,79)
y=c(2585,1819,1647,1496,921,2278,1810,1987,1612,1913)
plot(x,y)
cor(x,y)
lad=function(b)
{       b1<-b[1];
        b2<-b[2];
        return(sum(abs(y-(b1+b2*x))))
}
optlad=optim(c(1,2),lad)
lines(x,optlad$par[1]+x*optlad$par[2])
```

In der Praxis hat sich diese Modellwahl nicht durchgesetzt, da sich damit keine einfache statistische Überprüfung des Modells und der Modellparameter durchführen lässt.

Least Squares Error (LSE, L_2-Problem)

Eine weitere Möglichkeit besteht in der so genannten Kleinst-Quadrate-Schätzung. Dabei werden die entstehenden Schätzfehler quadriert und aufsummiert,

$$S_{\text{LSE}} : \mathbb{R}^2 \to \mathbb{R},\ (b_0, b_1) \mapsto S_{\text{LSE}}(b_0, b_1) := \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 .$$

In diesem Fall ist das Optimierungsproblem

$$\min_{b_0, b_1 \in \mathbb{R}} \{S_{\text{LSE}}(b_0, b_1)\}$$

zu lösen. Für diese Zielfunktion lassen sich die partiellen Ableitungen nach b_0 und b_1 be-

stimmen.

$$\begin{aligned}\frac{\partial S_{\text{LSE}}}{\partial b_0}(b_0,b_1) &= \sum_{i=1}^{n} 2(y_i-(b_0+b_1x_i))\cdot(-1)\\ &= -2\sum_{i=1}^{n} y_i + 2n\cdot b_0 + 2b_1\sum_{i=1}^{n} x_i \qquad (8.2)\\ \frac{\partial S_{\text{LSE}}}{\partial b_1}(b_0,b_1) &= \sum_{i=1}^{n} 2(y_i-(b_0+b_1x_i))\cdot(-x_i)\\ &= -2\sum_{i=1}^{n} x_iy_i + 2b_0\sum_{i=1}^{n} x_i + 2b_1\sum_{i=1}^{n} x_i^2 \qquad (8.3)\end{aligned}$$

Durch Nullsetzen der partiellen Ableitung (8.2) ergibt sich

$$b_0 = \bar{y} - b_1\bar{x}. \qquad (8.4)$$

Nullsetzen der partiellen Ableitung (8.3) und Einsetzen von (8.4) führt zu

$$\begin{aligned}&-\sum_{i=1}^{n} x_iy_i + (\bar{y}-b_1\bar{x})\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i^2 = 0\\ \Leftrightarrow\quad & b_1 = \frac{\bar{y}\sum\limits_{i=1}^{n} x_i - \sum\limits_{i=1}^{n} x_iy_i}{\bar{x}\sum\limits_{i=1}^{n} x_i - \sum\limits_{i=1}^{n} x_i^2} = \frac{\sum\limits_{i=1}^{n} x_i\sum\limits_{i=1}^{n} y_i - n\sum\limits_{i=1}^{n} x_iy_i}{\left(\sum\limits_{i=1}^{n} x_i\right)^2 - n\sum\limits_{i=1}^{n} x_i^2}. \qquad (8.5)\end{aligned}$$

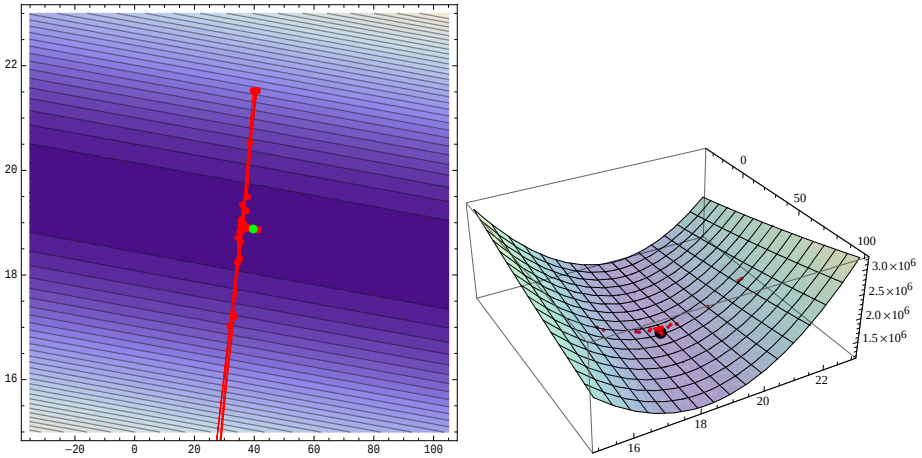
Die Hessematrix liefert den Nachweis dafür, dass die gefundenen Werte für b_0 und b_1 einen Minimalwert liefern.

$$H_{S_{\text{LSE}}} = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum\limits_{i=1}^{n} x_i^2 \end{pmatrix}$$

Diese Matrix ist für $x_i \neq \bar{x}$ für alle $i = 1,\ldots,n$ positiv definit, da $2n > 0$ und $4n\sum\limits_{i=1}^{n} x_i^2 - 4n^2\bar{x}^2 = 4n\sum\limits_{i=1}^{n}(x_i-\bar{x})^2 > 0$ ist. Wir berechnen für diesen Ansatz die Parameter:

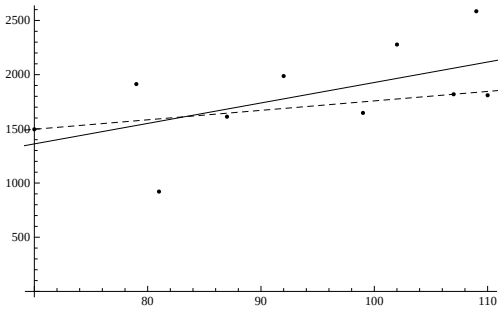
i	x_i	y_i	x_i^2	x_iy_i
1	109	2585	11881	281765
2	107	1819	11449	194633
3	99	1647	9801	163053
4	70	1496	4900	104720
5	81	921	6561	74601
6	102	2278	10404	232356
7	110	1810	12100	199100
8	92	1987	8464	182804
9	87	1612	7569	140244
10	79	1913	6241	151127
Σ	936	18068	89370	1724403

Damit ist $\bar{x} = 93.6$ und $\bar{y} = 1806.8$ und es ergibt sich $b_1 = \frac{936\cdot 18068 - 10\cdot 1724403}{936^2 - 10\cdot 89370} = 18.881$ bzw. $b_0 = 1806.8 - 18.881\cdot 93.6 = 39.5$.



Die Regressionsfunktion lautet $\hat{Y} = 39.5 + 18.881 \cdot X_1$. Für den Wert $x_1 = 109$ lässt sich ein Schätzwert $\hat{y}_1 = 2097.57$ schätzen. Der gemessene Wert liegt bei 2585, was zu einem quadrierten Fehler von 237588 führt. Durch Aufsummieren ergibt sich ein L_2-Fehler von 1188684.94.

```
lse=function(b)
{       b1<-b[1];
        b2<-b[2];
        return(sum((y-(b1+b2*x))^2))
}
optlse=optim(c(1,2),lse)
optlse=optim(c(885,9),lse)
mod1=lm(y~x)
mod1
mod1$fitted.values
lines(x,mod1$fitted.values)
```



Welche der beiden Regressionsgeraden ist „besser“? Jede ist im Sinne der gewählten Norm optimal. Ein zusätzlicher Besuch führt zu einer Absatzerhöhung um 8.73 bzw. 18.881 Einheiten. Dies ist jeweils eine folgerichtige Interpretation. Doch wie soll damit umgegangen werden? Dafür gibt es keine absolut richtige Antwort. Jedoch sollte man stets aufmerksam solche Interpretationen hinterfragen.

In wenigen Fällen genügt eine Einflussgröße zur Beschreibung einer Ergebnisgröße. Oftmals sind weitaus komplexere Zusammenhänge zu erwarten. Das Regressionsmodell kann dann auf das Modell (8.1) erweitert werden. Unter dem LSE-Modell ist zur Bestimmung der Parameter $b_0, \ldots, b_J$ die Funktion

$$S_{\text{LSE}} : \mathbb{R}^{J+1} \to \mathbb{R},\ (b_0, \ldots, b_J) \mapsto S_{\text{LSE}}(b_0, \ldots, b_J) := \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{J} b_j x_{ij} \right)^2 . \qquad (8.6)$$

zu minimieren, wobei x_{ij} die i-te Messung der j-ten Einflussgröße darstellt. Den Subtrahenden in der Klammer in (8.6) können wir schreiben als

$$b_0 + \sum_{j=1}^{J} b_j x_{ij} = \mathbf{x}_{i\cdot}^T \mathbf{b}$$

mit $\mathbf{x}_{i\cdot}^T = (1, x_{i1}, \ldots, x_{iJ})$ und $\mathbf{b}^T = (b_0, \ldots, b_J)$. Es lässt sich so eine Matrix $X \in \mathbb{R}^{n,J+1}$,

$$\begin{pmatrix} 1 & x_{11} & \ldots & x_{1J} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nJ} \end{pmatrix},$$

die Designmatrix, aufstellen. Mit $\mathbf{y}^T = (y_1, \ldots, y_n)$ folgt

$$S_{\text{LSE}}(b_0, \ldots, b_J) = (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) = \mathbf{y}^T\mathbf{y} - \mathbf{b}^T X^T\mathbf{y} - \mathbf{y}^T X\mathbf{b} + \mathbf{b}^T X^T X\mathbf{b}.$$

Der Gradient und die Hessematrix dieser Funktion ergeben sich aus den partiellen Ableitungen zu

$$\nabla S_{\text{LSE}}^T(b_0, \ldots, b_J) = -2X^T\mathbf{y} + 2X^T X\mathbf{b}, \tag{8.7}$$

$$H_{S_{\text{LSE}}}(b_0, \ldots, b_J) = 2X^T X. \tag{8.8}$$

Die Untersuchung der Hesse-Matrix $2X^T X$ der Funktion S_{LSE} zeigt wegen

$$2\mathbf{v}^T X^T X\mathbf{v} = 2(X\mathbf{v})^T(X\mathbf{v}) \geq 0$$

für alle $\mathbf{v} \neq \mathbf{0}$, $\mathbf{v} \in \mathbb{R}^{J+1}$, dass sie positiv-semidefinit ist. Die reelle symmetrische Matrix $X^T X$ ist diagonalisierbar. Ein Eigenwert 0 ergibt sich genau dann, wenn der Kern von $X^T X$ nicht nur aus dem Nullvektor besteht, also wenn $\text{rg}(X^T X) < J+1$ ist. Mit $\text{rg}(X^T X) = J+1$ ist $X^T X$ positiv definit. Nullsetzen des Gradienten ergibt im Falle der Existenz der Inversen von $X^T X$ als einzige Nullstelle

$$\hat{\mathbf{b}} = (X^T X)^{-1} X^T \mathbf{y}, \tag{8.9}$$

welche damit ein globales Minimum ist.

Sei in unserem Rechenbeispiel das Modell

„MENGE = $b_0 + b_1$·VERTRETERBESUCHE+b_2·PREIS+b_3·AUSGABEN“,

aufgestellt. Die Designmatrix ergibt sich zu

$$X = \begin{pmatrix} 1 & 109 & 12.5 & 2000 \\ 1 & 107 & 10 & 550 \\ 1 & 99 & 9.95 & 1000 \\ 1 & 70 & 11.5 & 800 \\ 1 & 81 & 12 & 0 \\ 1 & 102 & 10 & 1500 \\ 1 & 110 & 8 & 800 \\ 1 & 92 & 9 & 1200 \\ 1 & 87 & 9.5 & 1100 \\ 1 & 79 & 12.5 & 1300 \end{pmatrix}$$

Mit Hilfe der Gleichung (8.9) erhalten wir $\hat{\mathbf{b}}^T = (-6.866, 11.086, 9.927, 0.655)$. Die Parameter lassen sich folgendermaßen interpretieren: Eine Erhöhung einer Variablen um eine Einheit, verändert den Absatz um die entsprechende Komponente. Eine Erhöhung der Ausgaben für Marketing um eine Einheit erhöht den Absatz um 0.655 Einheiten. Ebenso erhöht die Anhebung des Preises um eine Einheit den Absatz um 9.927 Einheiten. Hier wird eine Schwierigkeit bei der Interpretation solcher Modelle klar. Eine Preis-Absatz-Funktion weist im Normalfall fallenden Charakter auf. Je höher der Preis, desto geringer wird der Absatz sein (vgl. [5]). Hier wird jedoch ein Anstieg herausgelesen.

