FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Interner Bericht

Statistische Verfahren für das Data Mining in einem Industrieprojekt

Thorsten Dickhaus

FZJ-ZAM-IB-2003-08

1. Auflage

(letzte Änderung: 23. Mai 2003)

Inhaltsverzeichnis

1 Problemstellung **1**

2 Korrelationsberechnung **3**
2.1 Motivation . . . 3
2.2 Theoretische Beschreibung . . . 3
2.3 Beispiele . . . 4
2.4 Anwendung . . . 7

3 Ausreißererkennung **9**
3.1 Motivation . . . 9
3.2 Theoretische Beschreibung . . . 11
3.3 Anwendung . . . 15
3.3.1 Naiver Ansatz . . . 15
3.3.2 Verwendung asymptotischer Verteilungsquantile . . . 16
3.3.3 Erkennen von „echten“ Ausreißern . . . 17
3.3.4 Mehrstufiges Vorgehen . . . 17
3.4 Ergebnisse . . . 18

4 Transformation von Variablen **21**
4.1 Motivation . . . 21
4.2 Anwendung . . . 22

5 Rangkorrelation **25**
5.1 Motivation . . . 25
5.2 Theorie . . . 28
5.3 Anwendung . . . 33

6 Hauptkomponentenanalyse **35**
6.1 Motivation . . . 35
6.2 Theoretischer Hintergrund . . . 35
6.3 Beispiel . . . 37
6.4 Anwendung . . . 37
6.5 Ausreißererkennung mit Hauptkomponenten . . . 39
6.5.1 Eindimensionale Ausreißer . . . 40
6.5.2 *p*-dimensionale Ausreißer . . . 41
6.5.3 Schlecht modellierte Beobachtungen . . . 42

7 Variablenselektion **43**
7.1 Motivation . . . 43
7.2 Auf Hauptkomponenten basierende Verfahren . . . 43
7.2.1 Selektion mit *p* Hauptkomponentenanalysen . . . 45

7.2.2 Selektion mit genau einer Hauptkomponentenanalyse 45
7.2.3 Elimination mit *(m-p)* Hauptkomponentenanalysen 46
7.2.4 Elimination mit genau einer Hauptkomponentenanalyse 46
7.3 Verfahren der „Principal variables“ 47
7.4 Anwendung 47
7.5 Ergebnisse 48

A Datenmaterial zu Beispiel 2.3.2 **49**

B Datenmaterial zu Beispiel 6.3.1 **55**

C Quantile der $\mathcal{N}$(0,1)-Verteilung zu Kapitel 6.5.3 **57**

Abbildungsverzeichnis

2.1 Körpergröße gegen Körpergewicht 5
2.2 Jahre in der „major league“ gegen Treffer im Jahre 1986 6
2.3 Menge Schädlingsbekämpfungsmittel gegen Ernteertrag 7

3.1 Ausreißer induziert signifikante Korrelation 10
3.2 Ausreißer verdecken signifikante Korrelation 10
3.3 Aufgedeckte Korrelation durch Vernachlässigung von Ausreißern 11
3.4 Ausreißer trotz geringen euklidischen Abstandes 12
3.5 Variable 373 gegen Variable 374 (alle 109 Beobachtungen) 17
3.6 Variable 373 gegen Variable 374 (restliche 107 Beobachtungen) 18

4.1 Menge Schädlingsbekämpfungsmittel gegen transformierten Ernteertrag 21

5.1 $Y = X^2 + \varepsilon$ 26
5.2 $Y = X^3 + \varepsilon$ 26
5.3 $Y = \exp(X) + \varepsilon$ 27
5.4 Streubild der Ränge zu Abbildungen 5.1 bis 5.3 27

Verzeichnis der Abkürzungen und Symbole

$\delta_{i,j}$	Kronecker-Symbol
$det(A)$	Determinante der Matrix A
$diag(\lambda_i)$	Diagonalmatrix mit den Einträgen λ_i
$\mathbb{E}$	Erwartungswertoperator
$i.i.d.$	independent and identically distributed
I_M	Indikatorfunktion der Menge M
$\|M\|$	Mächtigkeit der Menge M
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Parametern μ und σ^2
$\mathbb{P}$	Wahrscheinlichkeitsoperator
$\Phi(\cdot)$	Verteilungsfunktion der $\mathcal{N}$(0,1)-Verteilung
$sgn(\cdot)$	Vorzeichenoperator
$tr(A)$	Spur der Matrix A
A^t	Transponierte Matrix A
v^t	Transponierter Vektor v

Statistische Verfahren für das Data Mining in einem Industrieprojekt

In dem Prozess der Medikamentenentwicklung müssen aus einer sehr großen Anzahl von chemischen Substanzen diejenigen ausfindig gemacht werden, die im Organismus eine bestimmte, gewünschte Wirkung (wie z.B. die Linderung eines Schmerzreizes) erzielen. Der genaue Zusammenhang zwischen der chemischen Struktur einer Substanz und ihren daraus resultierenden biologischen Eigenschaften ist jedoch bisher weitgehend unbekannt. Die Aufgabenstellung, derartige unbekannte Zusammenhänge, also „versteckte“ Information, in erfassten Daten aufzudecken, wird auch als Data Mining bezeichnet. Am Anfang jedes Data Mining-Projektes steht eine Datenvorauswertung. Diese umfasst unter anderem eine Bereinigung sowie eine Reduktion der Daten.

In einer Diplomarbeit wurden im Rahmen eines von einem Aachener Pharmakonzern initiierten Projektes verschiedene statistische Verfahren für das Data Mining eingesetzt, um Daten aus der pharmazeutischen Forschung zu analysieren und (vor-)auszuwerten. Unter anderem wurden Algorithmen zur Korrelationsberechnung, Ausreißerbehandlung und Datenreduktion untersucht und programmiert. Die Implementierung erfolgte in der Programmiersprache C und wird zur Zeit schon in der Forschungsabteilung des Projektpartners eingesetzt.

Statistical methods for Data Mining in an industrial project

In the process of developing a new drug, chemical substances with special properties (i.e. the ability of easing a pain stimulus) have to be chosen from a very large amount of possible ones. However, the exact connection between the chemical structure of such a substance and its resulting biological properties is yet unknown. The task to detect such unknown relationships or “hidden“ information in collected data is called Data Mining. The beginning of each Data Mining project forms a preprocessing step. It consists of validation and reduction of data.

In a diploma thesis, statistical methods for Data Mining were used to analyse and (pre-)process pharmaceutical data within the scope of a project raised by a pharma company from Aachen. Among other things, algorithms for computation of correlations, treatment of outliers and data reduction were examined and programmed. The implementation was done in the C programming language and is already used in the research department of the project partner.

Kapitel 1

Problemstellung

Die Entwicklung eines neuen Medikamentes ist ein langwieriger Prozess: Aus einer sehr großen Anzahl von chemischen Substanzen müssen diejenigen ausgewählt werden, die im Organismus eine bestimmte, gewünschte Wirkung (wie z.B. die Linderung eines Schmerzreizes) erzielen.

Diese Fähigkeit ist eine biologische Eigenschaft der betreffenden Substanz und kann mit geeigneten Messverfahren bzw. Versuchen quantifiziert werden. Solche Messungen werden in der pharmazeutischen Forschung intensiv durchgeführt und ihre Ergebnisse in Datenbanken abgespeichert.

Hierbei treten verschiedene Probleme auf:

- Laborexperimente sind oft aufwändig und teuer.
- Die Datenhaltung ist nicht immer einfach zu organisieren, so dass es zu Daten-Inkonsistenzen, z.B. durch Fehleingaben, kommen kann.
- Die Anzahl an Messwerten wird leicht unüberschaubar groß.

Aus diesen Gründen ist die pharmazeutische Industrie daran interessiert, einerseits nach Möglichkeit nur die aussagekräftigsten Versuche wirklich durchzuführen und andererseits die entstehenden Daten von Fehlern bzw. „Ausreißern“ zu befreien.

Hier bietet sich ein Ansatzpunkt für statistische Verfahren:
Es ist möglich, Ähnlichkeiten zwischen Messreihen aufzudecken und damit gegebenenfalls Experimente einzusparen, die keine wesentlich neue Information gegenüber vorangegangenen Versuchen liefern. Diese Ähnlichkeiten werden mathematisch mit der Größe der **Korrelation** zwischen Variablen beschrieben (eine Messreihe bzw. ein Experiment wird also hier auf eine Zufallsvariable abgebildet).
Ferner lässt sich auch die Eigenschaft „Ausreißer“ einzelner Messpunkte (chemischer Strukturen, die in einer Messung untersucht worden sind) statistisch präzisieren und so geartete Daten können erkannt werden. Dies hat zum einen den Effekt, dass der Experimentator bzw. der mit der Datenauswertung beschäftigte Wissenschaftler auf eventuelle Messfehler aufmerksam wird (unerwünschte Ausreißer) und zum anderen können gegebenenfalls Strukturen mit besonders günstigen Eigenschaften, die für die Medikamentenentwicklung vielleicht besonders geeignet sind, schneller aus der großen erfassten Datenmenge herausgefiltert werden (gewünschte Ausreißer).

Eine andere Vorgehensweise besteht darin, dass die Ursachen der biologischen Eigenschaften einer Substanz in ihren physiko-chemischen Eigenschaften wie zum Beispiel der Anordnung der Atome und den Winkeln zwischen den Brückenbindungen in den Molekülen gesucht werden. Diese physiko-chemischen Eigenschaften lassen sich in der Regel viel leichter ermitteln und sind für viele relevante Substanzen häufig schon in großer Menge bekannt. Wenn es nun gelingt, aus den physiko-chemischen Eigenschaften die biologischen Eigenschaften vorherzusagen, lassen sich im

experimentativen Bereich zum Teil erhebliche Ressourcen einsparen.

Auch hierfür bieten sich Verfahren der mathematischen Statistik an, im Wesentlichen sind dies Klassifikationsalgorithmen.

Die beschriebene Aufgabenstellung, „versteckte“ Information in erfassten Daten aufzudecken, wird auch als **Data Mining** bezeichnet.
In der vorliegenden Arbeit wurden verschiedene statistische Verfahren des Data Mining eingesetzt, um Daten aus der pharmazeutischen Forschung zu analysieren und auszuwerten.
Die Implementation erfolgte in der Programmiersprache C und wird zur Zeit schon in der Forschungsabteilung eines großen Pharmakonzerns eingesetzt.

Kapitel 2

Korrelationsberechnung

2.1 Motivation

Wie bereits in den einleitenden Bemerkungen angeführt, ist es oftmals von Interesse, Ähnlichkeiten bzw. Abhängigkeiten von Messreihen zu erkennen und auszunutzen. Eine Möglichkeit dazu besteht darin, die beobachteten Werte als zweidimensionales Streubild darzustellen und die Form der sich ergebenden „Punktewolke“ zu analysieren. Dies ist bei der Analyse von einigen wenigen Messreihen ein durchaus gängiges und plausibles Vorgehen. Liegen jedoch, wie im vorliegenden Falle der Datenerfassung in der pharmazeutischen Forschung, sehr viele Messreihen vor, so wird die Anzahl der zu betrachtenden Streubilder (englisch: „scatter plots“) schnell unüberschaubar groß. Ist nämlich die Anzahl zu analysierender Messreihen gleich n, so ergeben sich $\frac{n\cdot(n-1)}{2}$ Streubilder, ihre Anzahl wächst also quadratisch mit der Anzahl der Messreihen.

Eine effizientere Form der Analyse dieser großen Datenmengen besteht darin, die Messreihen auf Zufallsvariablen abzubilden und statistische Kenngrößen der sich ergebenden Größen zu berechnen und systematisch zu analysieren. In der gegebenen Problemstellung des Messens von Ähnlichkeiten bzw. Abhängigkeiten von Zufallsgrößen (und damit der mit ihnen identifizierten Messreihen) ist insbesondere die Berechnung von Kovarianzen und Korrelationen ein geeignetes Hilfsmittel. Im Folgenden werden diese Konzepte zunächst theoretisch beschrieben und danach anhand von ausgewählten Beispielen das praktische Vorgehen erläutert.