```
absatz=y
vertreterbesuche=x
preis=c(12.5,10,9.95,11.5,12,10,8,9,9.5,12.5)
ausgaben=c(2000,550,1000,800,0,1500,800,1200,1100,1300)
plot(preis,absatz)
cor(preis,absatz)
plot(ausgaben,absatz)
cor(ausgaben,absatz)
mod2=lm(y~vertreterbesuche+preis+ausgaben)
mod2
mod2$fitted.values
plot(vertreterbesuche,absatz)
lines(vertreterbesuche,mod2$fitted.values,type="p",col=5)
plot(preis,absatz)
lines(preis,mod2$fitted.values,type="p",col=5)
plot(ausgaben,absatz)
lines(ausgaben,mod2$fitted.values,type="p",col=5)
plot(mod2$fitted.values,ausgaben)
plot(mod2$fitted.values,preis)
plot(mod2$fitted.values,absatz)
```

Das Modell mit seinen errechneten Koeffizienten muss überprüft werden. Dazu gibt es verschiedene Möglichkeiten. Das Bestimmheitsmaß ist beispielsweise ein Varianzkriterium, welches die Abweichungen der geschätzten Werte $\hat{y}_i$ von den gemessenen Werten y_i in Beziehung zu den Abweichungen der y_i zum Mittelwert $\bar{y}$ setzt. Um weitere Möglichkeiten der Überprüfung des Modells und seiner Annahmen treffen zu können, ist eine stochastische Modellierung notwendig.

8.3. Modell der multiplen Regression

Gegeben sei eine Stichprobe $\mathbf{y} \in M^n$, basierend auf einer Realisierung einer Zufallsvariablen $Y : \Omega \to M^n$. Seien weiter $\mathbf{x}_u$ dazugehörige Beobachtungsvektoren für $u = 1, \ldots, l$ mit $\mathbf{x}_u = (x_{1u}, \ldots, x_{nu})^T$ und $x_{iu} \in M_u$ für alle $i = 1, \ldots, n$, die unabhängigen Variablen. Betrachtet werden soll das erweiterte additive Modell

$$z(y_i) \quad = \quad f(g_1(x_{i1}, \ldots, x_{il}), \ldots, g_k(x_{i1}, \ldots, x_{il})) + \epsilon_i, \tag{8.10}$$

wobei

- $z : M \to \mathbb{R}$ eine messbare Transformation ohne Parametrisierung der y_i ist,
- der Vektor $\mathbf{z} := (z(y_1), \ldots, z(y_n))^T$ der Realisierungsvektor der Zufallsvariablen $Z := z(Y) : M^n \to \mathbb{R}^n$, der abhängigen Variablen, ist,

- $g_j : M_1 \times \ldots \times M_l \to \mathbb{R}$ für jedes $j = 1, \ldots, k$ eine Transformation ohne Parametrisierung der unabhängigen Variablen ist,
- die Matrix $X := (g_j(x_{i1}, \ldots, x_{il})) \in \mathbb{R}^{n,k}$ für $j = 1, \ldots, k$, $i = 1, \ldots, n$ die so genannte Designmatrix ist,
- $\boldsymbol{\epsilon} \in \mathbb{R}^n$ der Realisierungsvektor der Zufallsvariablen $E : \Omega \to \mathbb{R}^n$ ist und eine additive zufällige Komponente des Modells repräsentiert,
- die Abbildung $f : \mathbb{R}^k \to \mathbb{R}$ die systematische Komponente des Modells darstellt.

Eine weitere Annahme besteht nun darin,

- die Abbildung f als Linearkombination der g_j mit den Parametern $\beta_1, \ldots, \beta_k \in \mathbb{R}$ zu modellieren,

$$z_i = z(y_i) = \beta_1 g_1(x_{i1}, \ldots, x_{il}) + \ldots + \beta_k g_k(x_{i1}, \ldots, x_{il}) + \epsilon_i.$$

Ziel ist es dann, aus den gegebenen Daten die systematische Komponente „bestmöglich“ herauszuarbeiten und entsprechend die $\beta_1, \ldots, \beta_k$ nach noch vorzugebenden Kriterien zu bestimmen.

Beispiel 8.1

Sei unter der Annahme reellwertiger Merkmale $l = 1$, $k = 2$, $g_1(x_{i1}) = x_{i1}^0 = 1$, $g_2(x_{i1}) = x_{i1}$ und $z(y_i) = y_i$. Dann ergibt sich die Modellannahme

$$y_i = \beta_1 + \beta_2 x_{i1} + \epsilon_i.$$

Das Modell lässt sich für alle Daten als

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\beta} = (\beta_\mathbf{1}, \ldots, \beta_\mathbf{k})^\mathbf{T}, \tag{8.11}$$

für aus der selben Verteilungsfamilie stammende $Y_i \sim \mathbb{P}_{\bar{Y}, \psi \in \Psi}$ mit einer Zufallsvariablen $\bar{Y} : \Omega \to M$ und für identisch verteilte $E_i \sim \mathbb{P}_{\bar{E}, \phi \in \Phi}$ mit einer Zufallsvariablen $\bar{E} : \Omega \to \mathbb{R}$ und mit $\mathbf{x} = (g_1(x_1, \ldots, x_l), \ldots, g_k(x_1, \ldots, x_l)) \in \mathbb{R}^k$ mit fest gewählten $x_u \in M_u$ als

$$\bar{Z} = z(\bar{Y}) = \mathbf{x}^T\boldsymbol{\beta} + \bar{E}, \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T, \tag{8.12}$$

schreiben. Ist in der Designmatrix X gemäß Modell (8.11) eine Spalte der Matrix eine Linearkombination der anderen Spalten, so liefert diese Spalte keinerlei neue Information für das Modell. Sind weniger Beobachtungsdaten als Modellparameter vorhanden, so kann es passieren, dass keine eindeutige Festlegung der Parameter möglich ist. An der jeweiligen Stelle muss dann $n > k$ berücksichtigt werden. Somit sei eine weitere Forderung an die Designmatrix $X \in \mathbb{R}^{n,k}$ gegeben durch:

- $\mathrm{rg}(X) = k$.

Diese Forderung werden wir auch aus mathematischer Sicht begründen. In der Praxis ist die Rangforderung nicht immer haltbar oder es ist eine nahezu lineare Abhängigkeit zwischen verschiedenen Spalten gegeben. Letzteres führt zu Schwierigkeiten bei der Interpretation der Parameter, da kleine Änderungen in der Designmatrix zu großen Änderungen

bei den Parameterwerten führen können. Die letzte Überlegung betrifft die unabhängigen Variablen. Sie können entweder deterministisch oder stochastisch sein. Im Falle stochastischer unabhängiger Variabler können die nachfolgenden Untersuchungen stets unter der Bedingung der gegebenen Designmatrix durchgeführt werden. Eine stochastische Komponente im Modell (8.12) ist der Fehler. Um das Modell anpassen und überprüfen zu können, werden bzgl. des Fehlers zunächst mit $\sigma^2 > 0$ folgende klassische Annahmen getroffen:

- Der Fehler ist im Mittel Null, $\mathbb{E}_{(\boldsymbol{\beta}),\sigma^2}[E] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[E \mid X] = \mathbf{0}$,
- die Fehler sind untereinander unkorreliert, $\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E_i, E_j] = \mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E_i, E_j \mid X] = 0$ und
- die Varianz der Fehler bleibt konstant, $\mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[E_i] = \mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[E_i \mid X] = \sigma^2 > 0$ (Homoskedastizität).

Für homoskedastische und unkorrelierte Fehler lässt sich die Modellannahme in Form einer diagonalen Varianz-Kovarianzmatrix $\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E \mid X] = \sigma^2 I^{n,n}$ formulieren. Im Falle heteroskedastischer Fehler ist für jeden Fehler eine eigene Varianz anzusetzen, $\mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[E_i \mid X] = \sigma_i^2$. Sind die Fehler weiterhin unkorreliert, lässt sich das durch $\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E \mid X] = \sigma^2 \mathrm{diag}(w_{11}^2, \ldots, w_{nn}^2)$ darstellen. Bei zusätzlich korrelierten Fehlern gelte $\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E \mid X] = \sigma^2 W$ für eine positiv definite Matrix $W \in \mathbb{R}^{n,n}$.

Die Abhängigkeit von X bei stochastischen, unabhängigen Variablen soll im folgenden nicht mehr notiert werden. Der für die Mathematische Statistik als Ausgangspunkt dienende statistische Raum ist mittels der Zufallsvariablen $\bar{Z} = z(\bar{Y})$ durch $(\mathbb{R}, \mathcal{B}, \mathbb{P}_{\bar{Z},\theta\in\Theta})$ gegeben. Die Verteilung von $\bar{Z}$ hängt von den Parametern $\boldsymbol{\beta}$ und einer Verteilungsannahme für $\bar{E}$ ab. Denn es gilt bei Gültigkeit der Annahmen allgemein

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[\bar{Z}] &= \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[\mathbf{x}^T\boldsymbol{\beta} + \bar{E}] = \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[\mathbf{x}^T\boldsymbol{\beta}] + \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[\bar{E}] = \mathbf{x}^T\boldsymbol{\beta} + 0 \\
&= \mathbf{x}^T\boldsymbol{\beta}, && (8.13)\\
\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[Z] &= \mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[X\boldsymbol{\beta} + E] = X\boldsymbol{\beta} && (8.14)\\
\mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[\bar{Z}] &= \mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[\mathbf{x}^T\boldsymbol{\beta} + \bar{E}] = \mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[\bar{E}], && (8.15)\\
\mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[Z] &= \mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[X\boldsymbol{\beta} + E] = \mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[E] = \sigma^2(w_{11}, \ldots, w_{nn})^T, && (8.16)\\
\mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[Z] &= \mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[X\boldsymbol{\beta} + E] = \mathrm{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E] = \sigma^2 W. && (8.17)
\end{aligned}$$

Im Falle homoskedastischer Fehler kann die Varianz von $\bar{Z}$ geschätzt werden durch

$$\mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[\bar{Z}] = \mathbb{V}_{(\boldsymbol{\beta},\sigma^2)}[\bar{E}] = \sigma^2.$$

Da der wahre Wert von $\boldsymbol{\beta}$ nicht bestimmt werden kann, muss er durch einen Vektor $\hat{\boldsymbol{\beta}} \in \mathbb{R}^k$ geschätzt werden. Fassen wir die Gleichung (8.11) zeilenweise auf,

$$z_i = z(y_i) = \sum_{j=1}^{k} x_{ij}\beta_j + \epsilon_i,$$

so werden durch Einsetzen der Schätzwerte $\hat{\beta}_j$ in die Gleichung nicht mehr die ϵ_i, sondern Schätzwerte davon bestimmt:

$$z_i = z(y_i) = \sum_{j=1}^{k} x_{ij}\hat{\beta}_j + \hat{\epsilon}_i.$$

Dabei wird der geschätzte Fehler $\hat{\epsilon}_i = z_i - \sum_{j=1}^{k} x_{ij}\hat{\beta}_j$ Residuum genannt. $\hat{z}_i = \sum_{j=1}^{k} x_{ij}\hat{\beta}_j$ ist ein Schätzwert für den Erwartungswert von Z_i. Wir fassen zusammen.

Definition 8.2: Klassisches und allgemeines lineares Modell

Es sei $X \in \mathbb{R}^{n,k}$ eine Designmatrix mit $\text{rg}(X) = k$, $\boldsymbol{\beta} \in \mathbb{R}^k$ und $\sigma^2 > 0$. Sei $E : \Omega \to \mathbb{R}^n$ eine n-dimensionale reelle Zufallsvariable mit
Klassisches Modell:

$$\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[E] = \mathbf{0}, \quad \text{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E] = \sigma^2 I^{n,n}.$$

Allgemeines Modell:

$$\mathbb{E}_{(\boldsymbol{\beta},\sigma^2)}[E] = \mathbf{0}, \quad \text{Cov}_{(\boldsymbol{\beta},\sigma^2)}[E] = \sigma^2 W,\ W \in \mathbb{R}^{n,n} \text{ positiv definit.}$$

Dann heißt ein statistischer Raum

$$(\mathbb{R}^n, \mathcal{B}^n, \mathbb{P}_{Z,\theta\in\Theta}), \quad \theta = (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}_+ = \Theta$$

mit

$$Z = X\boldsymbol{\beta} + E$$

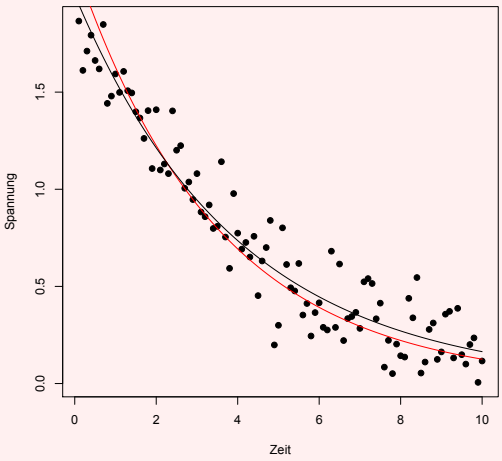
lineares Modell.

$\theta = (\boldsymbol{\beta}, \sigma^2)$ parametrisiert hierbei die Zufallsvariable Z selbst und nicht die Verteilung $\mathbb{P}_Z$ der Zufallsvariablen. Die Modellwahl besteht demnach in der Festlegung der Designmatrix. Die Modellparameter sind $\boldsymbol{\beta}$, σ^2 und W und sind nicht bekannt.

Beispiel 8.3: Exponentielle Regression

Die Entladespannung eines RC-Glieds wird durch $u(t) = u_0 e^{-\frac{t}{\tau}}$ für Konstanten u_0 und τ beschrieben. Auf Basis einer Messreihe (Zeit t, Spannung u) sollen diese Parameter bestimmt werden. Unter der Annahme, dass alle Werte für die Spannung größer Null gemessen wurden, kann durch Logarithmieren der Gleichung folgendes lineare Modell erzeugt werden:

$$\underbrace{\log(u)}_{z} = \underbrace{\log(u_0)}_{\beta_1} + \underbrace{\left(-\frac{1}{\tau}\right)}_{\beta_2} t + \epsilon.$$



Die rote Kurve zeigt die angepasste Regressionsfunktion, die schwarze Kurve die den Daten zugrunde liegende wahre Funktion.