2.2 Theoretische Beschreibung

Definition 2.2.1 (Kovarianz)
Gegeben seien zwei Zufallsvariablen X und Y.
Die Kovarianz von X und Y ist definiert als

$$Kov(X,Y) = \mathbb{E}((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))).$$

Sind die Verteilungen der zu Grunde liegenden Zufallsvariablen unbekannt und liegen lediglich Realisierungen $x_1, \ldots, x_n$ von X sowie $y_1, \ldots, y_n$ von Y vor, so verwendet man die folgende Schätzung:

Definition 2.2.2 (Empirische Kovarianz)
Gegeben seien zwei Zufallsvariablen X und Y und Realisierungen $x_1, \ldots, x_n$ von X sowie $y_1, \ldots, y_n$ von Y.

Die empirische Kovarianz von X und Y ist definiert als

$$Kov_n(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}((x_i - \bar{x}_n)\cdot(y_i - \bar{y}_n)).$$

Hierbei bezeichnet $\bar{x}_n$ bzw. $\bar{y}_n$ das arithmetische Mittel der beobachteten Werte der Zufallsvariablen X bzw. Y.

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Normiert man nun diese Kovarianzen auf das Intervall $[-1;1]$ bzw. berechnet man die Kovarianzen von auf Varianz 1 normierten Zufallsvariablen, so erhält man den (empirischen) **Korrelationskoeffizienten nach Pearson** oder die (empirische) **Produktmomentkorrelation** von X und Y, bezeichnet mit $\rho(X,Y)$ bzw. $\rho_n(X,Y)$.

Definition 2.2.3 (Pearson-Korrelationskoeffizient, Produktmomentkorrelation)

$$\begin{aligned}\rho(X,Y) &= \frac{Kov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}}\\ \rho_n(X,Y) &= \frac{Kov_n(X,Y)}{\sqrt{Var_n(X)}\cdot\sqrt{Var_n(Y)}}\end{aligned}$$

$Var(X)$ bzw. $Var_n(X)$ bezeichnet hierbei die (empirische) Varianz der Zufallsvariablen X, also

$$\begin{aligned}Var(X) &= \mathbb{E}((X-\mathbb{E}(X))^2) &&\equiv Kov(X,X)\\ Var_n(X) &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 &&\equiv Kov_n(X,X).\end{aligned}$$

Der so definierte Korrelationskoeffizient $\rho(X,Y)$ (im Folgenden auch kurz „die Korrelation zwischen X und Y“) ist ein Maß für den linearen Zusammenhang von X und Y. $|\rho(X,Y)| = 1$ bedeutet, dass eine eindeutige lineare Beziehung zwischen den beiden Zufallsvariablen X und Y besteht; analog lässt ein betragsmäßig großer Wert von $\rho_n(X,Y)$ auf eine lineare Abhängigkeit der beiden erfassten Stichproben schließen. Betragsmäßig kleine Werte der (empirischen) Korrelation hingegen deuten darauf hin, dass keine lineare Beziehung der betrachteten Zufallsgrößen bzw. deren Realisierungen zueinander besteht. Dies schließt jedoch nicht aus, dass ein nicht-linearer Zusammenhang zwischen X und Y vorliegt. Dazu mehr in den Kapiteln 4 und 5.

2.3 Beispiele

Beispiel 2.3.1 (Linearer Zusammenhang von *X* und *Y*)
Ein typisches Beispiel für eine lineare Beziehung zwischen zwei erfassten Merkmalen ist der Zusammenhang von Körpergröße X und Körpergewicht Y. Aus verschiedenen Studien ergibt sich, dass sich diese beiden Größen linear in Beziehung setzen lassen. Zum Beispiel hat K. Meyerbaum in [1] Daten von sechs zufällig ausgewählten Statistikprofessoren erhoben. Die Ergebnisse sind in der folgenden Tabelle aufgeführt:

x_i/cm	154	165	176	182	189	191
y_i/kg	73	76	85	101	92	108

In der folgenden Abbildung sind diese Werte in einem zweidimensionalen Streubild dargestellt:

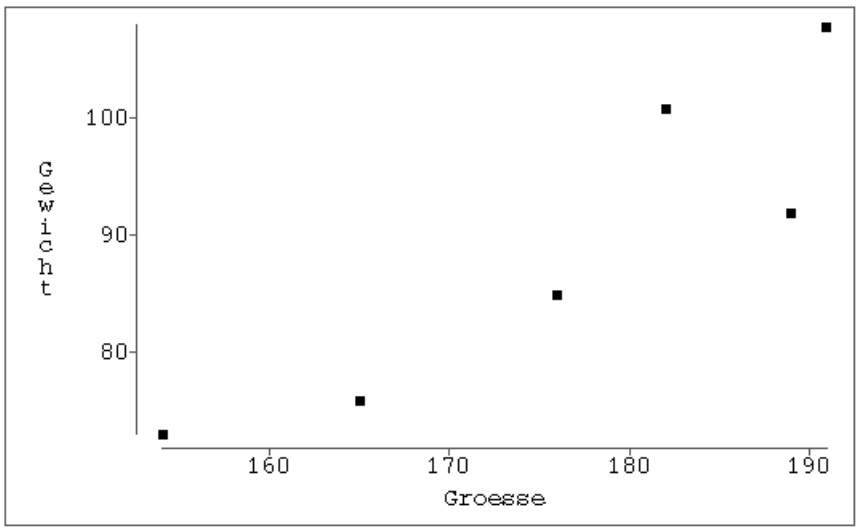

Abbildung 2.1: Körpergröße gegen Körpergewicht

Die Form der Punktewolke legt die Vermutung nahe, dass hier ein linearer Zusammenhang der beiden Merkmale gegeben ist. Will man dies nun auch mathematisch messen, so ergibt sich:

$$\begin{aligned} \bar{x}_6 &= \tfrac{1}{6} \cdot (154 + 165 + 176 + 182 + 189 + 191) \approx 176.17. \\ \bar{y}_6 &= \tfrac{1}{6} \cdot (73 + 76 + 85 + 101 + 92 + 108) \approx 89.17. \end{aligned}$$

$$\begin{aligned} Kov_6(X,Y) &= \tfrac{1}{5} \cdot ((154 - 176.17)(73 - 89.17) + (165 - 176.17)(76 - 89.17) \\ &\quad + (176 - 176.17)(85 - 89.17) + (182 - 176.17)(101 - 89.17) \\ &\quad + (189 - 176.17)(92 - 89.17) + (191 - 176.17)(108 - 89.17)) \\ &\approx 213.8. \end{aligned}$$

$$\begin{aligned} Var_6(X) = Kov_6(X,X) &= \tfrac{1}{5} \cdot ((154 - 176.17)^2 + (165 - 176.17)^2 \\ &\quad + (176 - 176.17)^2 + (182 - 176.17)^2 \\ &\quad + (189 - 176.17)^2 + (191 - 176.17)^2) \\ &\approx 248.36. \end{aligned}$$

$$\begin{aligned} Var_6(Y) = Kov_6(Y,Y) &= \tfrac{1}{5} \cdot ((73 - 89.17)^2 + (76 - 89.17)^2 \\ &\quad + (85 - 89.17)^2 + (101 - 89.17)^2 \\ &\quad + (92 - 89.17)^2 + (108 - 89.17)^2) \\ &\approx 229.16. \end{aligned}$$

$$\Rightarrow \rho_6(X,Y) = \frac{213.8}{\sqrt{248.36} \cdot \sqrt{229.16}} \approx 0.896.$$

Dieser hohe Wert der empirischen Korrelation von X und Y bestätigt den Eindruck, der aus dem scatter plot gewonnen wurde.

Beispiel 2.3.2 (Unabhängige Zufallsvariablen)
In der Publikation von Collier Books [2] wurden Daten der amerikanischen „major league baseball“ für das Spieljahr 1986 zusammengetragen. Unter anderem wurde für $n = 322$ Spieler ermittelt, wie viele Jahre sie bereits in der „major league“ spielen (Zufallsvariable X) und wieviele Treffer sie in dem betrachteten Spieljahr 1986 erzielt haben (Zufallsvariable Y).
Das nachfolgende Streubild dieser beiden Merkmale zeigt, dass sie offenbar nicht voneinander abhängen, da die Punktewolke keine erkennbare Struktur aufweist.

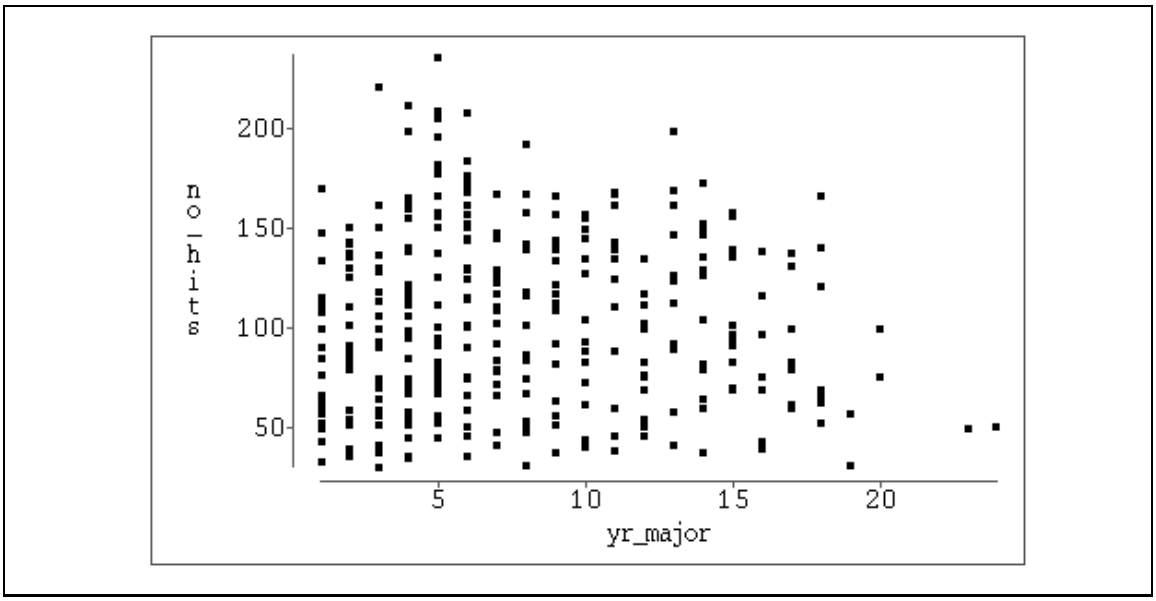

Abbildung 2.2: Jahre in der „major league“ gegen Treffer im Jahre 1986

Auch in diesem Fall wird der Eindruck des scatter plots durch den Wert der statistischen Größe $\rho_{322}(X,Y)$ gestützt. Die Berechnung von $\rho_{322}(X,Y)$ wurde mit der erstellten Statistik-Software durchgeführt und wird in Anhang A zusammen mit dem zu Grunde liegenden Datenmaterial aufgeführt.

Der berechnete Wert $\rho_{322}(X,Y) = -0.00803$ ist betragsmäßig sehr klein, was darauf hindeutet, dass kein linearer Zusammenhang von X und Y gegeben ist. In diesem Fall legt der scatter plot sogar die Vermutung nahe, dass überhaupt kein Zusammenhang zwischen den Merkmalen besteht, diese also voneinander unabhängig sind.

Dass man diesen Schluss jedoch nicht allein auf Basis des (empirischen) Korrelationskoeffizienten ziehen darf, soll das abschließende Beispiel verdeutlichen.

Beispiel 2.3.3 (Nichtlinearer Zusammenhang von *X* und *Y*)
Die folgende Tabelle aus [3] enthält Datenmaterial aus der Versuchsanlage einer Baumschule. Gemessen wurden die aufgebrachte Menge eines biologischen Schädlingsbekämpfungsmittels (Zufallsgröße X) und der Ernteertrag Y der damit behandelten Stachelbeersträucher:

$x_i/(g/m^2)$	1.3	0.3	2.5	2.1	2.3	0.2	1.8	0.6	2.6	0
$y_i/(kg/m^2)$	4.8	3.4	2.6	4.0	3.5	2.9	4.6	4.2	2.1	2.4

Auch hier wird im Folgenden nur das Ergebnis der Berechnung des empirischen Korrelationskoeffizienten angegeben (Programm-Output):

```
Feld  1 1.300000        4.800000
Feld  2 0.300000        3.400000
Feld  3 2.500000        2.600000
Feld  4 2.100000        4.000000
Feld  5 2.300000        3.500000
Feld  6 0.200000        2.900000
Feld  7 1.800000        4.600000
Feld  8 0.600000        4.200000
Feld  9 2.600000        2.100000
Feld 10 0.000000        2.400000

Die empirische Korrelation zwischen X und Y betraegt   -0.02490.
```

Der geringe Wert von $|\rho_{10}(X,Y)| = 0.02490$ ist zwar auch hier ein Hinweis darauf, dass der Einsatz des Schädlingsbekämpfungsmittels nicht linear mit dem Ernteertrag in Beziehung zu setzen

ist. Die in der folgenden Abbildung wiedergegebene Darstellung des zweidimensionalen Streubildes der beiden Merkmale legt jedoch nahe, hier einen quadratischen Zusammenhang von X und Y zu vermuten.

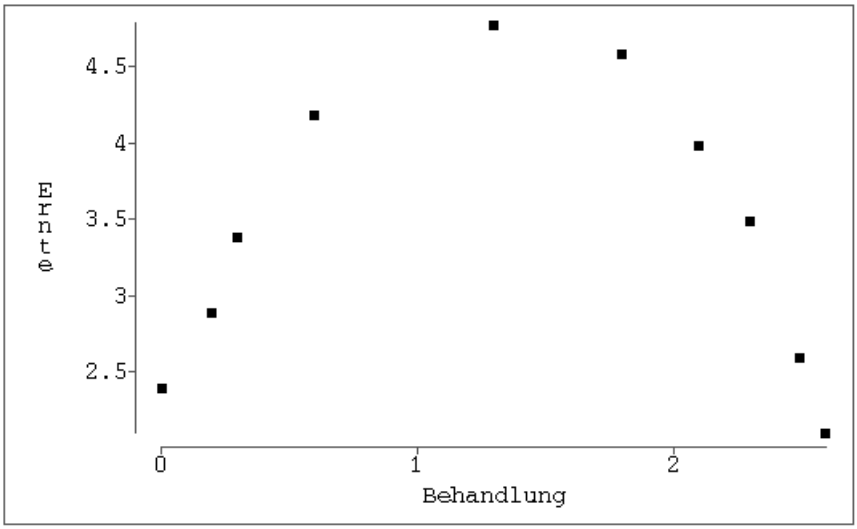

Abbildung 2.3: Menge Schädlingsbekämpfungsmittel gegen Ernteertrag

Dieser nicht-lineare Zusammenhang kann nicht aus der Berechnung der Korrelation abgeleitet werden.
In den Kapiteln 4 und 5 werden deswegen Verfahren erläutert, mit denen sich zumindest monotone, nicht-lineare Abhängigkeiten identifizieren lassen.

2.4 Anwendung

In der gegebenen Aufgabenstellung wurden die Experimente bzw. Laboruntersuchungen und ihre Ergebnisse als Messreihen interpretiert und durch Zufallsvariablen repräsentiert. Die Realisierungen dieser Größen ergeben sich durch die verschiedenen chemischen Substanzen, die in den Experimenten eingesetzt werden. So ergibt sich als Eingabedatensatz eine $(n \times m)$-Datenmatrix. Hierbei ist m die Anzahl der durchgeführten Experimente und n bezeichnet die Anzahl der eingesetzten Substanzen (chemische Strukturen). Es wurden nun paarweise die Korrelationen der Spalten dieser Eingabematrix berechnet und betragsmäßig signifikant große Werte in gesonderten Tabellen ausgegeben.

Probleme ergaben sich dabei hauptsächlich durch Inkonsistenzen in den Eingabedatensätzen (oft werden nicht alle Experimente für alle Strukturen durchgeführt), die zu fehlenden Werten (englisch: „missing values“) führen. Ist die Anzahl dieser „missing values“ relativ zur Gesamtanzahl n der insgesamt betrachteten Strukturen gesehen groß, so ergibt sich eine dünn besetzte Eingabe-Datenmatrix. Praktisch wurde dieses Problem dadurch gelöst, dass die Korrelationsberechnung nur auf der Basis derjenigen Strukturen durchgeführt wurde, für welche beide jeweils betrachteten Experimente Werte lieferten.

Ein weiteres Problem besteht darin, wie der Terminus „signifikant groß“ mathematisch präzisiert werden kann. Ein gängiges statistisches Vorgehen dazu besteht darin, Verteilungsannahmen über die Variablen (im vorliegenden Fall durch die Messreihen induziert) zu treffen und aufgrund dessen

Quantile der zugehörigen Verteilungsfunktionen als Schwellenwerte für die Größe der berechneten Statistiken zu verwenden. Dies war jedoch in der bearbeiteten Aufgabe nicht möglich, da Detailwissen über die Art der Messungen nicht weitergegeben wurde und somit komplett frei von Verteilungsannahmen gearbeitet werden musste. Deswegen wurde die erstellte Software so ausgelegt, dass der Benutzer (der analysierende Wissenschaftler) diese Schwellenwerte selbst eingeben und damit die Signifikanz der Korrelationen und in der Konsequenz auch die Größe der ausgegebenen Datensätze eigenhändig steuern kann.

In der praktischen Verwendung hat sich dies als ein praktikables Vorgehen erwiesen und es konnten in Testszenarien, bei welchen die Ähnlichkeiten der Messungen von vorne herein durch Expertenwissen belegt waren, diese auch hinreichend gut erkannt werden.

Kapitel 3

Ausreißererkennung

3.1 Motivation

Neben der in Kapitel 2 beschriebenen Aufgabe, Regelmäßigkeiten in den Daten aufzudecken, ist es oftmals ebenso von Interesse, Datenpunkte (Messwerte) zu erkennen, die signifikant aus dieser ermittelten Struktur herausfallen bzw. sich in der Punktewolke stark außerhalb des Gros' der gemessenen Daten befinden. Ein Datenpunkt mit dieser Eigenschaft wird im Folgenden als **„Ausreißer"** bezeichnet. Das Vorliegen von Ausreißern kann verschiedene Gründe haben:

1. Eventuell ist eine fehlerhafte Messung durchgeführt oder ein Messwert falsch eingegeben worden. Ein so gearteter Ausreißer ist sicherlich unerwünscht und der analysierende Wissenschaftler möchte diese Daten-Inkonsistenz natürlich gerne erkennen und gegebenfalls beheben.

2. Wie bereits eingangs beschrieben ist ein Ziel der durchgeführten Data Mining-Methoden, chemische Substanzen mit besonders günstigen Eigenschaften sytematisch aus der großen erfassten Datenmenge herauszufiltern. Diese günstigen Eigenschaften lassen sich in den experimentell gemessenen Daten als außergewöhnliche Werte und damit als „erwünschte Ausreißer" ablesen.

Ein weiterer Beweggrund dafür, Ausreißer aufzufinden und unter Umständen zu eliminieren (von der Analyse auszunehmen), soll durch die folgenden Abbildungen von Punktewolken mit den zugehörigen Korrelations-Werten motiviert werden.

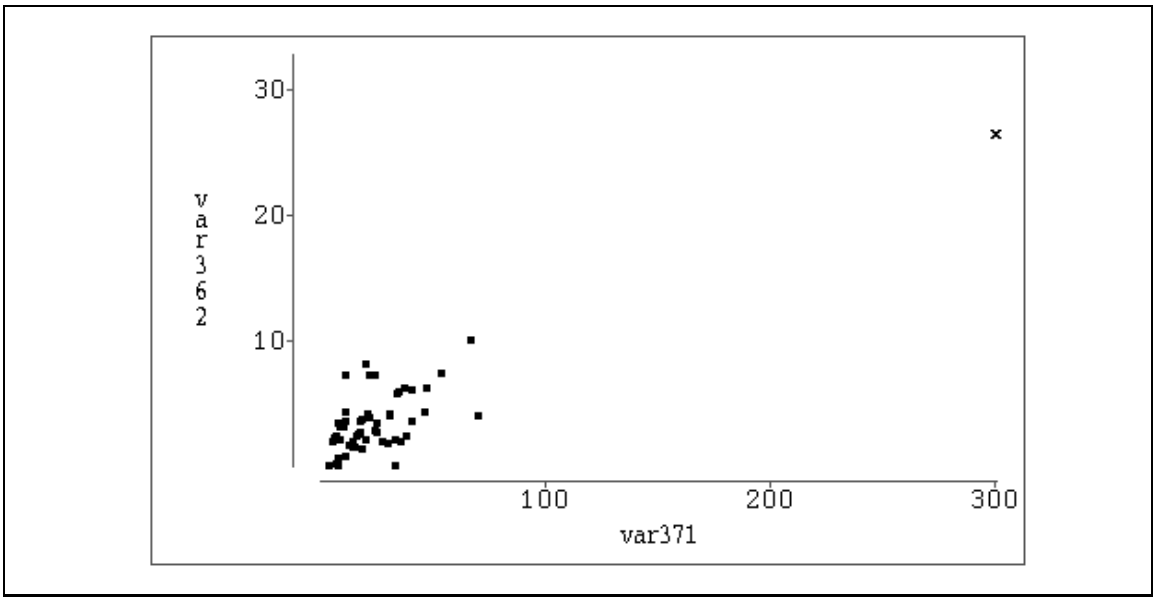

Abbildung 3.1: Ausreißer induziert signifikante Korrelation

In dem in Abbildung 3.1 dargestellten Fall liefert die Korrelationsberechnung einen hohen Wert von $\rho_n(X, Y) = 0.8703$, obwohl der überwiegende Teil der Beobachtungen eine unregelmäßige Struktur aufweist. Hier induziert also der in der Abbildung markierte, weit vom Datenzentrum entfernte Datenpunkt eine lineare Abhängigkeit, die jedoch kein Charakteristikum der Stichprobe insgesamt darstellt.