```
plot(Zeit,Spannung,pch=19)
l=log(Spannung)
m=lm(l~Zeit)
lines(Zeit,exp(m$coefficients[1])
  *exp(m$coefficients[2]*Zeit),col=2)
lines(Zeit,u)
```

8.4. Schätzung der Modellparameter

Das Modell der multiplen Regression enthält die unbekannten Parameter $\boldsymbol{\beta}$, σ^2 und W. Da diese nicht bekannt sind, müssen sie auf Basis der gegebenen Daten geschätzt werden. Die Schätzung der Modellparameter kann auf verschiedene Weisen geschehen. Für den Vektor $\boldsymbol{\beta}$ besteht eine Möglichkeit in der Betrachtung des Vektorraums $\mathbb{R}^k$. Wird dieser mit einer p-Norm

$$||\mathbf{v}||_p := \left(\sum_{j=1}^{k} |v_j|^p \right)^{\frac{1}{p}}, \quad \mathbf{v} \in \mathbb{R}^k$$

versehen, so lässt sich eine entsprechende Metrik

$$d_p(\mathbf{v}, \mathbf{w}) := ||\mathbf{v} - \mathbf{w}||_p, \quad \mathbf{v}, \mathbf{w} \in \mathbb{R}^k.$$

induzieren. Die Betrachtung des Gleichungssystems $\mathbf{y} = X\boldsymbol{\beta}$ bei gegebenem $\mathbf{y}$ und gegebener Designmatrix X führt zu der Überlegung, $\boldsymbol{\beta}$ so bestimmen, dass der Abstand bzgl. der gewählten Metrik zu den gegebenen $\mathbf{y}$ möglichst klein wird. Diese Vorgehensweise führt zu der Methode der kleinsten Quadrate bzw. zur robusten Schätzung. Eine andere Möglichkeit besteht in der Annahme einer bestimmten Randverteilung für die Fehler E. Dann kann die Maximum-Likelihood-Methode zur Anwendung kommen.

Methode der kleinsten Quadrate

Wir betrachten den normierten Raum $(\mathbb{R}^k, ||.||_2)$ mit der 2-Norm

$$||\mathbf{v}||_2 = \left(\sum_{j=1}^{k} |v_j|^2 \right)^{\frac{1}{2}} = (\mathbf{v}^T \mathbf{v})^{\frac{1}{2}}.$$

Sei $Z = X\boldsymbol{\beta} + E$ ein allgemeines lineares Modell. Dann lässt sich folgende Schätzfunktion für $\boldsymbol{\beta}$ wählen:

$$\begin{aligned} h : \mathbb{R}^n \to \mathbb{R}^k,\ \mathbf{z} \mapsto h(\mathbf{z}) &:= \underset{\beta \in \mathbb{R}^k}{\arg\min} \{||\mathbf{z} - X\boldsymbol{\beta}||_2\} \\ &= \underset{\beta \in \mathbb{R}^k}{\arg\min} \left\{ \left((\mathbf{z} - X\boldsymbol{\beta})^T (\mathbf{z} - X\boldsymbol{\beta}) \right)^{\frac{1}{2}} \right\}. \end{aligned} \tag{8.18}$$

Dabei bezeichnet der $\arg\min$-Ausdruck den eindeutig bestimmten Vektor $\hat{\boldsymbol{\beta}} \in \mathbb{R}^k$, für den die Funktion $||\mathbf{z} - X\boldsymbol{\beta}||_2$ minimal wird. Aufgrund der Monotonie der Wurzelfunktion ist dies gleichbedeutend mit der Minimierung der Funktion $||\mathbf{z} - X\boldsymbol{\beta}||_2^2$. Diese Vorgehensweise wird Methode der kleinsten Quadrate genannt. Ein notwendiges Kriterium hierfür ist, eine Nullstelle der ersten partiellen Ableitung nach $\boldsymbol{\beta}$ zu bestimmen. Wir erhalten zunächst die so genannten Normalgleichungen

$$\begin{aligned} \frac{\partial (\mathbf{z} - X\boldsymbol{\beta})^T (\mathbf{z} - X\boldsymbol{\beta})}{\partial \beta} &= -2X^T \mathbf{y} + 2X^T X\beta \stackrel{!}{=} \mathbf{0} \\ &\Leftrightarrow 2X^T X\beta = 2X^T \mathbf{z} \\ &\Leftrightarrow X^T X\beta = X^T \mathbf{z} \end{aligned} \tag{8.19}$$

Im Falle der Lösbarkeit des Gleichungssystems in $\boldsymbol{\beta}$ ergibt sich ein Schätzwert $\hat{\boldsymbol{\beta}}$ für die Schätzfunktion h. Gilt nun $\text{rg}(X) = k$, so ist $X^T X \in \mathbb{R}^{k,k}$ invertierbar und wir können die Gleichung (8.19) eindeutig nach $\hat{\boldsymbol{\beta}}$ auflösen zu

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{z}. \tag{8.20}$$

Die Untersuchung der Hesse-Matrix $2X^T X$ der Funktion $||\mathbf{z} - X\boldsymbol{\beta}||_2^2$ zeigt wegen $2\mathbf{v}^T X^T X\mathbf{v} = 2(X\mathbf{v})^T (X\mathbf{v}) \geq 0$ für alle $\mathbf{v} \neq \mathbf{0}$, dass sie positiv-semidefinit ist. Die reelle symmetrische Matrix $X^T X$ ist diagonalisierbar. Ein Eigenwert 0 ergibt sich genau dann, wenn der Kern von $X^T X$ nicht nur aus dem Nullvektor besteht, also wenn $\text{rg}(X^T X) < k$ ist. Mit $\text{rg}(X^T X) = k$ ist $X^T X$ positiv definit, $\boldsymbol{\beta}$ ist ein globales Minimum. Wir haben damit den einzig bestmöglichen Schätzwert $\hat{\boldsymbol{\beta}}$ für das Problem (8.18) gefunden.

Beispiel 8.4

Für aus demselben Material stammende Linsen soll die Brechzahl n des Materials berechnet werden, um das verwendete Glas zu bestimmen. Für unterschiedliche Krümmungsradien der Linse r_1, r_2 der Linse sei die jeweilige Brennweite f bei einer konstanten Linsendicke von $d = 10$mm und einer Beleuchtung mit Natriumlicht gemessen worden. Es gilt der Zusammenhang

$$\frac{1}{f} = (n-1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) + \frac{(n-1)^2}{n} \frac{d}{r_1 r_2}.$$

Um eine Regressionsanalyse durchführen zu können werden drei neue Variablen eingeführt: $y = \frac{1}{f}$, $x_1 = \left(\frac{1}{r_1} - \frac{1}{r_2}\right)$ und $x_2 = \frac{d}{r_1 r_2}$. Wir erhalten für das Modell ohne Konstante die beiden Koeffizienten $\beta_1 = 0.600$ und $\beta_2 = 0.229$. Daraus lässt sich ein Wert für die Brechzahl berechnen: Aus β_1: $n = 1.600$ bzw. aus β_2: $n = 1.607$. Es handelt sich wohl um Flintglas, das eine Brechzahl von $n = 1.603$ besitzt (bei Natriumlicht).

In R nutzen wir nun einige Plots zur Modelldiagnostik.

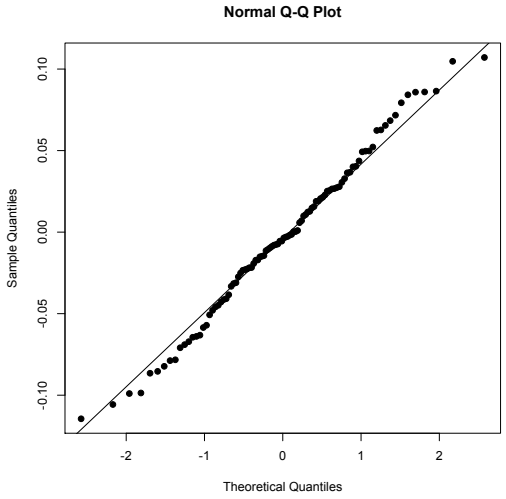
```
x1=1/r1-1/r2
x2=10/(r1*r2)
y=1/f
plot(y,x1,pch=19)
plot(y,x2,pch=19)
m=lm(y~0+x1+x2)
coefficients(m)
anova(m)
summary(m)
plot(residuals(m))
vcov(m)
qqnorm(residuals(m),pch=19)
qqline(residuals(m))
library(MASS)
sresid <- studres(m)
hist(sresid, freq=FALSE,main="")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```

Die Eingabe von „summary(m)“ führt zu folgender Ausgabe:

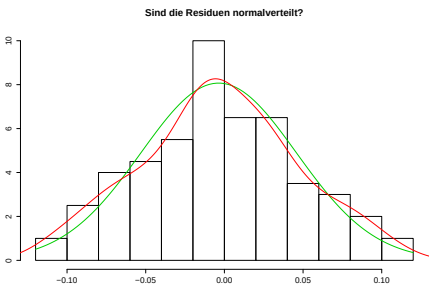
```
Call:
lm(formula = y ~ 0 + x1 + x2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.114461 -0.034526 -0.004545  0.026854  0.107109

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1 0.600499   0.013391   44.84   <2e-16 ***
x2 0.229437   0.003214   71.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01
  '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04981 on 98 degrees of freedom
Multiple R-squared: 0.9864, Adjusted R-squared: 0.9861
F-statistic:  3553 on 2 and 98 DF,  p-value: < 2.2e-16
```

Hierbei werden Tests für die einzelnen Parameter bzw. das Gesamtmodell durchgeführt. Das geschieht unter der Annahme normalverteilter Fehler im Modell mit Erwartungswert

Null und gleicher Varianz. Um das zu testen, kann ein sogenannter qq-Plot der Quantile der (Standard-)Normalverteilung gegen die Quantile der standardisierten Residuen durchgeführt werden. Dabei wird angenommen, dass die standardisierten Residuen standardnormalverteilt sind. Zu jedem einzelnen Wert kann der Wert der

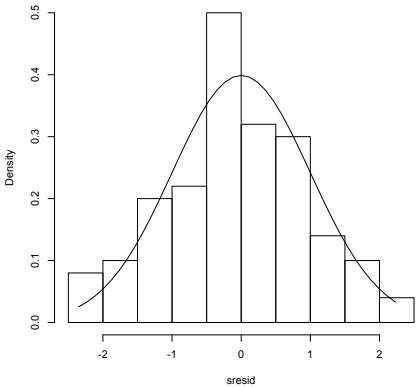


Verlaufen die Punkte im wesentlichen entlang einer Geraden, kann die Annahme akzeptiert werden, was hier der Fall ist. Auch ein Histogramm der Residuen oder studentisierten Residuen (standardisierte Residuen mit einer auf den Index bezogenen Skalierung) mit eingezeichneter Dichte der Standardnormalverteilung kann zur Überprüfung herangezogen wer-

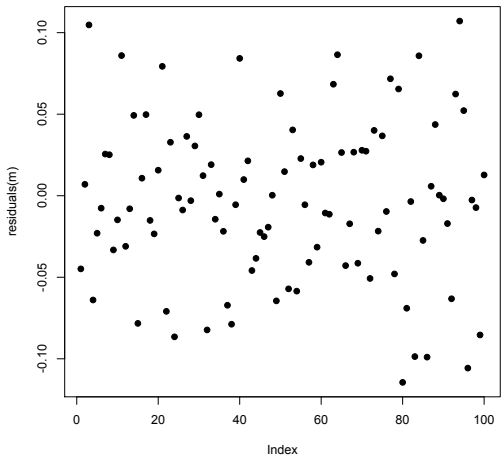


den.

6cm



Ein Plot der Residuen gegen ihren Index (vor allem bei Zeitreihen sinnvoll) oder gegen die Vorhersagewerte kann helfen aufzuzeigen, ob die Modellannahmen eines Erwartungswerts von Null und einer gleichbleibenden Varianz verletzt werden.

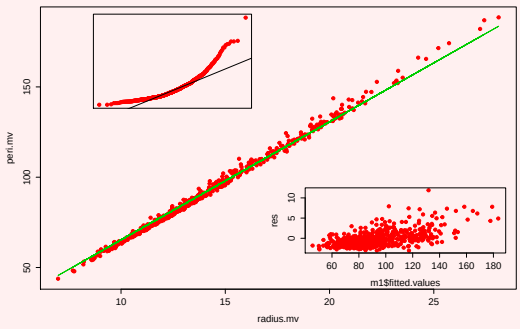


Beispiel 8.5

Zum Abschluss betrachten wir noch einmal das Brustkrebs-Beispiel. Dort werden der Radius und der Umfang von Zellen bestimmt un gemittelt. Wir überprüfen die Annahme, dass der Umfang dem eines Kreises entspricht und erhalten aus den Mittelwerten radius.mv un peri.mv

```
lm(formula = peri.mv ~ 0 + radius.mv)

Residuals:
Min       1Q   Median       3Q      Max
-3.0438 -1.7371 -0.8237  0.4853 11.8906

Coefficients:
Estimate Std. Error t value Pr(>|t|)
radius.mv 6.531686   0.005854    1116   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01
     '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.033 on 568 degrees of freedom
Multiple R-squared: 0.9995, Adjusted R-squared: 0.9995
F-statistic: 1.245e+06 on 1 and 568 DF,
     p-value: < 2.2e-16
```

Trotz des sehr hohen $\mathbb{R}$-Wertes und der signifikanten Tests kann die Modellannahme normalverteilter Fehler nicht akzeptiert werden:



Es zeigt sich, dass die Fehlervarianz nicht konstant und ein Muster im Residuenplot gegen die Vorhersagewerte erkennbar ist.

Teil IV.

Gruppierungen

9. Klassifikation

Warum?

Sollen Untersuchungsobjekte hinsichtlich einer ausgezeichneten kategorialen Eigenschaft in Gruppen eingeteilt werden, kann das geschehen, indem eine Entscheidung auf Basis gegebener Daten getroffen wird (supervised). Diese Entscheidung basiert auf einer Entscheidungsfunktion, die je nach Eingabedaten eine entsprechende Gruppenzuteilung für das jeweilige Untersuchungsobjekt ergibt. Dabei können die Eingabedaten sowohl metrisch als auch kategorial sein.

Wir betrachten in diesem Kapitel Datensätze, die ein ausgewiesenes qualitatives Merkmal C mit den Kategorien $c_1, \ldots, c_m$ besitzen, welches wir Klassenmerkmal nennen. Ziel unserer Betrachtungen ist es, aus vorhandenen Beobachtungen k verschiedener Merkmale $X_1, \ldots, X_k$ zusammen mit dem Klassenmerkmal ein Modell zu erzeugen, welches in der Lage ist bei gegebenen Werten $x_1, \ldots, x_k$ eine Kategorie $c \in C$ zu bestimmen, die mit großer Sicherheit einträte. Das Modell besteht aus einer Abbildung $\text{class}_{\mathcal{D}} : M^k \to C$, der Klassifikationsfunktion, der Wertebereich der Abbildung ist vorgegeben.

Die Daten, die zur Festlegung der Klassifikationsfunktion benutzt werden, heißt Trainingsdatensatz $\mathcal{D}$. Darin sind Beobachtungsdaten sowohl der Merkmale $X_1, \ldots, X_k$ als auch des Klassenmerkmals C enthalten. Die Klassifikationsfunktion wird anhand eines Testdatensatzes $\mathcal{T}$ überprüft und bewertet. Ein dabei häufig verwendetes Gütekriterium ist die Accuracy, die dem Quotienten von korrekt klassifizierten Fällen und allen Fällen entspricht.

9.1. Bayessche Klassifikation

Es seien $X_1, \ldots, X_k$ qualitative Merkmale. Zunächst können wir uns überlegen, dass aus wahrscheinlichkeitstheoretischer Sicht die Betrachtung der bedingten Wahrscheinlichkeit $\mathbb{P}(C = c_j | X_1 = x_1 \wedge \ldots \wedge X_k = x_k)$ für jedes $j = 1, \ldots, m$ vorzunehmen ist. Die größte Wahrscheinlichkeit dient als Entscheidungsregel. Das bedeutet nichts anderes als dass auf Basis der Trainingsdaten diejenige Assoziationsregel herangezogen wird, deren Konfidenz am größten ist.

$$\text{class}_{\mathcal{D}}(x_1, \ldots, x_k) = \text{md}((\text{conf}_{\mathcal{D}}(\{x_1, \ldots, x_k\} \to \{c_j\}))_{j=1,\ldots,m}).$$

Jedoch können wir bei einer großen Anzahl Items sämtliche Regeln und deren Konfidenzen nur mit größtem Aufwand bestimmen, so dass wir uns über Vereinfachungen Gedanken machen müssen.