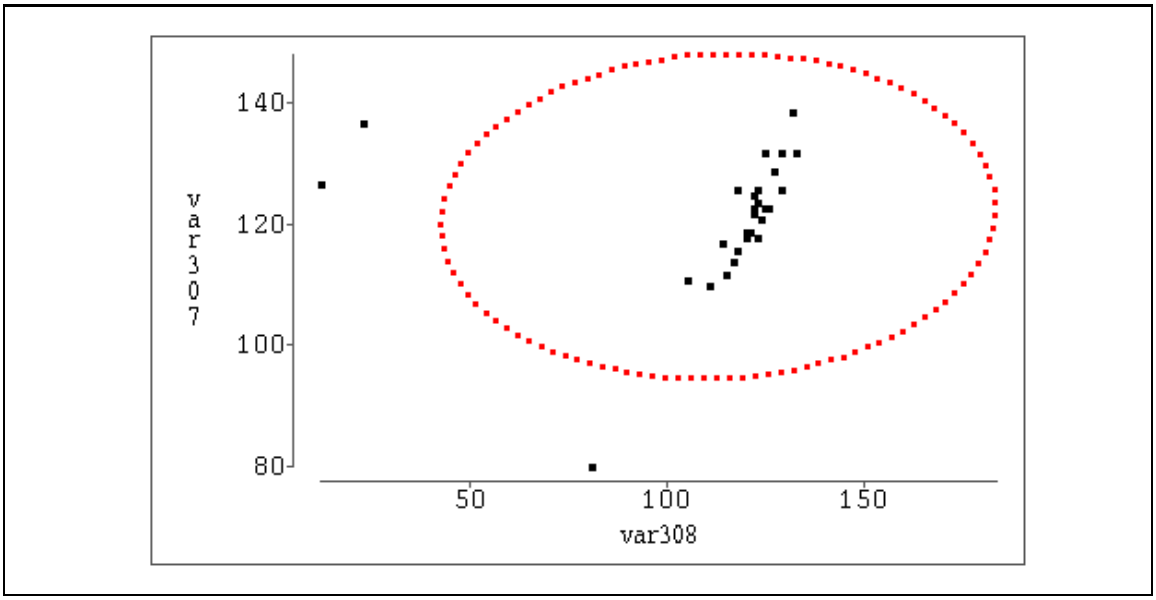

Abbildung 3.2: Ausreißer verdecken signifikante Korrelation

Der gegenteilige Fall ist in Abbildung 3.2 dargestellt. Das Gros der Daten zeigt eine lineare Beziehung zwischen VAR308 und VAR307. Die drei Datenpunkte, die außerhalb der eingezeichneten Ellipse liegen, „verschleiern“ jedoch diese Abhängigkeit und führen dazu, dass sich kein signifikant großer Wert für $\rho_n(VAR308, VAR307)$ errechnet. Der berechnete Wert für die empirische Korrelation ist in diesem Fall $\rho_{29}(VAR308, VAR307) = 0.0560$.

Nimmt man die drei Ausreißer von der Berechnung aus, so zeigt sich die lineare Beziehung des restlichen Datenmaterials eindeutig in der Kennzahl
$\rho_{26}(VAR308, VAR307) = 0.8609$. Grafisch wird dies in Abbildung 3.3 deutlich.

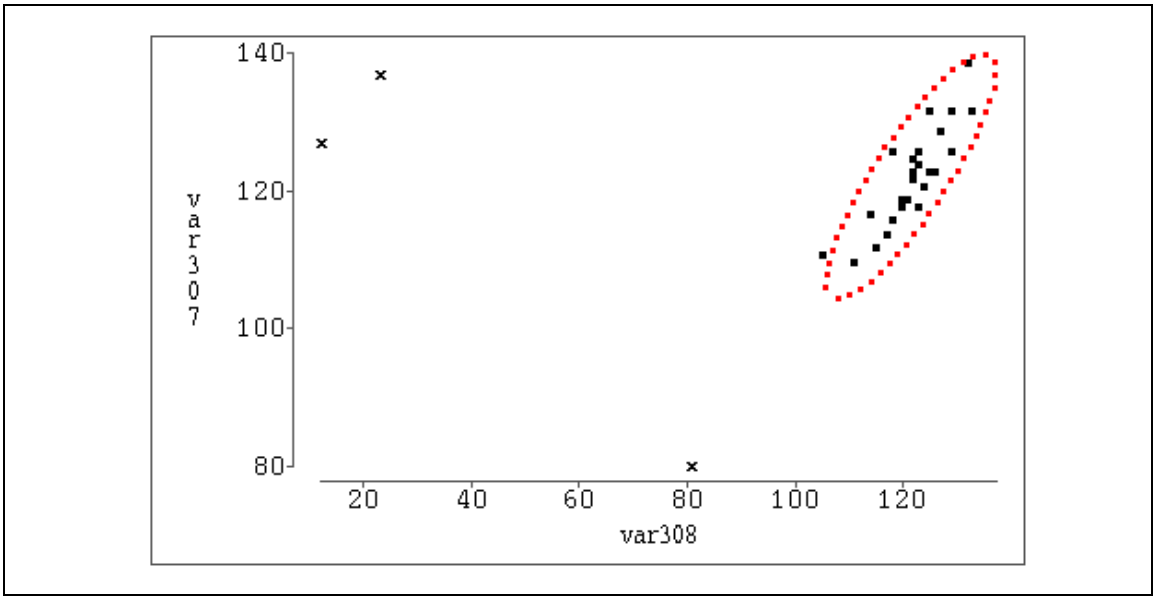

Abbildung 3.3: Aufgedeckte Korrelation durch Vernachlässigung von Ausreißern

Wie diese beiden Beispiele zeigen, können einzelne Ausreißer zum Teil erheblichen Einfluss auf die Korrelationsberechnung haben. Dies liegt im Wesentlichen daran, dass in der Formel für $\rho_n(X, Y)$ jedem Datenpunkt das gleiche Gewicht beigemessen wird und sich daher unverhältnismäßig große lokale Abweichungen vom Gesamtverhalten erheblich in der globalen Kenngröße der empirischen Korrelation niederschlagen.

Die Hauptschwierigkeit besteht nun darin, diese lokalen Inhomogenitäten mathematisch zu messen (es kann nicht jedes Streubild händisch betrachtet und ausgewertet werden) und zu beurteilen, wann diese als signifikant zu betrachten sind und nicht lediglich in der üblichen Schwankungsbreite der experimentellen Ungenauigkeit liegen.

3.2 Theoretische Beschreibung

Die zunächst naheliegendste Vorgehensweise zum Entdecken von Ausreißern besteht darin, für jeden Datenpunkt $z_i = (x_i, y_i), i = 1, \ldots, n$ der erhobenen zweidimensionalen Stichprobe dessen euklidischen Abstand vom Datenzentrum $\bar{z}_n = (\bar{x}_n, \bar{y}_n)$ zu ermitteln und als Abstandsmaß zu verwenden.

Definition 3.2.1 (Euklidischer Abstand im zweidimensionalen Raum)
Der euklidische Abstand $d(a, b)$ zweier Punkte $a = (a_1, a_2)$ und $b = (b_1, b_2) \in \mathbb{R}^2$ ist definiert als

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}.$$

Stammt $z_i = (x_i, y_i)$ aus einer Stichprobe zweier Merkmale X und Y vom Umfang n mit arithmetischen Mitteln $\bar{x}_n$ und $\bar{y}_n$, so misst

$$d(z_i, \bar{z}_n) = \sqrt{(x_i - \bar{x}_n)^2 + (y_i - \bar{y}_n)^2}$$

den euklidischen Abstand des Beobachtungspaares $z_i = (x_i, y_i)$ *vom Datenzentrum* $\bar{z}_n = (\bar{x}_n, \bar{y}_n)$.

Dass die Verwendung dieses Abstandsmaßes nicht immer sinnvoll ist, soll durch die folgende Abbildung 3.4 illustriert werden: Der markierte Datenpunkt hat zwar keinen großen euklidischen Abstand vom Datenzentrum, fällt aber doch aus der Punktewolke heraus, da er nicht in der generellen Streuungsrichtung liegt.

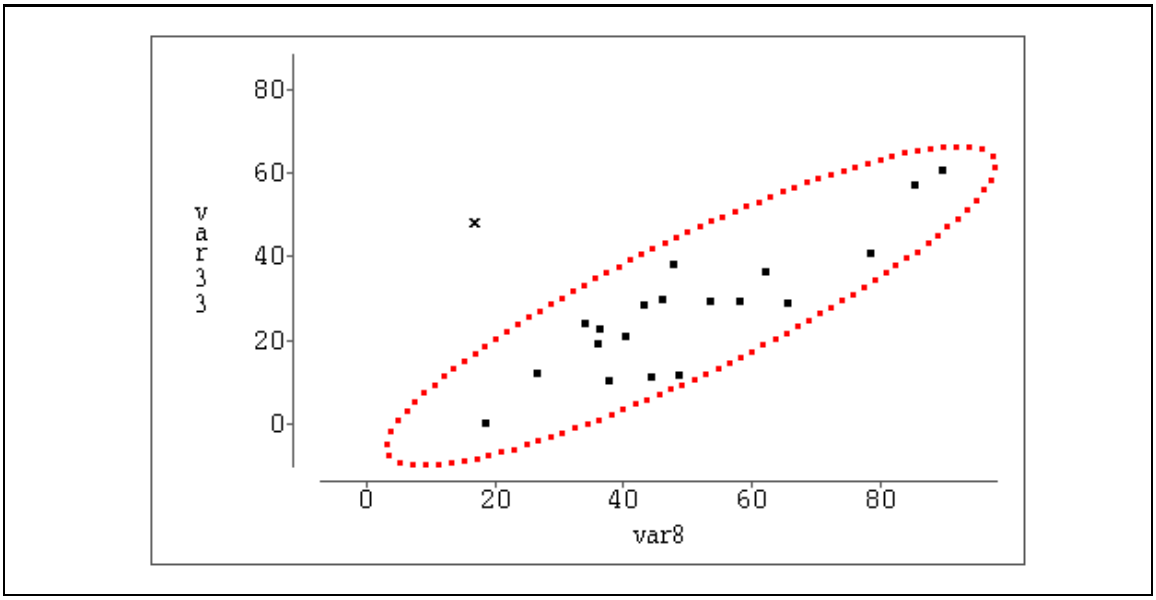

Abbildung 3.4: Ausreißer trotz geringen euklidischen Abstandes

Will man diese Zusatzinformation über die „Ausrichtung“ der Punktewolke mit in die Abstandsberechnung einfließen lassen, so bietet sich die Verwendung der sogenannten „Mahalanobis-Distanz“ $\Delta(z_i, \bar{z}_n)$ an. Hierbei wird in die Formel für den euklidischen Abstand eine Gewichtung mit der inversen Kovarianzmatrix eingerechnet, um der unterschiedlichen Streuung in x- und y-Richtung Rechnung zu tragen:

Definition 3.2.2 (Mahalanobis-Distanz)
Die quadrierte Mahalanobis-Distanz eines Punktes $z_i = (x_i, y_i)$ *aus einer Stichprobe zweier Merkmale X und Y vom Umfang n mit arithmetischen Mitteln* $\bar{x}_n$ *und* $\bar{y}_n$ *vom Datenzentrum* $\bar{z}_n = (\bar{x}_n, \bar{y}_n)$ *ist definiert als*

$$\Delta^2(z_i, \bar{z}_n) = \begin{pmatrix} (x_i - \bar{x}_n) \\ (y_i - \bar{y}_n) \end{pmatrix}^t \begin{pmatrix} Var_n(X) & Kov_n(X,Y) \\ Kov_n(X,Y) & Var_n(Y) \end{pmatrix}^{-1} \begin{pmatrix} (x_i - \bar{x}_n) \\ (y_i - \bar{y}_n) \end{pmatrix}.$$

Besonders einfach wird die Berechnung der Mahalanobis-Distanz in dem Fall, dass die zu Grunde liegenden Größen auf Mittelwert 0 und Varianz 1 standardisiert werden.

Folgerung 3.2.1 (Mahalanobis-Distanz standardisierter Größen)
Die quadrierte Mahalanobis-Distanz eines Punktes $z_i = (x_i, y_i)$ *aus einer Stichprobe zweier Merkmale X und Y vom Umfang n mit arithmetischen Mitteln* $\bar{x}_n = 0$ *und* $\bar{y}_n = 0$ *sowie* $Var_n(X) =$

$Var_n(Y) = 1$ *vom Datenzentrum* $\bar{z}_n = (0;0)$ *ergibt sich zu:*

$$\begin{aligned}
\Delta^2(z_i, \bar{z}_n) &= \begin{pmatrix} x_i \\ y_i \end{pmatrix}^t \begin{pmatrix} 1 & \rho_n(X,Y) \\ \rho_n(X,Y) & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \\
&= \frac{1}{1-\rho_n(X,Y)^2} \cdot \begin{pmatrix} x_i \\ y_i \end{pmatrix}^t \begin{pmatrix} 1 & -\rho_n(X,Y) \\ -\rho_n(X,Y) & 1 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \\
&= \frac{x_i^2 + y_i^2 - 2 \cdot x_i \cdot y_i \cdot \rho_n(X,Y)}{1-\rho_n(X,Y)^2}.
\end{aligned}$$

Aufgrund dessen wird man in der praktischen Berechnung die Größen X und Y zunächst standardisieren, damit die oben angegebene geschlossene Formel für $\Delta(z_i, \bar{z}_n)$ verwendet werden kann.