Satz und Definition 9.1: Bayessche Formel

Seien $(\Omega, \mathcal{F}, \mathbb{P})$ ein beliebiger Wahrscheinlichkeitsraum und $(A_n)_{n \in \mathbb{N}}$ mit $A_i \in \mathcal{F}$ für

alle $i = 1, \ldots, n$ eine Folge paarweise disjunkter Ereignisse in $\mathcal{F}$ mit $\bigcup_{i=1}^{\infty} A_i = \Omega$ und $\mathbb{P}(A_i) > 0$, dann gilt für ein Ereignis $A \in \mathcal{F}$

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i).$$

Ist zudem $\mathbb{P}(A) > 0$, dann folgt der so genannte Satz von Bayes

$$\mathbb{P}(A_i|A) = \frac{\mathbb{P}(A_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i)}{\sum_{i=1}^{\infty} \mathbb{P}(A_i) \cdot \mathbb{P}(A|A_i)}.$$

Diese Wahrscheinlichkeit wird auch als a-posteriori Wahrscheinlichkeit von A_i unter der Bedingung A bezeichnet.

Die bedingte Wahrscheinlichkeit, auf der die Konfidenz basiert, lässt sich zunächst mit Hilfe der Bayesschen Formel anders darstellen.

$$\begin{aligned} \mathbb{P}(C = c_j \quad | \quad & X_1 = x_1 \wedge \ldots \wedge X_k = x_k) \\ = \quad & \frac{\mathbb{P}(X_1 = x_1 \wedge \ldots \wedge X_k = x_k | C = c_j) \cdot \mathbb{P}(C = c_j)}{\mathbb{P}(X_1 = x_1 \wedge \ldots \wedge X_k = x_k)}. \end{aligned}$$

In der Situation der konkreten Realisierungen ist die im Nenner auftretende Wahrscheinlichkeit zur Entscheidung über die Klassifikation jeweils gleich, wir brauchen sie nicht zu berücksichtigen. Somit bleibt

$$\mathbb{P}(X_1 = x_1 \wedge \ldots \wedge X_k = x_k | C = c_j) \cdot \mathbb{P}(C = c_j)$$

zu bestimmen. Wir schreiben den ersten Ausdruck etwas um.

$$\begin{aligned} \mathbb{P}(X_1 = x_1 \wedge \ldots \wedge X_k = x_k \quad | \quad & C = c_j) \\ = \quad & \frac{\mathbb{P}(X_1 = x_1 \wedge \ldots \wedge X_k = x_k \wedge C = c_j)}{\mathbb{P}(C = c_j)} \\ = \quad & \frac{\mathbb{P}(X_1 = x_1 | X_2 = x_2 \wedge \ldots \wedge X_k = x_k \wedge C = c_j)}{\mathbb{P}(C = c_j)} \\ & \cdot \mathbb{P}(X_2 = x_2 \wedge \ldots \wedge X_k = x_k \wedge C = c_j) \\ & \ldots \\ = \quad & \frac{\mathbb{P}(X_1 = x_1 | X_2 = x_2 \wedge \ldots \wedge X_k = x_k \wedge C = c_j)}{\mathbb{P}(C = c_j)} \\ & \cdot \ldots \cdot \mathbb{P}(X_k = x_k | C = c_j) \cdot \mathbb{P}(C = c_j) \\ = \quad & \mathbb{P}(X_1 = x_1 | X_2 = x_2 \wedge \ldots \wedge X_k = x_k \wedge C = c_j) \\ & \cdot \ldots \cdot \mathbb{P}(X_k = x_k | C = c_j) \end{aligned}$$

Satz und Definition 9.2: Bedingte Unabhängigkeit von Zufallsvariablen

Seien $A : \Omega \to M_1$, $B : \Omega \to M_2$, $C : \Omega \to M_3$ Zufallsvaribien. Dann heißt A bedingt unabhängig von B gegeben C, wenn für alle $a \in \text{dom}(M_1), b_1, b_2 \in \text{dom}(M_2)$ und

$c \in \mathrm{dom}(C)$ gilt

$$\mathbb{P}(A = a|B = b_1 \cap C = c) = \mathbb{P}(A = a|B = b_2 \cap C = c) = \mathbb{P}(A = a|C = c).$$

Nun nehmen wir an, dass X_1 bedingt unabhängig von $X_2, \ldots, X_k$ gegeben C ist, X_2 bedingt unabhängig von $X_3, \ldots, X_k$ gegeben C usw., dann erhalten wir

$$\mathbb{P}(X_1 = x_1 \wedge \ldots \wedge X_k = x_k|C = c_j) = \prod_{i=1}^{k} \mathbb{P}(X_i = x_i|C = c_j)$$

Die Aufgabe der Klassifikation reduziert sich auf die Betrachtung von

$$\prod_{i=1}^{k} \mathbb{P}(X_i = x_i|C = c_j) \cdot \mathbb{P}(C = c_j). \tag{9.1}$$

Bei einer konkreten Beobachtung klassifizieren wir demnach durch Berechnung von

$$\mathrm{class}_{\mathcal{D}}(x_1, \ldots, x_k) = \mathrm{md}((\mathrm{supp}_{\mathcal{D}}(\{c_j\}) \cdot \prod_{i=1}^{k} \mathrm{conf}_{\mathcal{D}}(\{c_j\} \to \{x_i\}))_{j=1,\ldots,m}).$$

Dies nennen wir Bayessche Klassifikation.

In Titanic-Beispiel betrachten wir die Zuordnung zum Klassenmerkmal Survived auf Basis aller Daten. Die Kombination First, Adult und Female liefert folgende Situation:

$$\mathrm{conf}_{\mathcal{D}}(\{\text{First, Adult, Female}\} \to \{\text{Yes}\}) = \frac{140}{144} = 0.972,$$
$$\mathrm{conf}_{\mathcal{D}}(\{\text{First, Adult, Female}\} \to \{\text{No}\}) = \frac{4}{144} = 0.028.$$

Somit folgte $\mathrm{class}_{\mathcal{D}}(\{\text{First, Adult, Female}\} = \text{Yes}$. Die Bayessche Klassifikation liefert

$$\begin{aligned} &\mathrm{supp}_{\mathcal{D}}(\{\text{Yes}\}) \cdot \mathrm{conf}_{\mathcal{D}}(\{\text{Yes}\} \to \{\text{First}\}) \\ &\quad \cdot \mathrm{conf}_{\mathcal{D}}(\{\text{Yes}\} \to \{\text{Adult}\}) \\ &\quad \cdot \mathrm{conf}_{\mathcal{D}}(\{\text{Yes}\} \to \{\text{Female}\}) \quad = \frac{711}{2201} \cdot \frac{203}{711} \cdot \frac{654}{711} \cdot \frac{344}{711} = 0.041, \\ &\mathrm{supp}_{\mathcal{D}}(\{\text{No}\}) \cdot \mathrm{conf}_{\mathcal{D}}(\{\text{No}\} \to \{\text{First}\}) \\ &\quad \cdot \mathrm{conf}_{\mathcal{D}}(\{\text{No}\} \to \{\text{Adult}\}) \\ &\quad \cdot \mathrm{conf}_{\mathcal{D}}(\{\text{No}\} \to \{\text{Female}\}) \quad = \frac{1490}{2201} \cdot \frac{122}{1490} \cdot \frac{1438}{1490} \cdot \frac{126}{1490} = 0.004, \end{aligned}$$

und damit die gleiche Entscheidung.

Ein weiterer Vorteil der Bayesschen Klassifikation liegt darin, dass es manchmal Werte--Kombinationen unter den Beobachtungen nicht gibt und damit keine Konfidenz der entsprechenden Assoziationsregel bestimmbar ist. Das passiert bei den Titanic-Daten etwa bei {Crew, Child, Female}. Hiervon sind die Itemsets mit zwei Elementen, die bei der Bayesschen Klassifikation auftreten, weniger betroffen. Bei einer Beobachtung {Crew, Child, Female} würden wir uns für Yes entscheiden (Werte: Yes: 0.004, No: 0.001).

Allerdings kann auch der Fall eintreten, dass Itemsets mit zwei Elementen nicht beobachtet wurden und somit. Die führt zu einem Faktor Null in der Bildung des Produktes (9.1) und alle anderen Faktoren spielen keine Rolle mehr. Hier kann dann eine Korrektur vorgenommen werden, die sich folgendermaßen motivieren lässt. Aufgrund von Beobachtungen kann die Wahrscheinlichkeit $\mathbb{P}(A|B)$ eines Ereignisses A unter dem Ereignis B über den Support $\frac{n_{AB}}{n_{\cdot B}}$ geschätzt werden, wenn n_{AB} die Anzahl der Fälle, bei denen A und B auftreten und $n_{\cdot B}$ die Anzahl Fälle, bei denen B auftritt, darstellt. Dies kann als Vorwissen (apriori-Wissen) für weitere Beobachtungen angesehen werden. Ohne solches Vorwissen könnten wir die Wahrscheinlichkeit auf Basis einer gleichförmigen Verteilung mittels $\frac{1}{|A|}$ schätzen. Nun lassen sich beide Schätzungen durch eine Kovexkombination mit einem Parameter $\mu \in [0, 1]$,

$$\mu \cdot \frac{n_{AB}}{n_{\cdot B}} + (1 - \mu) \cdot \frac{1}{|A|},$$

kombinieren. Sei nun $\mu := \frac{n_{\cdot B}}{n_{\cdot B} + |A| \cdot \lambda}$ für ein $\lambda \in [0, \infty[$, so erhalten wir durch Einsetzen die Schätzung

$$\frac{n_{AB} + \lambda}{n_{\cdot B} + |A| \cdot \lambda},$$

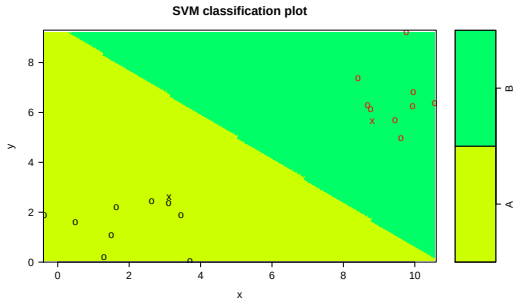
welche als Lidstone's law of succession bezeichnet wird. Ist $\lambda > 0$, so verschwindet keine Wahrscheinlichkeit $\mathbb{P}(A|B)$.

9.2. Support Vector Machines

Sei $\mathcal{D} := \{(\mathbf{x}_{1\cdot}, c_1), \ldots, (\mathbf{x}_{n\cdot}, c_n)\}$ ein Trainingsdatensatz mit einem Klassenmerkmal $C = \{-1, 1\}$ und quantitativen Merkmalen $X_1, \ldots, X_k$, $\mathbf{x}_{i\cdot} \in X_1 \times \ldots \times X_k$, $X_j = \mathbb{R}$ für alle $j = 1, \ldots, k$ und $c_i \in C$ für $i = 1, \ldots, n$. Sei $f_{\mathbf{w},b} : \mathbb{R}^k \to \mathbb{R}$ mit $\mathbf{w} \in \mathbb{R}^k$ und $b \in \mathbb{R}$, $\mathbf{x} \mapsto f_{\mathbf{w},b}(\mathbf{x}) := \mathbf{w}^T\mathbf{x} + b$, eine affin-lineare Funktion, dann suchen wir eine Funktion

$$\text{class}_{\mathcal{D}} : \mathbb{R}^k \to C, \mathbf{x} \mapsto \text{class}_{\mathcal{D}}(\mathbf{x}) := \begin{cases} 1, & f_{\mathbf{w},b}(\mathbf{x}) \geq 0 \\ -1, & f_{\mathbf{w},b}(\mathbf{x}) < 0 \end{cases}.$$

Die Modellierung führt zur Suche einer Hyperebene $H := \{\mathbf{x} \in \mathbb{R}^k;\ f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0\}$, welche die Trainingsdaten gemäß ihres Klassenmerkmals bestmöglich trennt[1]. Die folgende Abbildung zeigt ein fiktives Beispiel mit einer durch die beiden Farben gezeigten trennenden Geraden.



[1] [1]

Wir erkennen aber, dass es hier beliebig viele Möglichkeiten gibt, die beiden Klassen durch eine Gerade zu trennen.

Die Hyperebene erzeugt zwei Klassen $X^+ = \{\mathbf{x} \in \mathbb{R}^k;\ \mathbf{w}^T\mathbf{x} + b \geq 0\}$ und $X^- = \{\mathbf{x} \in \mathbb{R}^k;\ \mathbf{w}^T\mathbf{x}+b < 0\}$. Jede Beobachtung $\mathbf{x}_{i\cdot}$ kann genau einer der beiden Klassen zugeordnet werden. Nehmen wir an, dass keine der beiden Klassen leer ist. Dann gibt es je wenigstens eine Beobachtung, deren Abstand zur Hyperebene am kleinsten ist. Seien $\mathbf{x}^+ \in X^+$ und $\mathbf{x}^- \in X^-$ diese beiden Beobachtungen. Der Normalenvektor der Hyperebene ist $\mathbf{w}$, somit ist der Abstand von $\mathbf{x}^+$ zur Hyperebene über ein Vielfaches von $\mathbf{w}$, etwa $\lambda \cdot \mathbf{w}$ mit $\lambda \in \mathbb{R}$, bestimmbar. Weiter ist $\mathbf{x}^+ + \lambda \cdot \mathbf{w} \in H$, d.h.

$$\mathbf{w}^T(\mathbf{x}^+ + \lambda \cdot \mathbf{w}) + b = \mathbf{w}^T\mathbf{x}^+ + \lambda \cdot \mathbf{w}^T\mathbf{w} + b = 0.$$

Damit muss

$$\lambda = -\frac{\mathbf{w}^T\mathbf{x}^+ + b}{\mathbf{w}^T\mathbf{w}}$$

sein und es folgt für den Abstand von $\mathbf{x}^+$ zu H

$$||\mathbf{x}^+ - (\mathbf{x}^+ + \lambda \cdot \mathbf{w})|| = ||\lambda \cdot \mathbf{w}|| = |\lambda| \cdot ||\mathbf{w}|| = \frac{|\mathbf{w}^T\mathbf{x}^+ + b|}{||\mathbf{w}||}.$$