Geometrisch lässt sich der Übergang vom euklidischen Abstand zur Mahalanobis-Distanz so interpretieren, dass das zu Grunde liegende Koordinatensystem im $\mathbb{R}^2$ derart neu gewählt wird, dass die erste Achse in Richtung der stärksten Streuung (repräsentiert über den Richtungsvektor v_1) und die zweite Achse in Richtung $v_2 \perp v_1$ weist. Zudem wird der Koordinatenursprung in den Punkt $\bar{z}_n = (\bar{x}_n, \bar{y}_n)$ verschoben.
Überführt man nun den Zufallsvektor (X, Y) vermittels einer linearen Transformation so in den Zufallsvektor (C_1, C_2), dass dies äquivalent zu einem Basiswechsel von der kanonischen in die oben charakterisierte Basis ist, so bezeichnet man die transformierten Zufallsgrößen C_1 und C_2 als die „Hauptkomponenten“ der zu Grunde liegenden zweidimensionalen Verteilung.

Die Mahalanobis-Distanz ist nun definiert als der euklidische Abstand in dem veränderten Koordiantensystem, wobei eine Skalierung der neuen Koordinatenachsen in Einheiten der Standardabweichung der Hauptkomponenten stattfindet. Führt man diese Berechnung für standardisierte Größen durch, so ist dies äquivalent zur Lösung des Eigenwertproblems der Korrelationsmatrix

$$K_n := \begin{pmatrix} 1 & \rho_n(X,Y) \\ \rho_n(X,Y) & 1 \end{pmatrix},$$

denn es gilt:

$$\begin{aligned}
v_1 &= \arg(\max_{a \in \mathbb{R}^2} Var_n(a^t \cdot (X,Y)^t)) \\
&= \arg(\max_{a \in \mathbb{R}^2} a^t \cdot \begin{pmatrix} Var_n(X) & Kov_n(X,Y) \\ Kov_n(X,Y) & Var_n(Y) \end{pmatrix} \cdot a)
\end{aligned}$$

und (X, Y) wird hier als standardisiert vorausgesetzt, so dass die Varianz- / Kovarianzmatrix in die oben angegebene Korrelationsmatrix K_n übergeht.

Wie Ergebnisse aus der linearen Algebra (Maximierung von quadratischen Formen) zeigen, ergibt sich v_1 als Eigenvektor zum größeren Eigenwert λ_1 und v_2 als Eigenvektor zum kleineren Eigenwert λ_2 von K_n. Ferner ist die Varianz- / Kovarianzmatrix bezüglich des neuen Koordinatensystems gleich der diagonalisierten Korrelationsmatrix

$$K_D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

Anders ausgedrückt ist $Var_n(C_1) = \lambda_1$ und $Var_n(C_2) = \lambda_2$. Daraus ergibt sich, dass die oben angeführte Skalierung entlang der Hauptkomponenten dadurch zu geschehen hat, dass die Normierungsfaktoren $\frac{1}{\sqrt{\lambda_i}}, i = 1, 2$ für die Richtungen $v_i, i = 1, 2$ verwendet werden.

Die konkrete Berechnung ergibt die geschlossene Formel in Folgerung 3.2.1. Bezeichne hierzu $\chi_{K_n}(\lambda)$ das charakteristische Polynom der standardisierten Korrelationsmatrix K_n, so ergibt sich:

$$\chi_{K_n}(\lambda) = det(K_n - \lambda E) = \begin{vmatrix} (1-\lambda) & \rho_n(X,Y) \\ \rho_n(X,Y) & (1-\lambda) \end{vmatrix} = (1-\lambda)^2 - \rho_n(X,Y)^2.$$

$$\chi_{K_n}(\lambda) = 0 \Leftrightarrow (1-\lambda)^2 = \rho_n(X,Y)^2 \Leftrightarrow |(1-\lambda)| = |\rho_n(X,Y)|.$$

$$\Rightarrow \lambda_1 = 1 + |\rho_n(X,Y)| \wedge \lambda_2 = 1 - |\rho_n(X,Y)| \wedge \lambda_1 > \lambda_2.$$

$$(K_n - \lambda_1 \cdot E)v_1 = 0 \Leftrightarrow$$

$$\begin{pmatrix} -|\rho_n(X,Y)| & \rho_n(X,Y) \\ \rho_n(X,Y) & -|\rho_n(X,Y)| \end{pmatrix} v_1 = 0 \Leftrightarrow$$

$$\begin{aligned} -|\rho_n(X,Y)|v_{11} + \rho_n(X,Y)v_{12} &= 0 \\ \rho_n(X,Y)v_{11} - |\rho_n(X,Y)|v_{12} &= 0 \end{aligned}$$

$$\Rightarrow v_{11} = sgn(\rho_n(X,Y)) \cdot v_{12} \Rightarrow v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ sgn(\rho_n(X,Y)) \end{pmatrix}.$$

$$(K_n - \lambda_2 \cdot E)v_2 = 0 \Leftrightarrow$$

$$\begin{pmatrix} |\rho_n(X,Y)| & \rho_n(X,Y) \\ \rho_n(X,Y) & |\rho_n(X,Y)| \end{pmatrix} v_2 = 0 \Leftrightarrow$$

$$\begin{aligned} |\rho_n(X,Y)|v_{21} + \rho_n(X,Y)v_{22} &= 0 \\ \rho_n(X,Y)v_{21} + |\rho_n(X,Y)|v_{22} &= 0 \end{aligned}$$

$$\Rightarrow v_{22} = -sgn(\rho_n(X,Y)) \cdot v_{21} \Rightarrow v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -sgn(\rho_n(X,Y)) \end{pmatrix}.$$

Dadurch lassen sich nun die Matrix T der entsprechenden Basistransformation sowie die transformierten Koordianten eines Vektors $(x_i, y_i)^t \in \mathbb{R}^2$ als $(x_i^*, y_i^*) = (x_i, y_i) \cdot T$ berechnen:

$$T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ sgn(\rho_n(X,Y)) & -sgn(\rho_n(X,Y)) \end{pmatrix}$$

$$(x_i^*, y_i^*) = (x_i, y_i) \cdot T = \frac{1}{\sqrt{2}}(x_i, y_i) \begin{pmatrix} 1 & 1 \\ sgn(\rho_n(X,Y)) & -sgn(\rho_n(X,Y)) \end{pmatrix}$$

$$= \frac{1}{\sqrt{2}} \begin{pmatrix} x_i + sgn(\rho_n(X,Y)) \cdot y_i \\ x_i - sgn(\rho_n(X,Y)) \cdot y_i \end{pmatrix}^t.$$

Diese Darstellung gestattet nun die Berechnung von $\Delta(z_i, \bar{z}_n)$ eines Punktes (x_i, y_i) aus einer Stichprobe zweier Merkmale X und Y vom Umfang n mit arithmetischen Mitteln $\bar{x}_n = 0$ und $\bar{y}_n = 0$

sowie $Var_n(X) = Var_n(Y) = 1$ vom Datenzentrum $(0;0)$ als euklidischen Abstand des transformierten Punktes vom Koordinatenursprung unter Beachtung der Normierung entlang der Hauptkomponenten:

$$\begin{aligned}
\Delta^2(z_i, \bar{z}_n) &= (\frac{x_i + sgn(\rho_n(X,Y)) \cdot y_i}{\sqrt{2}\sqrt{1+|\rho_n(X,Y)|}})^2 + (\frac{x_i - sgn(\rho_n(X,Y)) \cdot y_i}{\sqrt{2}\sqrt{1-|\rho_n(X,Y)|}})^2 \\
&= \frac{x_i^2 + 2sgn(\rho_n(X,Y))x_iy_i + y_i^2}{2(1+|\rho_n(X,Y)|)} + \frac{x_i^2 - 2sgn(\rho_n(X,Y))x_iy_i + y_i^2}{2(1-|\rho_n(X,Y)|)} \\
&= \frac{(x_i^2 + 2sgn(\rho_n(X,Y))x_iy_i + y_i^2)(1-|\rho_n(X,Y)|)}{2(1-\rho_n(X,Y)^2)} \\
&\quad + \frac{(x_i^2 - 2sgn(\rho_n(X,Y))x_iy_i + y_i^2)(1+|\rho_n(X,Y)|)}{2(1-\rho_n(X,Y)^2)} \\
&= \frac{1}{2(1-\rho_n(X,Y)^2)} \cdot [2x_i^2 + 2y_i^2 - |\rho_n(X,Y)| \cdot (x_i^2 + 2sgn(\rho_n(X,Y))x_iy_i + y_i^2 \\
&\qquad - x_i^2 + 2sgn(\rho_n(X,Y))x_iy_i - y_i^2)] \\
&= \frac{2x_i^2 + 2y_i^2 - 4 \cdot |\rho_n(X,Y)| \cdot sgn(\rho_n(X,Y)) \cdot x_i \cdot y_i}{2(1-\rho_n(X,Y)^2)} \\
&= \frac{x_i^2 + y_i^2 - 2 \cdot x_i \cdot y_i \cdot \rho_n(X,Y)}{1-\rho_n(X,Y)^2}.
\end{aligned}$$

Dieses Konzept der Zerlegung der Gesamtstreuung in orthogonale Streukomponenten wird in Kapitel 6 auf Dimensionen $m \geq 3$ zur sogenannten „Hauptkomponentenanalyse“ erweitert.

Als „Ausreißer“ werden nun wie eingangs dieses Kapitels beschrieben Datenpunkte mit einer großen Mahalanobisdistanz zu $(\bar{x}_n, \bar{y}_n)$ deklariert. In konkreten Anwendungen besteht dann die grundlegende Fragestellung darin, ab wann $\Delta^2(z_i, \bar{z}_n)$ als so groß angesehen werden soll, dass dem zugehörigen Punkt das Attribut „Ausreißer“ zugeordnet wird. Dazu gibt es eine ganze Reihe möglicher Ansätze. Der folgende Abschnitt beschreibt die Methoden, die in der vorliegenden Arbeit Anwendung gefunden haben.

3.3 Anwendung

3.3.1 Naiver Ansatz

Der erste, naive Ansatz zur Festlegung einer kritischen Schranke Δ_{krit} für den Wert der Mahalanobisdistanz bestand darin, besonders markante Streubilder wie etwa in Abbildung 3.1 oder Abbildung 3.2 wiedergegeben zu betrachten. Hier ist es wie bereits beschrieben in jedem Fall wünschenswert, dass die vorhandenen Ausreißer automatisch von Softwareseite erkannt werden, da durch ihr Vorhandensein weitergehende Berechnungen wie die der empirischen Korrelation empfindlich gestört werden. Aufgrund dessen wurden mehrere Beispiele, in denen dies der Fall ist, analysiert und festgestellt, wie Δ_{krit} jeweils zu wählen wäre, um diese Ausreißer ausfindig zu machen. Es ergab sich, dass in der Regel bei diesen augenscheinlichen Ausreißern eine Mahalanobisdistanz von $\Delta \geq 2$, häufig sogar von $\Delta \geq 2.5$ vorlag.
Von diesen empirischen Ergebnissen ausgehend wurde daraufhin der konstante Wert $\Delta_{krit} = 2$ (für fast sicheres Erkennen) bzw. $\Delta_{krit} = 2.5$ als kritische Mahalanobisdistanz vereinbart. Der Vorteil dieser Herangehensweise besteht darin, dass sie völlig frei von Verteilungsannahmen ist und daher zu ihrer Anwendbarkeit nichts über die Herkunft des zu Grunde liegenden Datenmaterials bekannt sein muss.