Wir suchen nun diejenigen Belegungen für $\mathbf{w}$ und b, für welche die Distanz zwischen den zu H parallelen Hyperebenen, die durch die beiden Punkte $\mathbf{x}^+$ und $\mathbf{x}^-$ gehen und so die Partitionierung erhalten, maximiert wird. Sei $H^+ := \{\mathbf{x} \in \mathbb{R}^k;\ \mathbf{w}^T\mathbf{x} + b = a_1,\ a_1 > 0\}$ diejenige Hyperebene parallel zu H mit $\mathbf{x}^+ \in H^+$ und sei entsprechend $H^- := \{\mathbf{x} \in \mathbb{R}^k;\ \mathbf{w}^T\mathbf{x}+b = a_2,\ a_2 < 0\}$ mit $\mathbf{x}^- \in H_2$. Da jede zwischen H^+ und H^- liegende und dazu parallele Hyperebene $\tilde{H}$ die Klassen trennt, nehmen wir an, dass H genau in der Mitte liegt und mit der günstigen Wahl von $\mathbf{w}$ und b damit $a_2 = -a_1$ ist (das hat eine Drehung zur Folge). Da für jedes $\mu \neq 0$ und für jedes $\mathbf{x} \in H$ dann $(\mu\mathbf{w})^T\mathbf{x} + \mu b = \mu(\mathbf{w}^T\mathbf{x} + b) = 0$ gilt, ändert eine gemeinsame Skalierung von $\mathbf{w}$ und b mit μ nichts an der Hyperebene H. Deswegen können wir mit $\tilde{b} := \frac{b}{a_1}$ und $\tilde{\mathbf{w}} := \frac{\mathbf{w}}{a_1}$ $(\mu = \frac{1}{a_1})$ schreiben:

$$\begin{aligned} H^+ &= \{\mathbf{x} \in \mathbb{R}^k;\ \tilde{\mathbf{w}}^T\mathbf{x} + \tilde{b} = 1\}, \\ H &= \{\mathbf{x} \in \mathbb{R}^k;\ \tilde{\mathbf{w}}^T\mathbf{x} + \tilde{b} = 0\}, \\ H^- &= \{\mathbf{x} \in \mathbb{R}^k;\ \tilde{\mathbf{w}}^T\mathbf{x} + \tilde{b} = -1\}. \end{aligned}$$

Der zu maximierende Abstand zwischen den beiden Hyperbenen H^+ und H^- beträgt

$$d(H^+, H^-) = \frac{|\mathbf{w}^T\mathbf{x}^- + b| + |\mathbf{w}^T\mathbf{x}^+ + b|}{||\mathbf{w}||} = \frac{|-a_1| + |a_1|}{||\mathbf{w}||} = \frac{2}{||\tilde{\mathbf{w}}||}.$$

Für $c_i = 1$ gilt nun $\tilde{\mathbf{w}}^T\mathbf{x}_i + \tilde{b} \geq 1$ und entsprechend für $c_i = -1$ dann $\tilde{\mathbf{w}}^T\mathbf{x}_i + \tilde{b} \leq -1$. Für alle $i = 1, \ldots, n$ können wir $c_i \cdot (\tilde{\mathbf{w}}^T\mathbf{x}_i + \tilde{b}) \geq 1$ schreiben.

Einer Maximierung von $\frac{2}{||\tilde{\mathbf{w}}||}$ entspricht einer Minimierung von $\frac{||\tilde{\mathbf{w}}||}{2}$. Zudem ändert sich an der Lösung nichts, wenn wir statt $||\tilde{\mathbf{w}}||$ ohne Wurzel den Ausdruck $\tilde{\mathbf{w}}^T\tilde{\mathbf{w}}$ minimieren. Zur einfacheren Notation schreiben wir wieder $\mathbf{w}$ und b anstelle von $\tilde{\mathbf{w}}$ bzw. $\tilde{b}$. Wir suchen eine Lösung von

$$\begin{aligned} &\min_{\mathbf{w} \in \mathbb{R}^k,\ b \in \mathbb{R}} \left\{ \frac{\mathbf{w}^T\mathbf{w}}{2} \right\} \\ &\text{unter } c_i \cdot (\mathbf{w}^T\mathbf{x}_i + b) \geq 1,\ i = 1, \ldots, n. \end{aligned} \tag{9.2}$$

Bevor wir uns um die Aufgabe der Lösung des Minimierungsproblems kümmern, betrachten wir zwei mögliche Probleme, die auftreten können. In der Praxis

- werden die Trainingsdaten nicht auf diese Weise trennbar sein. Dies gilt insbesondere für gestörte Daten. Wollen wir weiterhin eine trennende Hyperebene, können wir versuchen, Fehler in der Klassifikation zuzulassen, müssen diese aber in irgendeiner Form „bestrafen“,
- werden die Klassen trotz zugelassener Fehler in den Daten nicht sinnvoll durch eine Hyperebene trennbar sein. Wir sollten uns überlegen, ob es eine Möglichkeit der Bestimmung einer nichtlinearen Grenze zwischen den Klassen gibt.

Um Fehler in der Klassifikation zuzulassen, können wir für jede Beobachtung $\mathbf{x}_i$, $i = 1, \ldots, n$ so genannte Schlupfvariablen $\xi_i \geq 0$ einführen und die Nebenbedingungen des Optimierungsproblems etwas aufweichen:

$$c_i \cdot (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i.$$

Ist $\xi_i = 1$, so liegen die Daten genau auf der Grenze der Entscheidungsfunktion ($f_{\mathbf{w},b} = 0$). Für Werte von ξ_i, die größer als Eins sind, wird falsch klassifiziert. Die Bestrafung der Fehler erfolgt in der Zielfunktion, indem wir z.B. einen additiven Term $p \cdot \sum_{i=1}^{n} \xi_i$ mit einem wählbaren Gewichtungsfaktor $p > 0$ hinzufügen. Das Problem wird dadurch in der Anzahl an Parametern aufgebläht und lässt sich wie folgt formulieren:

$$\begin{array}{ll} \min\limits_{\mathbf{w}\in\mathbb{R}^k,\, b\in\mathbb{R},\, \xi_i\in\mathbb{R}_0^+} & \left\{ \frac{\mathbf{w}^T\mathbf{w}}{2} + p \cdot \sum\limits_{i=1}^{n} \xi_i \right\} \\ \text{unter} & c_i \cdot (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ i = 1, \ldots, n. \end{array} \tag{9.3}$$

Wir gehen noch einen Schritt weiter. Wie schon angedeutet, genügt es oftmals nicht, Fehler wie in der eben beschriebenen Art zuzulassen. Der Fehlerterm wird möglicherweise so groß, dass das Ergebnis keinen Sinn macht, d.h. es werden zu viele Beobachtungen falsch klassifiziert. Abhilfe kann hier die Betrachtung nichtlinearer Grenzen bei der Klassifikation schaffen. Neben dem Aufblähen des Problems durch Hinzufügen neuer Parameter können wir auch den Beobachtungsraum transformieren. Erfolgt die Transformation so, dass die transformierten Beobachtungsdaten im neuen Raum, genannt Featureraum, durch eine Hyperebene (unter Berücksichtigung einiger weniger Fehler) trennbar sind, können wir das letzte Minimierungsproblem auf die transformierten Beobachtungsdaten anwenden.

Sei $\phi : \mathbb{R}^k \to \mathbb{R}^m$ mit $m \geq k$ eine Funktion, die Beobachtungsdaten in einen Featureraum abbildet,

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ik}) \mapsto \phi(\mathbf{x}_i) = \phi(x_{i1}, \ldots, x_{ik}),\ i = 1, \ldots, n.$$

Aus dem Trainingsdatensatz $\mathcal{D}$ entsteht der Featuredatensatz

$$\mathcal{FD} = \{(\phi(\mathbf{x}_1), c_1), \ldots, (\phi(\mathbf{x}_n), c_n)\}.$$

Damit können wir das Minimierungsproblem (9.3) umformulieren zu

$$\begin{array}{ll} \min\limits_{\mathbf{w}\in\mathbb{R}^m,\, b\in\mathbb{R},\, \xi_i\in\mathbb{R}_0^+} & \left\{ \frac{\mathbf{w}^T\mathbf{w}}{2} + p \cdot \sum\limits_{i=1}^{n} \xi_i \right\} \\ \text{unter} & c_i \cdot (\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0,\ i = 1, \ldots, n. \end{array} \tag{9.4}$$

Die Zielfunktion des Minimierungsproblems (9.4) ist quadratisch und konvex, die Restriktionen linear in den Parametern. In der Formulierung (MP) können wir keine Lösung suchen, da wir die Restriktionen nicht einfach berücksichtigen können. Wir versuchen durch Formulierung des Lagrange-Dualen Problems gemäß (4.2) eine Lösung zu finden. Das zu (9.4) Lagrange-Duale Problem lautet

$$\begin{array}{ll} \max\limits_{\mathbf{u},\, \mathbf{z}\in\mathbb{R}^n} & \left\{ \Theta(\mathbf{u},\mathbf{z}) := \inf\limits_{\mathbf{w}\in\mathbb{R}^m,\, b\in\mathbb{R},\, \xi_i\in\mathbb{R}_0^+} \right. \\ & \left. \left\{ \frac{\mathbf{w}^T\mathbf{w}}{2} + p\cdot\sum\limits_{i=1}^{n}\xi_i + \sum\limits_{i=1}^{n} u_i\cdot(1-\xi_i - c_i(\mathbf{w}^T\phi(\mathbf{x_i})+b)) - \sum\limits_{i=1}^{n} z_i\xi_i \right\} \right\} \\ \text{unter} & u_i, z_i \geq 0 \quad \forall i = 1,\ldots,n. \end{array} \tag{9.5}$$

mit $\Theta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Die Zielfunktion und die Restriktionen in (9.5) sind differenzierbar, die Restriktionen linear in den Parametern und damit konvex. Somit können wir über die KKT-Bedingungen versuchen, einen Optimalpunkt zu finden. Hierzu bestimmen wir

$$\begin{aligned} \nabla f(\mathbf{w},b,\xi)^T &= (w_1,\ldots,w_m,0,p,\ldots,p), \\ \nabla g_i(\mathbf{w},b,\xi)^T &= (-c_i\phi(\mathbf{x}_i)_1,\ldots,-c_i\phi(\mathbf{x}_i)_m,-c_i,0,\ldots,0,\underbrace{-1}_{i\text{-te Stelle}},0,\ldots,0), \\ & \quad i=1,\ldots,n, \\ \nabla g_i(\mathbf{w},b,\xi)^T &= (0,\ldots,0,\underbrace{-1}_{i\text{-te Stelle}},0,\ldots,0), \\ & \quad i=n+1,\ldots,2n. \end{aligned}$$

Die erste Gleichung (ein Gleichungssystem) der KKT-Bedingungen lautet

$$\begin{pmatrix} w_1 \\ \vdots \\ w_m \\ 0 \\ p \\ \\ \\ \vdots \\ \\ \\ p \end{pmatrix} + \sum_{i=1}^{n} u_i \begin{pmatrix} -c_i\phi(\mathbf{x}_i)_1 \\ \vdots \\ -c_i\phi(\mathbf{x}_i)_m \\ -c_i \\ 0 \\ \vdots \\ 0 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \sum_{i=1}^{n} z_i \begin{pmatrix} 0 \\ \\ \\ \vdots \\ \\ \\ 0 \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0},$$

Ein Optimalpunkt $(\mathbf{w},b,\xi)$ erfüllt das Gleichungssystem. Deswegen ersetzen wir in der Zielfunktion wo es möglich ist, Terme durch Bedingungen des Gleichungssystems. Das

Ziel ist, $\mathbf{w}, b$ und ξ zu eliminieren:

$$\begin{aligned}
& \frac{\mathbf{w}^T\mathbf{w}}{2} + p \cdot \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} u_i \cdot (1 - \xi_i - c_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b)) - \sum_{i=1}^{n} z_i\xi_i \\
= \quad & \frac{1}{2}\left(\sum_{i=1}^{n} u_i c_i \phi(\mathbf{x}_i)\right)^T \left(\sum_{j=1}^{n} u_j c_j \phi(\mathbf{x}_j)\right) + (u_i + z_i)\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n} u_i - \sum_{i=1}^{n} u_i\xi_i \\
& - \sum_{i=1}^{n} u_i c_i \left(\sum_{j=1}^{n} u_j c_j \phi(\mathbf{x}_j)\right)^T \phi(\mathbf{x}_i) - \sum_{i=1}^{n} u_i c_i b - \sum_{i=1}^{n} z_i\xi_i \\
= \quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i u_j c_i c_j \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) + \sum_{i=1}^{n} u_i
\end{aligned} \tag{9.6}$$

Die Koeffizienten z_i treten in der verbleibenden Zielfunktion nicht mehr auf. Wegen $p = u_i + z_i$ und $u_i, z_i \geq 0$ muss $0 \leq u_i \leq p$ für alle u_i erfüllt sein. Wir notieren noch für $i = 1, \ldots, n$ die weiteren KKT-Bedingungen.

$$\begin{aligned}
\sum_{i=1}^{n} \left(u_i \cdot (1 - \xi_i - c_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b)) - z_i\xi_i\right) &= 0 \\
1 - \xi_i - c_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) &\leq 0 \\
-\xi_i &\leq 0 \\
u_i, z_i &\geq 0
\end{aligned}$$

Wegen der Bedingungen $\mathbf{g}(\mathbf{x}) \leq 0$ und $\mathbf{u} \geq 0$ folgt für $\mathbf{u} = (u_1, \ldots, u_n, z_1, \ldots, z_n)^T$ unmittelbar, dass $\mathbf{u}^T\mathbf{g}(\mathbf{x}) = 0$ genau dann erfüllt ist, wenn $u_i g_i(\mathbf{x}) = 0$ bzw. $z_i g_{i+n}(\mathbf{x}) = 0$ ist. Wir erhalten folgende Aufgabenstellung:

$$\begin{aligned}
\max_{\mathbf{u}\in\mathbb{R}^n} \quad & \left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_i u_j c_i c_j \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) + \sum_{i=1}^{n} u_i\right\} \\
\text{unter} \quad & 0 \leq u_i \leq p, \\
& \sum_{i=1}^{n} u_i c_i = 0.
\end{aligned} \tag{9.7}$$

Zu maximieren ist eine quadratische Funktion unter linearen Ungleichungsnebenbedingungen. Dies kann auf Basis numerischer Verfahren erfolgen. Nach der Bestimmung der u_i lässt sich $\mathbf{w}$ durch

$$\mathbf{w} = \sum_{i=1}^{n} u_i c_i \phi(\mathbf{x}_i)$$

berechnen. Die Entscheidungsfunktion f ergibt sich zu

$$\begin{aligned}
f_{\mathbf{w},b}(\mathbf{x}) &= \mathbf{w}^T\phi(\mathbf{x}) + b \\
&= \sum_{i=1}^{n} \left(u_i c_i \phi(\mathbf{x}_i)\right)^T \phi(\mathbf{x}) + b \\
&= \sum_{i=1}^{n} \left(u_i c_i \phi(\mathbf{x}_i)^T\phi(\mathbf{x})\right) + b
\end{aligned}$$

Die Entscheidung kann ohne $\mathbf{w}$ oder b bestimmen zu müssen erfolgen. Wir unterscheiden für $p > 0$ drei Fälle:

$$\begin{array}{rcll} 0 < u_i < p & \Rightarrow & p > z_i > 0 & \Rightarrow \xi_i = 0 \quad \Rightarrow c_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) = 1, \\ u_i = 0 & \Rightarrow & z_i = p & \Rightarrow \xi_i = 0 \quad \Rightarrow c_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1, \\ u_i = p & \Rightarrow & z_i = 0 & \Rightarrow \xi_i \geq 0 \quad \Rightarrow c_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) = 1 - \xi_i \leq 1. \end{array}$$

Die $\mathbf{x}_i$, für die $u_i \neq 0$ ist, werden Support-Vektoren genannt, da nur sie Einfluss in der Entscheidungsfunktion haben. Die dazugehörigen Daten liegen entweder auf dem Rand der abgrenzenden Hyperebenen oder es handelt sich um Datenpunkte, die sich im inneren Bereich befinden und entweder falsch $(\xi_i > 1)$ oder richtig $(\xi_i \leq 1)$ klassifiziert werden können.

Den Parameter b können wir berechnen, indem wir für die Support-Vektoren

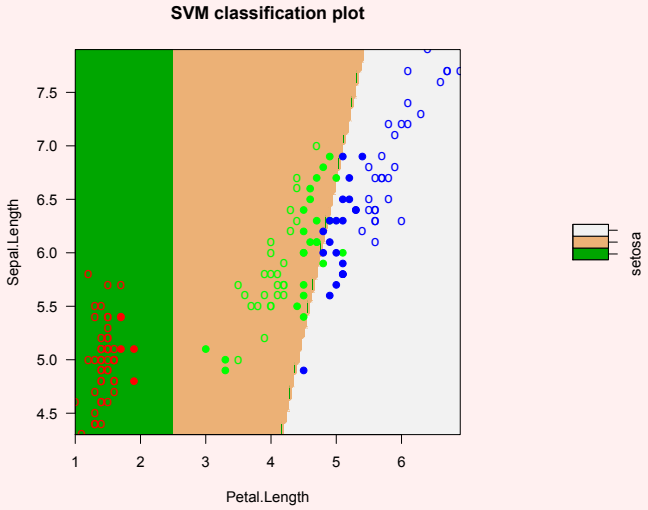
$$c_i \left(\left(\sum_{j=1}^{n} u_j c_j \phi(\mathbf{x}_j) \right)^T \phi(\mathbf{x}_i) + b \right) = 1$$

$$\overset{c_i^2=1}{\Leftrightarrow} \quad b = c_i - \left(\sum_{j=1}^{n} u_j c_j \phi(\mathbf{x}_j) \right)^T \phi(\mathbf{x}_i) \tag{9.8}$$

benutzen. Da es für jeden Support-Vektor zu verschiedenen Werten von b kommen kann, kann über eine Mittelung oder Medianbildung ein robusterer Wert für b festgelegt werden.

Beispiel 9.3

Von drei verschiedenen Irisiarten wurden Kelchblattlängen und -breiten sowie Blütenblattlängen und -breiten erfasst. Mit den jeweiligen Längen soll auf Basis der Irisart eine (lineare) Einteilung in drei Klassen erfolgen.

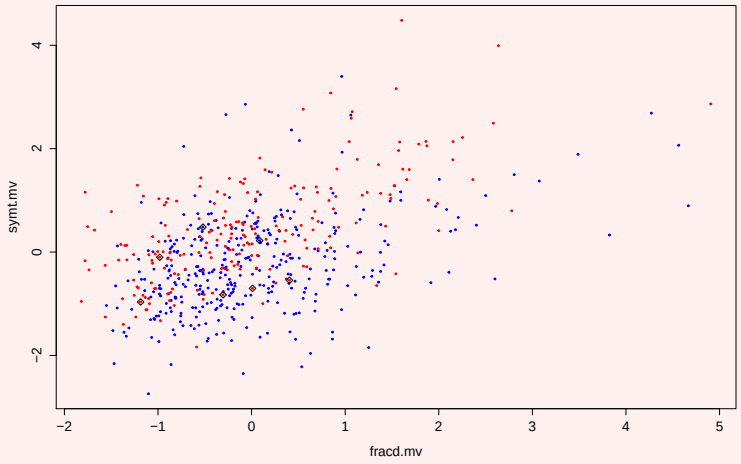


Wir erhalten 49 Support-Vektoren (4,24,21). Es werden vier Pflanzen falsch eingeteilt.

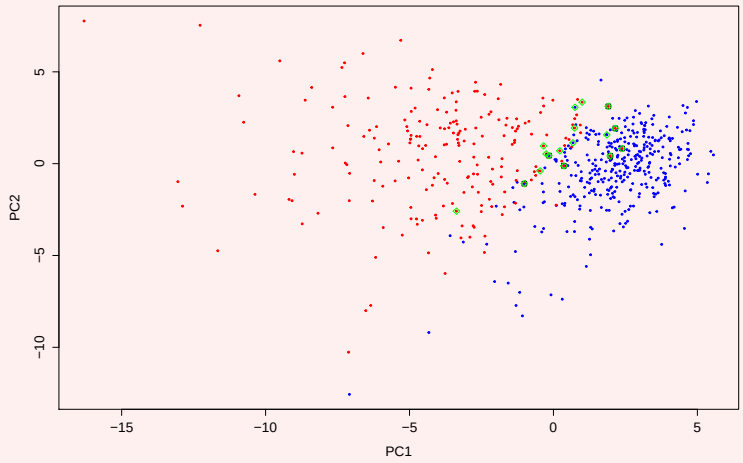
```
library(e1071)
q=as.data.frame(cbind(Petal.Length,Sepal.Length))
mod=svm(Species~.,q,kernel="linear")
print(mod)
summary(mod)
plot(mod,q,Sepal.Length~Petal.Length,svSymbol=19,
   color.palette=terrain.colors,symbolPalette=rainbow(3),
   grid=150,main="")
table(predict(mod), true=Species)
```

Beispiel 9.4

Bei den untersuchten Patienten im Brustkrebsdatensatz wurden bösartige oder nur gutartige Zellen gefunden und die Patienten entsprechend eingeteilt (Merkmal diagnose, Merkmalswerte M bzw. B). Damit kann eine lineare SVM anhand der übrigen Merkmale durchgeführt werden. Dabei zeigt sich, dass für den kompletten Datensatz (569 Patienten) nur 7 falsch eingeteilt werden (2 B$\rightarrow$ M und 5 M $\rightarrow$;B). Die falsch eingeordneten Patienten sind in folgendem Scatterplot gekennzeichnet.



Nehmen wir als Grundlage die ersten sechs Hauptkomponenten der standardisierten Daten aus Beispiel 5.12 und führen eine lineare SVM durch, so erhalten wir nicht dasselbe Ergebnis, jedoch sind insgesamt nur 17 der 569 Patienten falsch klassifiziert. In nachfolgendem Scatterplot der ersten beiden Hauptkomponenten sind die falsch klassifizierten Patienten nach dem ersten Modell mit denjenigen aus dem zweiten verglichen.



Kernel-SVM

Eine Transformation vom $\mathbb{R}^k$ in einen höher-dimensionalen Raum kann zu Schwierigkeiten führen. Dazu stellen wir uns vor, wir haben im Intervall $[0, 1]$ 100 Realisierungen einer gleichverteilten Zufallszahl. Das Intervall wird dadurch gut abgedeckt. Das Einheitsquadrat jedoch wird von 100 Punkten nicht besonders gut abgedeckt. Es werden bereits 100^2 Realisierungen benötigt.

Müssen wir nun jedes $\mathbf{x}_i$ auf $\phi(\mathbf{x}_i)$ abbilden und dann ein Skalarprodukt $\phi^T\phi$ bilden? Unter gewissen Voraussetzungen können wir uns das ersparen. Dies liegt an einer praktischen Wahl einer so genannten Kernelfunktion

$$k : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}, (\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{y}) := \phi(\mathbf{x})^T\phi(\mathbf{y}), \tag{9.9}$$

da in den Berechnungen nichts anderes als ein Skalarprodukt im entsprechenden Featureraum benötigt wird. Nehmen wir z.B. die Abbildung

$$\phi : \mathbb{R}^2 \to \mathbb{R}^3,\ (x, y) \mapsto \phi(x, y) := (x^2, y^2, \sqrt{2}xy)^T.$$

Dann gilt

$$\begin{aligned} \phi(x, y)^T\phi(z, w) &= (x^2, y^2, \sqrt{2}xy)(z^2, w^2, \sqrt{2}zw)^T = x^2z^2 + y^2w^2 + 2xyzw \\ &= (xz + yw)^2. \end{aligned}$$

Wir setzen $k((x, y), (z, w)) := (xz + yw)^2$ und das Skalarprodukt lässt sich deutlich einfacher berechnen.

Durchführung und Bewertung einer Klassifikation

Die Daten sind in ein Trainings- und ein Testdaten eingeteilt. Ein Modell, das eine hohe Genauigkeit (Accuracy) bei den Trainingsdaten, jedoch im Gegensatz eine geringere Genauigkeit bei den Testdaten besitzt, wird als overfitted bezeichnet. Für die Erzeugung von Trainingsdaten können verschiedene Methoden benutzt werden. Eine zufällige Auswahl

oder zeitlich bedingte Zusammensetzung sind möglich, weiter kann auch die Kreuzvalidierung als Methode benutzt werden.

10. Clusteranalyse

Warum?

Manchmal sollen Untersuchungsobjekte gruppiert werden, ohne jedoch vorab die einzelnen Gruppen oder deren Anzahl zu kennen (unsupervised). Auf Basis der erhobenen Daten kann versucht werden, eine Gruppierung zu finden. Dabei wird versucht, die Objekte, die sich möglichst ähnlich sind, in einer Gruppe zusammenzufassen, unähnliche Objekte in verschiedenen Gruppen.

Ziel einer Clusteranalyse ist das Auffinden einer empirischen Klassifikation oder einer hierarchischen Ähnlichkeitsstruktur. Die Clusteranalyse kann Merkmalsträger-orientiert oder merkmalsorientiert durchgeführt werden. Eine Klassifikation von Merkmalen kann dahingehend interpretiert werden, dass es wie bei der Faktorenanalyse latente Merkmale gibt, die nicht beobachtbar sind. Die am Merkmalsträger orientierte Clusteranalyse gruppiert Merkmalsträger anhand der Realisierungen der Merkmale.
Sei S eine endliche Menge von Objekten. Eine Abbildung

$$c : \mathcal{P}(S) \to \{0, 1\}, \tag{10.1}$$

für die gilt

- $c(\emptyset) = 0$,
- $\forall \omega \in S : \exists M \in \mathcal{P}(S) : \omega \in M \Rightarrow c(M) = 1$
  jedes Objekt ist wenigstens einer Klasse zugeordnet,
- verschiedene Objekte in disjunkten Elementen von $c^{-1}(\{1\})$ sind möglichst unähnlich,
- verschiedene Objekte in einem Element von $c^{-1}(\{1\})$ sind möglichst ähnlich,

heißt Clusteranalyse. Jedes Element von $c^{-1}(\{1\})$ heißt Cluster. Ein zentrales Element der Clusteranalyse sind die Ähnlichkeit bzw. Distanz zweier Objekte. Um eine Ähnlichkeit oder Distanz bestimmen zu können, benötigen wir eine Datenmatrix, anhand welcher die Objekte unterschieden werden können. Sei $g := |\mathcal{S}|,\ \mathcal{S} \subseteq \mathcal{P}_*(S)$. Gilt für

$$\bigcup_{i=1}^{g} M_i = S,\ M_i \cap M_j = \emptyset,\ i \neq j, \tag{10.2}$$

so heißt $\mathcal{S} = \{M_1, \ldots, M_g\}$ Partition von S. Die Elemente $M \in \mathcal{S}$ einer Partition heißen Klassen. Eine Clusteranalyse, die als Ergebnis eine Partition liefert, heißt partitionierendes Verfahren, andernfalls nicht-partitionierendes Verfahren.

10.1. Hierarchische Clusteranalyse

Wir betrachten eine endliche Menge S von Objekten mit $n = |S|$ und ein Mengensystem $\mathcal{S} \subset \mathcal{P}_*(S)$ (ohne die leere Menge). Soll für die Daten eine hierarchische Ähnlichkeitsstruktur gefunden werden, kommen hierarchische Methoden der Clusteranalyse zum Einsatz. Wir konzentrieren uns zunächst auf eine am Merkmalsträger orientierte Sichtweise. Die hierarchische Clusteranalyse kann in agglomerative und divisive Verfahren unterschieden werden. Bei agglomerativen Verfahren wird davon ausgegangen, dass alle Merkmalsträger zunächst ein eigenes Cluster bilden, d.h.

$$\mathcal{S} = \{\{\omega_{1\cdot}\}, \ldots, \{\omega_{n\cdot}\}\}, \quad \omega_{i\cdot} \in S \; \forall i = 1, \ldots, n. \tag{10.3}$$

Sukzessive werden dann die bestehenden Cluster zu weniger Clustern zusammengefasst. Führen wir das bis zum Ende durch, erhalten wir ein Cluster, in dem alle Objekte enthalten sind, d.h.

$$\mathcal{S} = \{\{\omega_{1\cdot}, \ldots, \omega_{n\cdot}\}\}, \quad \omega_{i\cdot} \in S \; \forall i = 1, \ldots, n. \tag{10.4}$$

Umgekehrt verlaufen divisive Verfahren. Anfangs gibt es ein Cluster, in dem alle Objekte enthalten sind. Dann wird dieses Cluster in kleinere Cluster zerlegt. Schließlich gibt es so viele Cluster wie es Objekte gibt. Betrachtet werden bei beiden Arten von Verfahren Hierarchien.

Definition 10.1: Hierarchie, Klasse

Ein Mengensystem $\mathcal{S} \subseteq \mathcal{P}_*(S)$ heißt Hierarchie von S, wenn für zwei Mengen $M_l, M_m \in \mathcal{S}$ genau eine der folgenden drei Möglichkeiten zutrifft:

$$M_l \cap M_m = \emptyset, \quad M_l \subset M_m, \quad M_m \subset M_l.$$

Jedes Element einer Hierarchie heißt Klasse.