Für zuerst ausschließlich betrachtete „moderate“ Stichprobenumfänge wie z.B. $n = 30$ oder $n = 50$

lieferte dieses Vorgehen auch durchaus brauchbare Ergebnisse. Im Falle von Messreihen mit erheblich größerem Stichprobenumfang zeigte sich allerdings der große Nachteil des Verfahrens: Mit wachsendem n wurde der Anteil der so bestimmten Ausreißer immer größer und führte dazu, dass die erzeugte Information für den Analysten nicht mehr überschaubar blieb. Dies lässt sich damit begründen, dass bei kleinen Stichprobenumfängen die Kenngrößen arithmetisches Mittel und empirische Varianz durch einzelne Ausreißer stark beeinflusst werden. Dies führt dazu, dass die Mahalanobisdistanz dieser Ausreißer zu dem berechneten Datenzentrum nur eine bestimmte Maximalgröße annehmen kann und der kritische Wert Δ_{krit} daher nicht zu groß gewählt werden darf. Liegt hingegen ein großer Stichprobenumfang n vor, so nimmt der Einfluss einzelner Beobachtungen auf die Kenngrößen ab und deswegen wird bei Beibehaltung der ursprünglichen kritischen Schranke ein wesentlich größerer Anteil der Beobachtungen als Ausreißer deklariert. Nachteilhaft daran ist vor allen Dingen, dass viele der so deklarierten Beobachtungen gar keine Ausreißer im ursprünglichen Sinne sind. Es wäre also wünschenswert, dass auch Δ_{krit} mit wachsendem n größer wird, damit nicht etwa bei z.B. $n = 1000$ fast 50 % $\triangleq$ 500 Datenpunkte als Ausreißer aufgeführt werden. Dies hätte im Falle der Beibehaltung des „naiven Ansatzes“ jedoch zur Folge, dass für jede Analyse eine neue Konstante gewählt werden müsste und dass diese überdies einen Kausalbezug zu n haben sollte.

Es ist daher also naheliegend, Δ_{krit} nicht als konstant, sondern als Funktion des Stichprobenumfangs n anzusetzen. Im Folgenden werden nun zwei Verfahren beschrieben, die einen solchen Bezug zwischen n und Δ_{krit} herstellen.

3.3.2 Verwendung asymptotischer Verteilungsquantile

Wie in Kapitel 3.2 hergeleitet misst die Mahalanobisdistanz den euklidischen Abstand standardisierter Zufallsgrößen in einem speziellen Koordinatensystem. Bezeichne also $C = (C_1, C_2)$ die zugrunde liegenden Hauptkomponenten und $\mu = (\mu_1, \mu_2)$ den zweidimensionalen Erwartungswert der zu Grunde liegenden Verteilung, so gilt:

$$\Delta^2(C, \mu) = \frac{(C_1 - \mu_1)^2}{Var(C_1)} + \frac{(C_2 - \mu_2)^2}{Var(C_2)}.$$

Bei diesem Ausdruck handelt es sich um eine Summe von Quadraten zweier standardisierter Zufallsgrößen. Unterliegen diese Zufallsgrößen nun einer Normalverteilung, so ist die Verteilung von $\Delta^2(C, \mu)$ bekannt als χ^2-Verteilung mit zwei Freiheitsgraden oder kurz χ^2_2-Verteilung. Im Falle der entsprechenden empirischen Größen gilt dies zwar nur asymptotisch, aber es ist dennoch möglich, Quantile der χ^2_2-Verteilung für Δ_{krit} zu verwenden, wenn man eine normalverteilte Grundgesamtheit unterstellt.
Praktisch kann damit z.B. das oben angedeutete Ziel, den mittleren Anteil an Ausreißern zu kontrollieren, verfolgt werden. Die Forderung lautet in mathematischer Notation

$$\mathbb{P}(\Delta > \Delta_{krit}) = \alpha(n) \stackrel{!}{=} \frac{\mu(n)}{n}$$

mit $\mu(n)$ als gewünschter Anzahl an Ausreißern bei Stichprobengröße n und kann durch die Wahl von $\Delta^2_{krit}(n) = \chi^2_{2;1-\alpha(n)}$ realisiert werden. $\chi^2_{2;1-\alpha(n)}$ bezeichnet hierbei das $(1 - \alpha(n))$-Quantil der χ^2-Verteilung mit zwei Freiheitsgraden.

In dem konkreten Anwendungsbeispiel der Analyse von Daten der pharmazeutischen Forschung wurde $\mu(n) = \sqrt{n/10}$ festgelegt, d.h. aus 1000 Datenpunkten sollten im Mittel etwa 10 Ausreißer herausgefunden werden. Es sei hier aber noch einmal darauf hingewiesen, dass in die Berechnung die Normalverteilungsannahme eingeht und die für $\Delta^2(z_i, \bar{z}_n)$ verwendete Verteilung nur asymptotisch gilt.

3.3.3 Erkennen von „echten“ Ausreißern

Ein ambitionierteres Vorgehen besteht darin, dass der Forderung nachgegangen wird, nur „echte“ Ausreißer zu erkennen. Hier wird also nicht die Anzahl der sich ergebenden Ausreißer zu steuern versucht, sondern es wird eine Signifikanzschranke für die Verlässlichkeit des Urteils „Datenpunkt (x_i, y_i) ist Ausreißer“ gesucht. Präziser formuliert lautet die Forderung, dass mit Wahrscheinlichkeit $(1 - \alpha)$ keine Ausreißer erkannt werden sollen, wenn keine „echten“ Ausreißer vorliegen. Um dies bei gegebenem Signifikanzniveau α zu gewährleisten, wählt man Δ_{krit} derart, dass $\mathbb{P}(\Delta_{max} > \Delta_{krit}) \stackrel{!}{=} \alpha$ erfüllt ist. Hierbei bezeichnet Δ_{max} die größte in der Stichprobe auftretende Mahalanobisdistanz. Liegt also eine homogene Stichprobe ohne „echte“ Ausreißer vor, so wird der Messwert mit der größten Mahalanobisdistanz (welche ja nicht zu dem Urteil „Ausreißer“ führen sollte !) auch nur in $\alpha \cdot 100\%$ der Fälle als Ausreißer deklariert.

Schwierigkeiten bereitet hierbei die konkrete Berechnung von $\Delta_{krit}(n, \alpha)$ bei gegebenem Stichprobenumfang n und vorgegebenem Signifikanzniveau α. In [4] ist diesem Problem nachgegangen worden und es finden sich auf den Seiten 516f. Tabellen für $\Delta_{krit}(n, \alpha)$ mit n im Bereich von 3 bis 1000 und $\alpha = 0.01$ sowie $\alpha = 0.05$. Diese Tabellenwerte wurden in der erfolgten Implementierung übernommen und für $1000 < n \leq 4000$ wurde die Tabelle durch Simulationsrechnungen erweitert. Ab $n > 4000$ wurde dann zur asymptotischen Verteilung, die sich als χ^2_2-Verteilung zur n-ten Potenz ergibt, übergegangen (siehe hierzu auch Abschnitt 6.5.2 mit $p = 2$).

3.3.4 Mehrstufiges Vorgehen

Wie bereits in den einleitenden Bemerkungen zu diesem Kapitel beschrieben, kann sich das Vorhandensein von Ausreißern verfälschend auf die berechneten Werte von $\rho_n(X, Y)$ auswirken. Des Weiteren kann es unter Umständen dazu kommen, dass Ausreißer mit sehr großen Mahalanobisdistanzen solche Ausreißer überdecken, für die Δ zwar im Verhältnis zu den erkannten Ausreißern klein ist, die aber dennoch erkannt werden sollten. Als Beispiel dafür sei der folgende scatter plot von VAR373 und VAR374 angeführt:

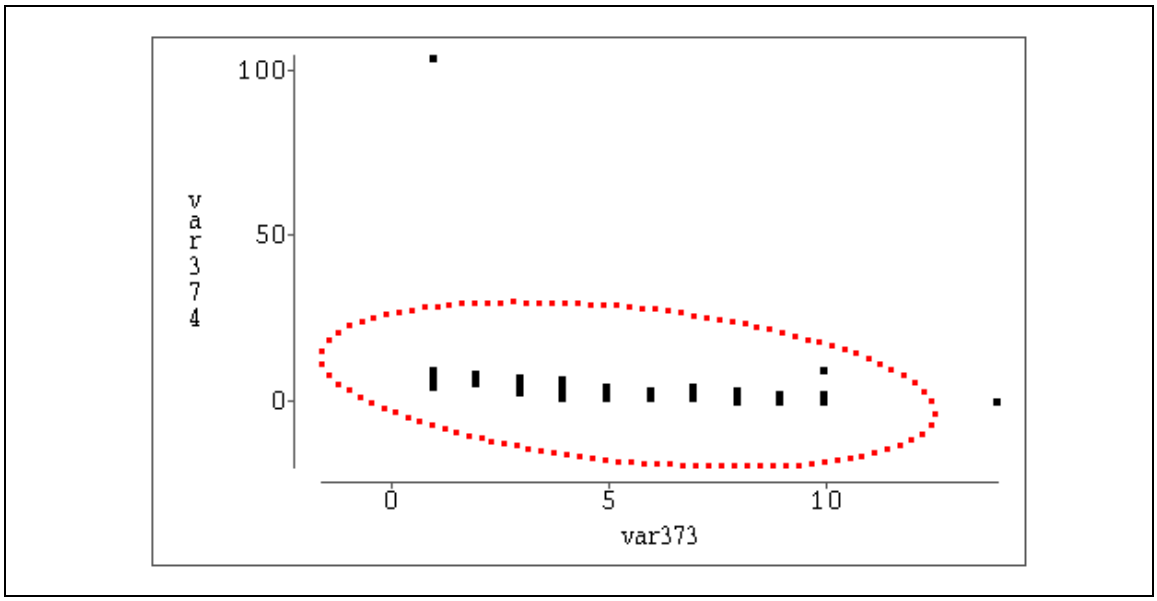

Abbildung 3.5: Variable 373 gegen Variable 374 (alle 109 Beobachtungen)

Es sind eindeutig zwei Ausreißer auszumachen, die außerhalb der eingezeichneten Konfidenzellipse zur Mahalanobisdistanz 2.5 liegen. Der Rest des Datenmaterials weist jedoch optisch keine

weiteren Ausreißer auf. Nimmt man die zwei erkannten Datenpunkte nun von der Berechnung aus (eliminiert man die zwei erkannten Ausreißer), so ergibt sich folgendes Streubild der verbleibenden 107 Datenpaare:

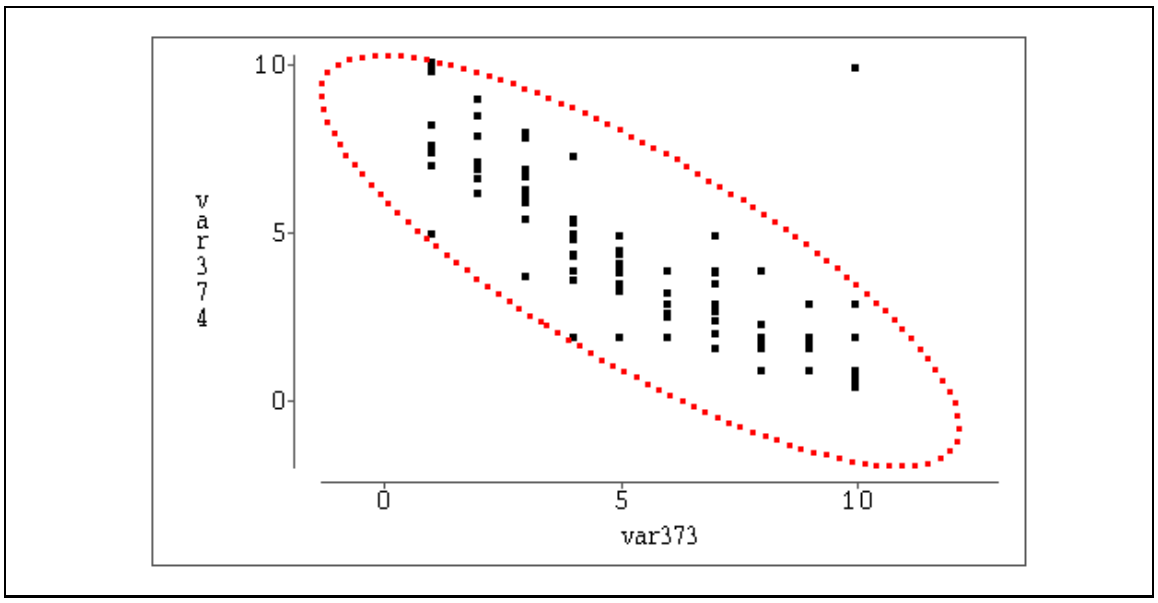

Abbildung 3.6: Variable 373 gegen Variable 374 (restliche 107 Beobachtungen)

Hier zeigt sich nun (aufgrund der veränderten Skalierung des plots klar zu erkennen) ein weiterer Ausreißer, für den beide Variablen etwa den Wert 10 annehmen. Dieser Ausreißer wäre unerkannt geblieben, falls die Ausreißererkennung für diese Variablenkombination nur einmalig durchgeführt worden wäre.

Deswegen wurde in der vorliegenden Arbeit ein iteratives Vorgehen der Ausreißerdetektion und -elimination durchgeführt. Nachdem auf der Basis aller vorliegender Datenpaare $z_i = (x_i, y_i), i = 1, \ldots, n$ die empirische Korrelation $\rho_n(X, Y)$ berechnet wurde, wurden die Mahalanobisdistanzen $\Delta(z_i, \bar{z}_n)$ dieser Messpunkte vom Datenzentrum bestimmt und diejenigen Datenpunkte $z_i^* = (x_i^*, y_i^*)$, bei welchen $\Delta(z_i^*, \bar{z}_n)$ eine signifikante Größe hatte, gekennzeichnet. Danach geschah eine Neuberechnung der empirischen Korrelation, wobei die Paare (x_i^*, y_i^*) nicht mehr berücksichtigt wurden. Dies kann als eine „Robustifizierung“ der Berechnung von $\rho_n(X, Y)$ angesehen werden. Eine Möglichkeit der Erweiterung der erstellten Software besteht darin, von vorne herein robuste Schätzungen der Korrelationsmatrizen zu verwenden.

Ferner wurden pro Iteration die kritischen Werte für Δ angepasst und die Ausreißerbestimmung mit dem neuen Δ_{krit} fortgesetzt, um gegebenenfalls noch unerkannte Ausreißer ermitteln zu können.

3.4 Ergebnisse

In der folgenden Tabelle sollen die Ergebnisse der drei verwendeten Methoden anhand eines Beispiel-Datensatzes quantitativ gegenüber gestellt werden. Dieser Beispiel-Datensatz verfügt über 1017 Merkmale mit 11363 Beobachtungen.

	Ausreißer	davon 4 häufigste	Korrelationen ≥ 0.8
$\Delta_{krit} = 2.0$	≈ 600.000	5 %	≈ 20.000
$\Delta_{krit} = 2.5$	≈ 180.000	5 %	≈ 11.000
$\Delta_{krit}(n) = \chi^2_{2;1-\frac{1}{\sqrt{10n}}}$	≈ 180.000	9 %	≈ 17.500
„Echt“ mit $\alpha = 0.05$	≈ 70.000	11 %	≈ 17.000
„Echt“ mit $\alpha = 0.01$	≈ 40.000	16 %	≈ 16.000

Diese Ergebnisse illustrieren die Vorteile der dritten durchgeführten Methode. Die absolute Anzahl an Ausreißern wird drastisch reduziert und dabei wird augenscheinlich auch der Effekt erzielt, dass nur noch wirklich auffällige Beobachtungen in Erscheinung treten. Dies wird daran erkennbar, dass der Prozentsatz der vier Beobachtungen, die in allen Variablenkombinationen insgesamt am häufigsten als Ausreißer auftreten und von daher wohl in der Tat aus dem Rahmen fallen, im Falle der Anwendung der Methoden zur Erkennung „echter“ Ausreißer stark ansteigt; es erfolgt also eine Konzentration auf die auffälligsten Datenpunkte.
Ferner wird der Wert von ρ_n und damit die Anzahl der als signifikant betrachteten Korrelationen durch die Wahl der so gearteten Verfahren kaum beeinflusst. Es gibt allerdings einzelne Fälle, in denen durch diese Art der Ausreißerbehandlung signifikante Korrelationen verloren gehen.

In Kapitel 5 wird mit der Rangkorrelation nach Spearman ein Konzept vorgestellt, mit dem dies verhindert werden kann.

Kapitel 4

Transformation von Variablen

4.1 Motivation

Wie bereits in Kapitel 2 erläutert, ist der (empirische) Korrelationskoeffizient nach Pearson nur ein Maß für lineare Abhängigkeiten zwischen Zufallsvariablen bzw. Messreihen.
In Beispiel 2.3 etwa wurde der Zusammenhang zwischen dem Einsatz eines Schädlingsbekämpfungsmittels und dem damit erzielten Ernteertrag als quadratisch angenommen. Über diese nichtlineare Abhängigkeit gibt der berechnete Wert $\rho_{10}(X, Y) = -0.02490$ keinerlei Auskunft.

Um diese Abhängigkeit dennoch aufzudecken bzw. die getroffene Annahme einer quadratischen Beziehung zu überprüfen, wurde ein Regressionspolynom der Ordnung 2 nach dem Modell $Y = a_2 \cdot X^2 + a_1 \cdot X + a_0 + \varepsilon$ an das vorhandene Datenmaterial angepasst und dann die beobachteten Realisierungen des Ernteertrags durch die Werte der inversen Regressionsfunktion, ausgewertet an den jeweiligen Messpunkten, ersetzt. Das Ergebnis zeigt die folgende Abbildung:

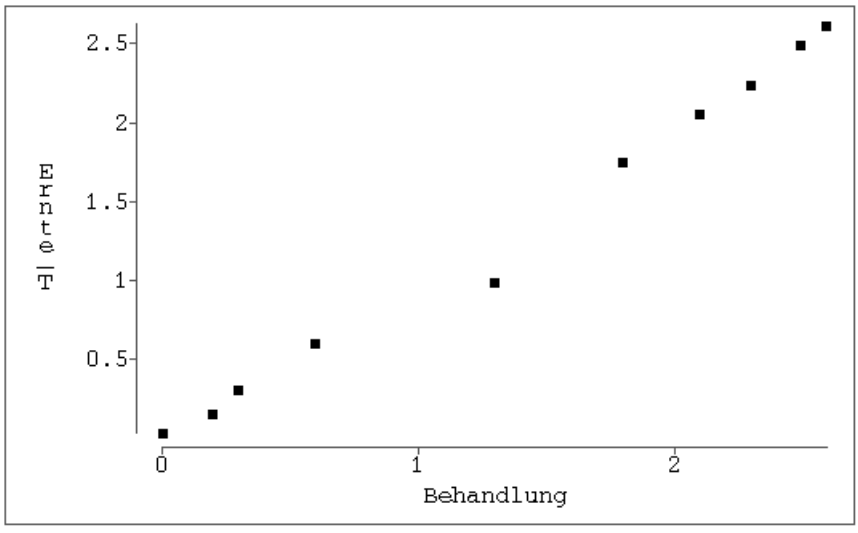

Abbildung 4.1: Menge Schädlingsbekämpfungsmittel gegen transformierten Ernteertrag

Berechnet man nun die empirische Produktmomentkorrelation zwischen den beiden Merkmalen auf der Basis der transformierten Werte für den Ernteertrag Y, so ergibt sich

$$\rho^*_{10}(X, Y) = 0.9952 \approx 1.$$

Es ist also gelungen, durch **Transformation** der Werte von Y die Abhängigkeit zwischen den Messreihen zu linearisieren und damit von dem (empirischen) Korrelationskoeffizienten nach Pearson messbar zu machen.

Allgemein lassen sich nicht-lineare Abhängigkeiten in lineare Abhängigkeiten überführen, indem eine oder gegebenenfalls auch beide zu Grunde liegenden Zufallsgrößen geeigneter Transformationen unterzogen wird bzw. werden. In der Regel ist es jedoch keineswegs offensichtlich, wie diese linearisierenden Transformationen zu wählen sind, denn der Typ der vorliegenden Beziehung geht maßgeblich in die Wahl der Transformation ein.

Es sei an dieser Stelle noch darauf hingewiesen, dass das Konzept der Variablentransformation auch anderen Absichten dienen kann. Wie unter anderem in [5] ausführlich beschrieben, kann einer eindimensionalen Verteilung durch geeignete Transformation die Schiefe genommen oder durch Maßstabsänderung (z.B. Logarithmierung) eine bessere Auflösung bzw. Darstellbarkeit gemessener Daten erreicht werden. Da dies aber nicht zu den Hauptzielen der vorliegenden Arbeit zählt, soll hierauf an dieser Stelle nicht näher eingegangen werden.

In den im nachfolgenden Abschnitt beschriebenen Fallbeispielen aus der pharmazeutischen Forschung war eine Identifikation des zu wählenden Transformationstyps aufgrund von Expertenwissen bzw. physikalischen Grundzusammenhängen möglich und es konnte erreicht werden, dass signifikant große Korrelationen nach getätigten Transformationen berechnet wurden, falls Abhängigkeiten des angenommenen Typs vorlagen.

4.2 Anwendung

Neben der offensichtlichen Anwendung der Transformation von Volumendaten durch Übergang zur dritten Wurzel ist auch im Falle von Konzentrationsdaten das Konzept der Transformation eingesetzt worden, um die Abhängigkeit von Variablen zu linearisieren. Hierzu sind Vorüberlegungen aus dem Bereich der Reaktionskinetik erforderlich. Hier werden nur die notwendigen Ergebnisse wiedergegeben; die zu Grunde liegenden biologischen Zusammenhänge werden genauer z.B. in [6] sowie [7] erläutert.

Betrachtet wird die reversible Bindung von Substanzen an Rezeptormoleküle. Wird einem Organismus eine chemische Substanz verabreicht, so kann es dazu kommen, dass sich diese an sogenannte „Rezeptormoleküle“ in dem Organismus bindet und die beiden Moleküle reversibel zu einem neuen Molekül reagieren. Es ist nun möglich, die Konzentrationen der an diesen Reaktionsvorgängen beteiligten Ausgangsmolekülen mit der der resultierenden Moleküle mathematisch in Beziehung zu setzen. Dies geschieht im Wesentlichen durch die Einführung der sogenannten „Reaktionskonstanten“ k_1 und k_{-1}. Das in diesem Fall benutzte Modell entstammt dem Massenwirkungsgesetz. Es besagt, dass die Reaktionsvorgänge proportional von den Ausgangskonzentrationen und die inversen Reaktionsvorgänge (Substanz und Rezeptor lösen sich wieder voneinander) proportional von der Konzentration reagierter (verbundener) Moleküle abhängen, wobei die Proportionalitätskonstanten gerade die oben angeführten Zahlen k_1 bzw. k_{-1} sind.