Jede Hierarchie ist damit eine Vereinigungsmenge von Partitionen. Die Homogenität von Klassen kann über einen Index bestimmt werden.

Definition 10.2: Index

Ein Index zur Hierarchie $\mathcal{S}$ ist eine für alle Klassen $M \in \mathcal{S}$ definierte nichtnegative Funktion $h : \mathcal{S} \to \mathbb{R}$ mit $M_l \subseteq M_m \Rightarrow h(M_l) \leq h(M_m)$ und $h(M) = 0 \Leftrightarrow \omega_{i\cdot} = \omega_{j\cdot} \; \forall \omega_{i\cdot}, \omega_{j\cdot} \in M$.

Für die beiden folgenden Abschnitte betrachten wir die Datenmatrix

$$X^T = \begin{pmatrix} 1 & 5 & 2 & 6 & 4 \\ 1 & 5 & 1 & 4 & 2 \end{pmatrix}, \quad X \in \mathbb{R}^{5,2}$$

mit 5 Objekten $\omega_1, \ldots, \omega_5$ und 2 Merkmalen und mit den folgenden Distanzmatrizen:

City-Block	Euklidisch
$D_1 = \begin{pmatrix} 0 & 8 & 1 & 8 & 4 \\ & 0 & 7 & 2 & 4 \\ & & 0 & 7 & 3 \\ & & & 0 & 4 \\ & & & & 0 \end{pmatrix}$	$D_2 = \begin{pmatrix} 0.00 & 5.66 & 1.00 & 5.83 & 3.16 \\ & 0.00 & 5.00 & 1.41 & 3.16 \\ & & 0.00 & 5.00 & 2.24 \\ & & & 0.00 & 2.82 \\ & & & & 0.00 \end{pmatrix}$
Quadriert Euklidisch	**Chebychev**
$D_3 = \begin{pmatrix} 0 & 32 & 1 & 34 & 10 \\ & 0 & 25 & 2 & 10 \\ & & 0 & 25 & 5 \\ & & & 0 & 8 \\ & & & & 0 \end{pmatrix}$	$D_4 = \begin{pmatrix} 0 & 4 & 1 & 5 & 3 \\ & 0 & 4 & 1 & 2 \\ & & 0 & 4 & 2 \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix}$

Definition 10.3

Sei S eine endliche Menge und seien $\mathcal{S}_1, \mathcal{S}_2$ Partitionen. Eine Abbildung $\mathcal{S}_1 \to \mathcal{S}_2$ heißt agglomerative (divisive) Clusterbildung, wenn $\mathcal{S}_1 \cup \mathcal{S}_2$ eine Hierarchie ist und es für jedes $M_1 \in \mathcal{S}_1$ genau ein $M_2 \in \mathcal{S}_2$ gibt mit $M_1 \subseteq M_2$ ($M_2 \subseteq M_1$).

Agglomerative Verfahren

Gegeben sei eine n, k-Datenmatrix X mit n Merkmalsträgern, zusammengefasst in der Menge S und k Merkmalen und eine Distanzmatrix $D = (d_{ij}) \in \mathbb{R}^{n,n}$. Weiter sei das Mengensystem (10.3), also eine Hierarchieebene, gegeben. Agglomerative Verfahren sollen anhand des Single-Linkage-Verfahrens vorgestellt werden. In jedem Schritt werden dabei Klassen mit den kleinsten Distanzen zu einer Klasse zusammengefasst. Die Distanz zweier Klassen M_l, M_m lautet

$$d(l, m) = \min_{\omega_i \in M_l, \omega_j \in M_m} d_{ij} \tag{10.5}$$

und die minimale Distanz zweier Klassen ist

$$\min_{M_l, M_m \in \mathcal{S},\ m > l} d(l, m) \tag{10.6}$$

Beispiel 10.4

Wir führen das Verfahren auf Basis der vier Distanzmatrizen durch. Zunächst haben

wir die Hierarchie $\mathcal{S} = \{M_1, \ldots, M_5\}$ mit $M_i := \{\omega_{i.}\},\ i = 1, \ldots, 5$. Wir erhalten

Dist.funktion	Schritt	Zusammengefasste Klassen	Partition von S
City-Block	1	$M_{13} = M_1 \cup M_3$	$\{M_{13}, M_2, M_4, M_5\}$
	2	$M_{24} = M_2 \cup M_4$	$\{M_{13}, M_{24}, M_5\}$
	3	$M_{135} = M_{13} \cup M_5$	$\{M_{135}, M_{24}\}$
	4	$M_{12345} = M_{135} \cup M_{24}$	$\{M_{12345}\}$
Euklid	1	$M_{13} = M_1 \cup M_3$	$\{M_{13}, M_2, M_4, M_5\}$
	2	$M_{24} = M_2 \cup M_4$	$\{M_{13}, M_{24}, M_5\}$
	3	$M_{135} = M_{13} \cup M_5$	$\{M_{135}, M_{24}\}$
	4	$M_{12345} = M_{135} \cup M_{24}$	$\{M_{12345}\}$
Quadr. Euklid	1	$M_{13} = M_1 \cup M_3$	$\{M_{13}, M_2, M_4, M_5\}$
	2	$M_{24} = M_2 \cup M_4$	$\{M_{13}, M_{24}, M_5\}$
	3	$M_{135} = M_{13} \cup M_5$	$\{M_{135}, M_{24}\}$
	4	$M_{12345} = M_{135} \cup M_{24}$	$\{M_{12345}\}$
Chebychev	1	$M_{13} = M_1 \cup M_3, M_{24} = M_2 \cup M_4$	$\{M_{13}, M_{24}, M_5\}$
	2	$M_{12345} = M_{13} \cup M_5 \cup M_{24}$	$\{M_{12345}\}$

Die Distanzmatrizen D_i^j nach dem j-ten Schritt mit der i-ten Distanzfunktion von oben lauten

$$D_1^1 = \begin{pmatrix} 0 & 7 & 7 & 3 \\ & 0 & 2 & 4 \\ & & 0 & 4 \\ & & & 0 \end{pmatrix} \qquad D_1^2 = \begin{pmatrix} 0 & 7 & 3 \\ & 0 & 4 \\ & & 0 \end{pmatrix} \qquad D_1^3 = \begin{pmatrix} 0 & 4 \\ & 0 \end{pmatrix}$$

$$D_2^1 = \begin{pmatrix} 0.00 & 5.00 & 5.00 & 2.24 \\ & 0.00 & 1.41 & 3.16 \\ & & 0.00 & 2.82 \\ & & & 0.00 \end{pmatrix} \qquad D_2^2 = \begin{pmatrix} 0.00 & 5.00 & 2.24 \\ & 0.00 & 2.82 \\ & & 0.00 \end{pmatrix} \qquad D_2^3 = \begin{pmatrix} 0.00 & 2.82 \\ & 0.00 \end{pmatrix}$$

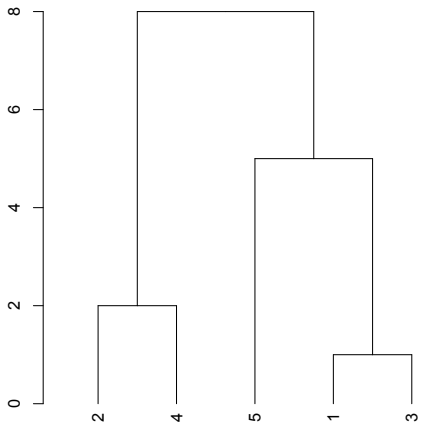
$$D_3^1 = \begin{pmatrix} 0 & 25 & 25 & 5 \\ & 0 & 2 & 10 \\ & & 0 & 8 \\ & & & 0 \end{pmatrix} \qquad D_3^2 = \begin{pmatrix} 0 & 25 & 5 \\ & 0 & 8 \\ & & 0 \end{pmatrix} \qquad D_3^3 = \begin{pmatrix} 0 & 8 \\ & 0 \end{pmatrix}$$

$$D_4^1 = \begin{pmatrix} 0 & 4 & 2 \\ & 0 & 2 \\ & & 0 \end{pmatrix}$$

Das Single Linkage Verfahren besitzt die so genannte Verkettungseigenschaft, bei der Klassen, die zwar deutlich getrennt aber durch einzelne Objekte wie bei Kettengliedern verbunden sind. Damit lassen sich Ausreißer feststellen. Es stellt sich die Frage, wie die hierarchische Struktur dargestellt und wie eine bestmögliche Klassenzahl gefunden werden kann. Da die Klassenbildung in jedem Schritt aufeinander aufbaut, bietet sich eine Baumdarstellung an, die wir nun ansehen.

Darstellung der Hierarchie

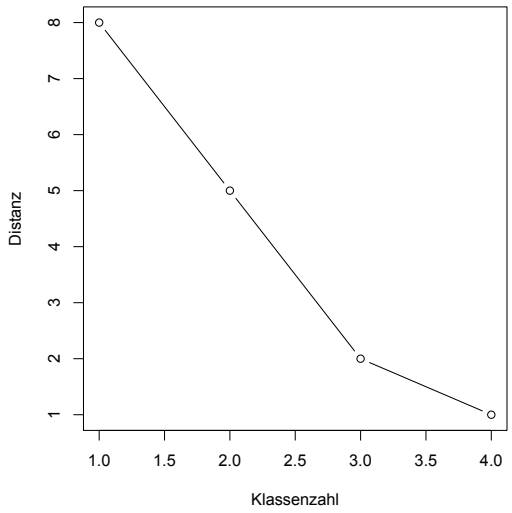
Die sich aus einem hierarchischen Verfahren ergebende Hierarchie kann in einem so genannten Dendrogramm dargestellt werden. Dabei werden die Objekte gegen die im jeweiligen Schritt zur Trennung/Vereinigung entscheidenden Distanzen angetragen und dabei die Objekte in einer Baumstruktur verbunden.




```
x=c(1,5,2,6,4)
y=c(1,5,1,4,2)
d <- dist(cbind(x,y), method = "euclidean")^2
fit <- hclust(d, method="single")
plot(as.dendrogram(fit))
```

Bestimmung der Klassenzahl

Die Bestimmung der Klassenzahl kann über das Ellbogenkriterium erfolgen. Dabei wird die Klassenzahl gegen die im jeweiligen Schritt entscheidende Distanz geplottet und die Punkte miteinander verbunden. Der subjektiv größte Knick im Diagramm entscheidet über die Klassenzahl, indem der Klassenwert vor dem Knick genommen wird.




```
plot(4:1, fit$height, type="b", xlab="Klassenzahl",
  ylab="Distanz")
```

Bei hierarchischen Clusteranalyseverfahren lässt sich das Manko feststellen, dass einmal begangene Aufteilungen bzw. Zusammenfassungen von Klassen nicht mehr rückgängig gemacht werden können. Dies erscheint dann sinnvoll, wenn wir eine Hierarchie in den Klassen suchen. Sollen Partitionierungen rückgängig gemacht werden können, empfiehlt es sich, ein klassifikatorisches Verfahren durchzuführen.

10.2. Klassifikatorische Clusteranalyse

Soll für die Daten eine empirische Klassifikation gefunden werden, kommen klassifikatorische Methoden der Clusteranalyse zum Einsatz. Eine Variante solcher Verfahren bilden die K-Means-Verfahren. Dabei werden Klassenzentren derart gebildet, dass die Streuungsquadratsummen in den Klassen minimal werden. Ausgangspunkt ist wiederum eine Datenmatrix $X \in \mathbb{R}^{n,k}$. Abkürzend notieren wir nun i anstelle von ω_i.. Es werden bei K Klassen $M_1, \ldots, M_K$ und k Merkmalen K Klassenzentren $\bar{x}_p = (\bar{x}_{p1}, \ldots, \bar{x}_{pk})^T$, mit $p = 1, \ldots, K$, $j = 1, \ldots, k$ durch

$$\bar{x}_{pj} = \frac{1}{|M_p|} \sum_{i \in M_p} x_{ij} \tag{10.7}$$

gebildet und die Streuungsquadratsumme

$$\text{SQ}_{in}(K) := \sum_{p=1}^{K} \sum_{i \in M_p} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{pj})^2 \tag{10.8}$$

innerhalb der Klassen minimiert. Die dritte Summe entspricht der quadrierten Euklidischen Distanz,

$$\sum_{j=1}^{k}(x_{ij}-\bar{x}_{pj})^2 =: d(\omega_{i\cdot},\bar{x}_p)^2$$

und deshalb schreiben wir

$$\min\{\mathrm{SQ}_{in}(K)\} \text{ mit } \mathrm{SQ}_{in}(K)=\sum_{p=1}^{K}\sum_{i\in M_p} d(\omega_{i\cdot},\bar{x}_p)^2. \tag{10.9}$$

Die konstante Gesamtstreuungsquadratsumme zwischen allen Objekten,

$$\mathrm{SQ}_{ges} := \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-\bar{x}_j)^2,$$

mit dem Merkmalsmittelwert $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ hängt mit den Streuungsquadratsummen innerhalb der Klassen und den Streuungsquadratsummen

$$\mathrm{SQ}_{zw}(K) := \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}(\bar{x}_{pj}-\bar{x}_j)^2 = \sum_{p=1}^{K}|M_p|\sum_{j=1}^{k}(\bar{x}_{pj}-\bar{x}_j)^2$$

zwischen den Klassen zusammen. Es gilt

Satz 10.5

$$\mathrm{SQ}_{ges} = \mathrm{SQ}_{zw}(K) + \mathrm{SQ}_{in}(K)$$

Beweis.

$$
\begin{aligned}
\mathrm{SQ}_{ges} &= \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_j)^2 \\
&= \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_{pj}+\bar{x}_{pj}-\bar{x}_j)^2 \\
&= \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}\left[(x_{ij}-\bar{x}_{pj})^2+2(x_{ij}-\bar{x}_{pj})(\bar{x}_{pj}-\bar{x}_j)+(\bar{x}_{pj}-\bar{x}_j)^2\right] \\
&= \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}\left[(x_{ij}-\bar{x}_{pj})^2+(\bar{x}_{pj}-\bar{x}_j)^2\right] \\
&\quad +2\sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}(x_{ij}-\bar{x}_{pj})(\bar{x}_{pj}-\bar{x}_j) \\
&= \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}\left[(x_{ij}-\bar{x}_{pj})^2+(\bar{x}_{pj}-\bar{x}_j)^2\right] \\
&\quad +2\sum_{p=1}^{K}\sum_{j=1}^{k}(\bar{x}_{pj}-\bar{x}_j)\underbrace{\sum_{i\in M_p}(x_{ij}-\bar{x}_{pj})}_{=0} \\
&= \sum_{p=1}^{K}\sum_{i\in M_p}\sum_{j=1}^{k}\left[(x_{ij}-\bar{x}_{pj})^2+(\bar{x}_{pj}-\bar{x}_j)^2\right] \\
&= \mathrm{SQ}_{in}(K)+\mathrm{SQ}_{zw}(K) \qquad (10.10)
\end{aligned}
$$

□

Da die Gesamtstreuungsquadratsumme konstant ist, könnten wir die Aufgabenstellung (10.9) auch formulieren als

$$\max\{\mathrm{SQ}_{zw}(K)\}. \qquad (10.11)$$

Ein zu klassifizierendes Objekt wird derjenigen Klasse zugeordnet, zu dessen Klassenzentrum der geringste Euklidische Abstand besteht, da damit der kleinstmögliche Streuungsbeitrag erzeugt wird. Sind somit die Klassenzahl K und die Klassenzentren vorgegeben, kann jedes Objekt durch Bestimmung von

$$\omega_{i\cdot} \in M_p \Leftrightarrow p = \min_{l\in\{1,\ldots,K\}}\{d(\omega_{i\cdot},\bar{x}_l)^2\}$$

einer Klasse zugewiesen werden. Bei mehreren minimalen Distanzen kann das Objekt einer beliebigen Klasse zugeordnet werden, da die Zuordnung nicht fest bleiben muss. Durch die Zuordnung aller Objekte zu genau einer Klasse können dann durch Gleichung (10.7) neue Klassenzentren bestimmt und eine neue Zuordnung durchgeführt werden. Diese Schritte können solange iteriert werden, bis es keine Änderung der Zuordnung der Objekte mehr gibt.

Beispiel 10.6

Gegeben sei wieder die Datenmatrix $X^T = \begin{pmatrix} 1 & 5 & 2 & 6 & 4 \\ 1 & 5 & 1 & 4 & 2 \end{pmatrix}$. Es sollen aufgrund der Betrachtung des Scatterplots zwischen X_1 und X_2 zwei Klassen mit dem K-Means-Verfahren erzeugt werden. Wir starten mit den beliebig gewählten Merkmalsträgern $\bar{x}_1 := \omega_{1\cdot} = (1,1)$ und $\bar{x}_2 := \omega_{4\cdot} = (6,4)$ als initiale Klassenzentren und erhalten folgende Distanzen der Merkmalsträger von den Klassenzentren:

Merkmalsträger	$d(\omega_{i\cdot}, \bar{x}_1)^2$	$d(\omega_{i\cdot}, \bar{x}_2)^2$	Zuordnung
$\omega_{1\cdot}$	0	34	Klasse 1
$\omega_{2\cdot}$	32	2	Klasse 2
$\omega_{3\cdot}$	1	25	Klasse 1
$\omega_{4\cdot}$	34	0	Klasse 2
$\omega_{5\cdot}$	10	8	Klasse 2

Als neue Klassenzentren ergeben sich

$$\bar{x}_1 = \left(\frac{1}{2}(1+2), \frac{1}{2}(1+1)\right) = (1.50, 1.00) \text{ und}$$
$$\bar{x}_2 = \left(\frac{1}{3}(5+6+4), \frac{1}{3}(5+4+2)\right) = (5.00, 3.67).$$

Die neue Zuordnung ergibt sich wie folgt:

Merkmalsträger	$d(\omega_{i\cdot}, \bar{x}_1)^2$	$d(\omega_{i\cdot}, \bar{x}_2)^2$	Zuordnung
$\omega_{1\cdot}$	0.25	23.11	Klasse 1
$\omega_{2\cdot}$	28.25	1.78	Klasse 2
$\omega_{3\cdot}$	0.25	16.11	Klasse 1
$\omega_{4\cdot}$	21.25	0.11	Klasse 2
$\omega_{5\cdot}$	7.25	3.78	Klasse 2

Da sich keine Änderung gegenüber dem ersten Durchlauf ergeben hat, wird die Iteration beendet. Werden nicht nur zwei, sondern auch eine, drei, vier oder fünf Klassen gebildet, ergeben sich folgende Zuordnungen:

K	Zentren	$\omega_{1\cdot}$	$\omega_{2\cdot}$	$\omega_{3\cdot}$	$\omega_{4\cdot}$	$\omega_{5\cdot}$
1	$(4.5, 3.25)$	Kl. 1	Kl. 1	Kl. 1	Kl. 1	Kl. 1
2	$(1.5,1), (5, 3.67)$	Kl. 1	Kl. 2	Kl. 1	Kl. 2	Kl. 2
3	$(1.5,1), (5.5, 4.50), (4,2)$	Kl. 1	Kl. 2	Kl. 1	Kl. 2	Kl. 3
4	$(1.5,1), (5,5), (4,2), (6,4)$	Kl. 1	Kl. 2	Kl. 1	Kl. 4	Kl. 3
5	$(1,1), (5,5), (2,1), (4,2), (6,4)$	Kl. 1	Kl. 2	Kl. 3	Kl. 4	Kl. 5

Es stellt sich die Frage, welche Klassenzahl genommen wird. Ist das K-Means-Verfahren für mehrere Klassenzahlen durchgeführt worden, kann die erkärte Streuung $SQ_{in}(K)$ als

Grundlage für eine Prüfgröße des Modells verwendet werden. Dabei geben

$$V_K := 1 - \frac{\text{SQ}_{in}(K)}{\text{SQ}_{ges}} \text{ und } V_K^{K-1} := 1 - \frac{\text{SQ}_{in}(K)}{\text{SQ}_{in}(K-1)} \tag{10.12}$$

die Verbesserung des Modells mit K Klassen gegenüber dem Nullmodell (mit $\text{SQ}_{ges} = \text{SQ}_{in}(1)$) und die Verbesserung des Modells mit K Klassen gegenüber dem Vormodell mit $K-1$ Klassen an.

Beispiel 10.7

Das letzte Beispiel liefert folgende Prüfgrößen:

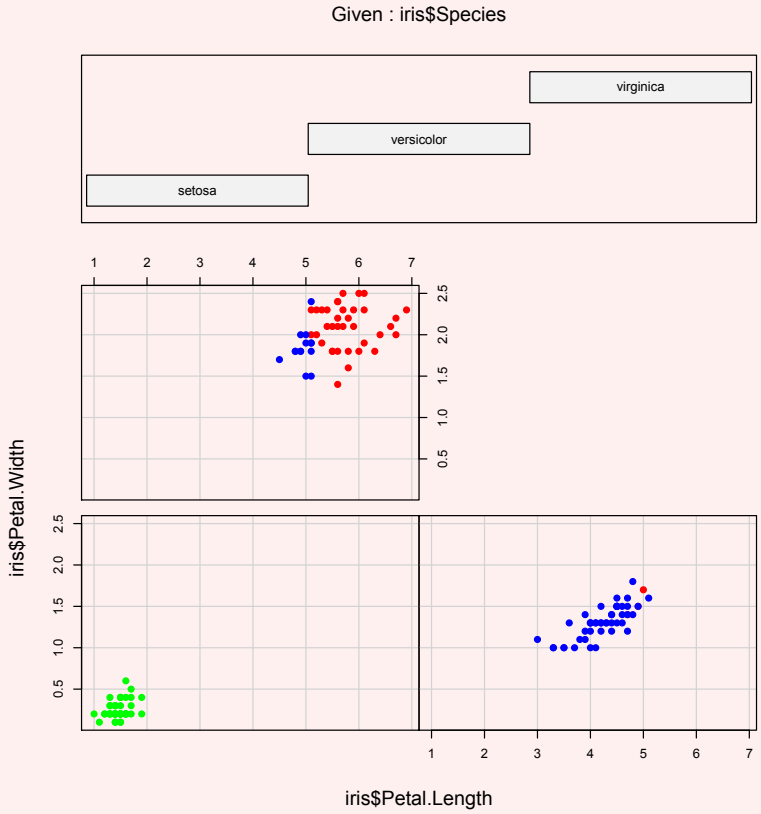
K	$\text{SQ}_{in}(K)$	V_K	V_K^{K-1}
1	36.56	0	–
2	7.17	0.80	0.80
3	1.5	0.96	0.79
4	0.5	0.99	0.67
5	0	1	1

Bei den wenigen Daten ist keine klare Aussage möglich. Gehen wir nach der größten Verbesserung, behalten wir das 2-Klassen-Modell bei.

```
x=c(1,5,2,6,4)
y=c(1,5,1,4,2)
fit <- kmeans(cbind(x,y),2,nstart=1,algorithm="Lloyd")
aggregate(cbind(x,y),by=list(fit$cluster),FUN=mean)
plot(x,y,col=c("red","blue")[fit$cluster])
wss <- (nrow(cbind(x,y))-1)*sum(apply(cbind(x,y),2,var))
for (i in 1:4) wss[i] <- sum(kmeans(cbind(x,y),
  centers=i)$withinss)
plot(1:4, wss, type="b", xlab="Clusterzahl␣K",
  ylab="SQin(K)")
```

Beispiel 10.8

Wir betrachten das K-Means-Verfahren am Beispiel der Iris-Daten und wollen versuchen, die Pflanzen in drei Klassen (Spezies) zu Clustern. Um die Zuordnung zu den gegebenen Spezies-Werten zu überprüfen, erzeugen wir einen so genannten Coplot. Dieser erzeugt anhand der Kategorien eines nominal skalierten Merkmals Scatterplots der entsprechenden Teilmengen des Datensatzes. Hier betrachten wir beispielsweise die Länge und Breite der Blütenblätter.



Es werden 12 Pflanzen gemäß der Spezies nicht richtig eingeteilt.

```
fit <- kmeans(cbind(iris[1:4]),3,nstart=100,
  algorithm="Lloyd")
aggregate(cbind(iris[1:4]),by=list(fit$cluster),FUN=mean)
coplot(iris$Petal.Width~iris$Petal.Length|iris$Species,
  pch=19,col=c("red","blue","green")[fit$cluster])
```

Literatur

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] L. Fahrmeir, A. Hamerle und G. Tutz. *Multivariate statistische Verfahren*. Berlin: de Gruyter Verlag, 1996. ISBN: 3-11-013806-9.

[3] R. Kabacoff. *R in Action: Data Analysis and Graphics With R*. Manning Pubs Co Series. Manning Publications Company, 2011. ISBN: 9781935182399. URL: http://books.google.de/books?id=qWpWRwAACAAJ.

[4] S. Schäffler und D. Meintrup. *Stochastik*. Berlin Heidelberg: Springer Verlag, 2005. ISBN: 3-540-21676-6.

[5] J. Schira. *Statistische Methoden der VWL und BWL*. München Boston: Pearson Studium, 2009. ISBN: 978-3-86894-020-6.

[6] R. Schlittgen. *Multivariate Statistik*. Lehr- und Handbücher der Statistik : Fachgebiet Biometrie. Oldenbourg Wissensch.Vlg, 2009. ISBN: 9783486585957. URL: http://books.google.de/books?id=Qn2tPAAACAAJ.

[7] J.W. Tukey. *Exploratory Data Analysis*. New York: Addison-Wesley, 1977. ISBN: 0-201-07616-0.

[8] I. H. Witten, E. Frank und M. A. Hall. *Data Mining*. 3. Aufl. Morgan Kaufmann, 2011.

[9] H. Witting. *Mathematische Statistik I*. Teubner, 1985.

Index

σ-Algebra, 6
    Borelsche, 8
Äquivarianz, 24

Abfrage, 40
Ausreißer, 25

Bindung, 41
Boxplot, 29

Daten
    kleine standardisierte, 45
Datenmatrix, 18
Dichte, 9
    diskrete, 7
    Normalverteilung, 9
    Standardnormalverteilung, 9
Distanz
    Chebychev-, 85
    City-Block-, 85
    Euklidische-, 85
    Mahalanobis-, 87
    Quadrierte Euklidische-, 85

Elementarereignis, 5
Empirische Varianz
    verallgemeinerte, 43
Ereignis, 5
Ereignismenge, 5
Ergebnisraum, 5
Erwartungswert, 11

Faktorladung, 71
Funktion
    konkave, 66
    konvexe, 66

Grundgesamtheit, 17

Häufigkeiten
    absolute, 19
    relative, 19

Hauptachsen, 69
Hauptachsentransformation, 69
Hauptkomponenten, 70
Highlighting, 40
    linked, 40
Histogramm, 30

Interquartilsabstand, 29

Kaiserkriterium, 71
Kardinalskala, 18
Kerndichteschatzer, 32
Kernfunktion, 32
Kommunalität, 72
Korrelationskoeffizient
    empirischer, 36
Korrelationsmatrix
    empirische, 37
Kovarianz, 11
    empirische, 36

Lageparameter, 24
Linking, 40

MAD, 29
Median, 10
    empirischer, 29
Medianpunkt, 33
Merkmal, 18
Merkmalsraum, 18
Merkmalstrager, 17
Merkmalswert, 18
Mittelwert
    empirischer, 23
    empirischer getrimmter, 29
Modalwert, 10
Modus, 10

Nominalskala, 18

Ordnialskala, 18
Ordnungsstatistik, 29

i-te, 29
Orthogonalprojektion, 34

Parameterraum, 11

qq-Plot, 121
Quantil
α-, 9
empirisches α-, 29
Quantilfunktion, 9

Randverteilung, 14
Rang, 41
Rangkorrelationskoeffizient
Spearmanscher, 42
Rangvarianz, 42
Realisierungen, 19
Residuenkovarianzmatrix
empirische, 74
Robustheit, 25

Schätzfunktion, 23
Schätzwert, 23
Schiefe
empirische, 32
Schwerpunkt, 33
Scree-Test, 71
Selektion, 40
Sensitivitätsdiagramm, 25
Spur, 44
Standardabweichung
empirische, 23
Stichprobe, 18
Stichprobenmenge, 18
Stichprobenumfang, 18
stochastisch unabhangig, 14
Streuparameter, 24

Totalvariation, 43
Träger, 7

Varianz, 11
empirische, 23
Varianz-Kovarianz-Matrix
empirische, 36
Verteilung
Poisson, 7
Verteilungsfunktion, 8
diskrete, 8
empirische, 41

Wahrscheinlichkeit, 6
Wahrscheinlichkeitsdichtefunktion
gemeinsame, 13
Wahrscheinlichkeitsmaß, 6
diskretes, 7
stetiges, 8
Wahrscheinlichkeitsraum, 7
diskreter, 7
stetiger, 7
Warnung, 40
Whisker
oberer, 29
unterer, 29

Zähldichte, 7
Zufallsexperiment, 5
Zufallsvariable, 10