Bezeichnet also S die verabreichte Substanz und R den zugehörigen Rezeptor, so kann der beschriebene Reaktionsvorgang wie folgt symbolisiert werden:

$$a \cdot S + R \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} SR$$

Mit den folgenden Bezeichnungen:

$C_R(t)$	Konzentration freier Rezeptoren zum Zeitpunkt t
$C_S(t)$	Konzentration freier Substanz-Moleküle zum Zeitpunkt t
$C_{SR}(t)$	Konzentration blockierter Rezeptoren (verbundener Moleküle) zum Zeitpunkt t
$C_{R,ges} = C_R(t) + C_{SR}(t)$	gesamte Rezeptorkonzentration

ergibt sich unter Verwendung des Massenwirkungsgesetzes für die zeitliche Änderung von $C_{SR}(t)$ die folgende Differentialgleichung:

$$\dot{C}_{SR}(t) = k_1 \cdot C_S(t)^a \cdot C_R(t) - k_{-1} C_{SR}(t).$$

Geht man nun davon aus, dass sich diese Differentialgleichung für $t \to \infty$ stabilisiert und in einen stationären Endzustand mit $\dot{C}_{SR}(t) \equiv 0$ übergeht, so gilt in diesem eingeschwungenen Zustand:

$$\begin{aligned} 0 &= k_1 C_S^a \cdot C_R - k_{-1} \cdot C_{SR} \\ &= k_1 C_S^a \cdot (C_{R,ges} - C_{SR}) - k_{-1} C_{SR} \end{aligned}$$

Das Argument $t \to \infty$ wurde hier aus Gründen der Notationsvereinfachung bei den Konzentrationswerten fortgelassen, da diese im eingeschwungenen Zustand konstant sind. Führt man nun noch die Größe

$$\tilde{C}_{SR} := \frac{C_{SR}}{C_{R,ges}} \stackrel{\wedge}{=} \text{Anteil blockierter Rezeptoren}$$

ein, so vereinfacht sich die stationäre Gleichung weiter, indem man beide Seiten durch $C_{R,ges}$ dividiert:

$$\begin{aligned} & k_1 C_S^a \cdot (1 - \tilde{C}_{SR}) - k_{-1} \cdot \tilde{C}_{SR} &&= 0 \\ \iff\ & \tilde{C}_{SR} \cdot (k_{-1} + k_1 C_S^a) &&= k_1 C_S^a \\ \iff\ & \tilde{C}_{SR} = \tfrac{k_1 C_S^a}{k_{-1} + k_1 C_S^a} &&= \tfrac{C_S^a}{k + C_S^a} \end{aligned}$$

mit $\tilde{k} := \frac{k_{-1}}{k_1}$. Hieraus ergibt sich, dass $\tilde{k}^{\frac{1}{a}}$ genau die Substanzkonzentration ist, bei der die Hälfte der Rezeptoren blockiert wird.

In den Laborexperimenten der pharmazeutischen Forschung werden nun häufig folgende Messwerte ermittelt:

X	Anteil der Rezeptoren, die bei gegebener Konzentration C_0 der Substanz blockiert wird.
Y	Substanzkonzentration, bei der die Hälfte der Rezeptoren blockiert wird (**k_i-Wert**)

Aus der hergeleiteten Beziehung zwischen $\tilde{C}_{SR}$ und dem zugehörigen k_i-Wert kann nun ein Modellierungsansatz für den Zusammenhang von X und Y abgeleitet werden:

$$\begin{aligned} X &= \frac{C_0^a}{Y^a + C_0^a} \\ X \cdot (Y^a + C_0^a) &= C_0^a \\ X \cdot Y^a &= C_0^a (1 - X) \\ Y^a &= C_0^a \cdot \frac{1 - X}{X} \\ a \cdot \ln(Y) &= a \cdot \ln(C_0) + \ln(\frac{1 - X}{X}) \\ \text{logit}(X) &= a \cdot (\ln(Y) - \ln(C_0)) \end{aligned}$$

Um eine Linearisierung der Abhängigkeit zwischen X und Y zu erreichen, wurde also in den entsprechenden Fällen die Variable Y einer Transformation mit der natürlichen Logarithmusfunktion und die Variable X der sogenannten Logit-Transformation

$$\text{logit}(X) = \ln\left(\frac{1-X}{X}\right)$$

unterzogen.

Kapitel 5

Rangkorrelation

5.1 Motivation

Im vorherigen Kapitel wurde die Möglichkeit vorgestellt, mit Hilfe geeigneter Transformationen eine nicht-lineare Abhängigkeit zwischen zwei Zufallsgrößen X und Y zu linearisieren und somit von der Produktmomentkorrelation messbar zu machen. Das Hauptproblem bei dieser Vorgehensweise besteht jedoch darin, dass die Wahl der zu tätigen Transformation(en) von der zu Grunde liegenden Beziehung zwischen X und Y abhängt. Häufig ist jedoch (wie auch z.T. in der vorliegenden Aufgabe der Datenanalyse im Rahmen der pharmazeutischen Forschung) Detailwissen über die Herkunft des Datenmaterials nicht bekannt. Damit kann kein Modell für den Zusammenhang von X und Y hergeleitet werden und somit bleiben evtl. linearisierende Transformationen unentdeckt.

Es wäre also wünschenswert, eine „universelle“ Transformation zur Verfügung zu haben, die möglichst viele Abhängigkeiten linearisiert. Im Bereich der nichtparametrischen Statistik hat sich hier das Konzept der **Rangkorrelation nach Spearman** etabliert, mit welchem sich zumindest alle monotonen Abhängigkeiten identifizieren lassen. Das Vorgehen besteht darin, dass nicht der Pearson'sche Korrelationskoeffizient der Originaldaten, sondern die Produktmomentkorrelation der zugehörigen Positionen der Messwerte in der jeweiligen geordneten Stichprobe (der Ränge der Beobachtungen zu X und Y) berechnet wird. Damit spielt also der absolute Abstand der gemessenen Werte keine Rolle, sondern lediglich die Ordnung der Messwerte geht in die Berechnung ein. Aufgrund dessen liefert die Spearman'sche Rangkorrelation für alle Variablenkombinationen (X, Y), in denen X über eine monotone (ggfs. auch nicht-lineare) Transformation in Y überführt werden kann, betragsmäßig den Wert 1. Um dies zu illustrieren, wurden im Folgenden drei zweidimensionale Stichproben simuliert.

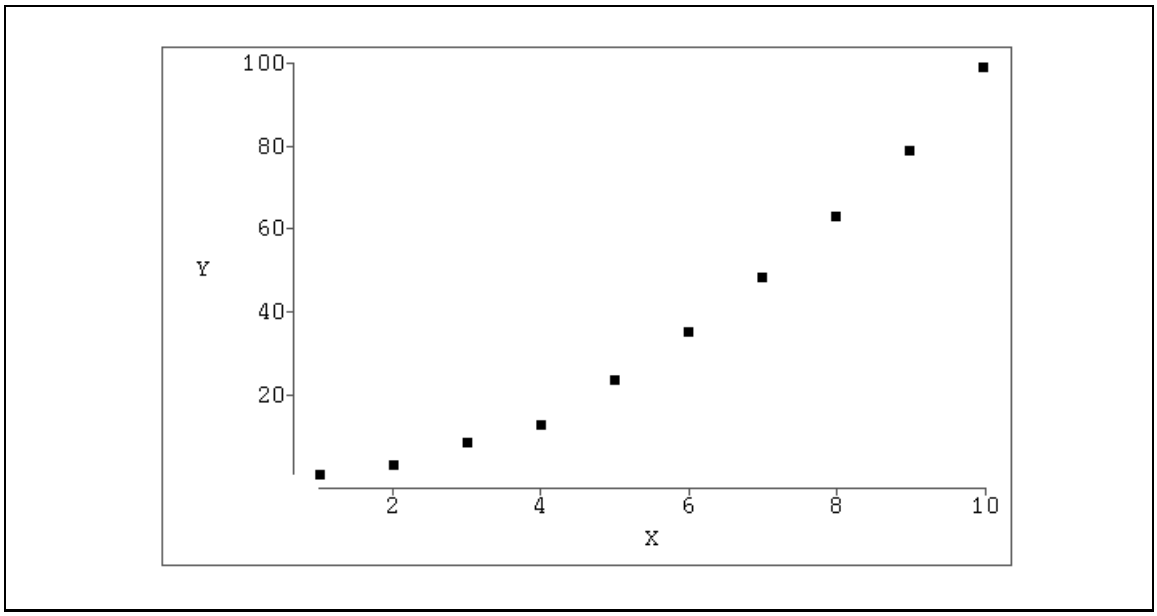

Abbildung 5.1: $Y = X^2 + \varepsilon$

Abbildung 5.1 liegt das Modell eines quadratischen Zusammenhangs von X und Y mit einem standardnormalverteilten Fehlerterm zu Grunde. Hier fällt in dem betrachteten Intervall $[1; 10]$ für X die Nichtlinearität der gegebenen Abhängigkeit bei der Berechnung von $\rho_n(X, Y)$ nicht besonders ins Gewicht, es ergibt sich der hohe Wert $\rho_n(X, Y) = 0.9723$.

Geht man nun zu einem kubischen Zusammenhang $Y = X^3 + \varepsilon, \varepsilon \sim \mathcal{N}(0,1)$ über, so ergibt sich das folgende Streubild

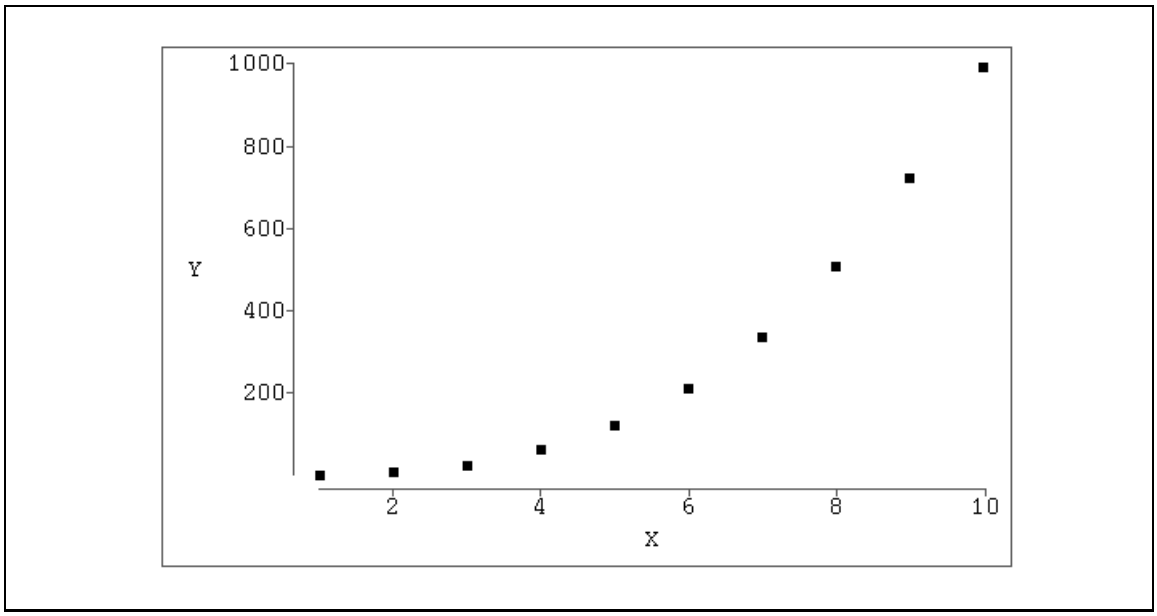

Abbildung 5.2: $Y = X^3 + \varepsilon$

und aufgrund der stärkeren Abweichung vom linearen Zusammenhang sinkt der Wert der Produktmomentkorrelation auf $\rho_n(X, Y) = 0.9280$.

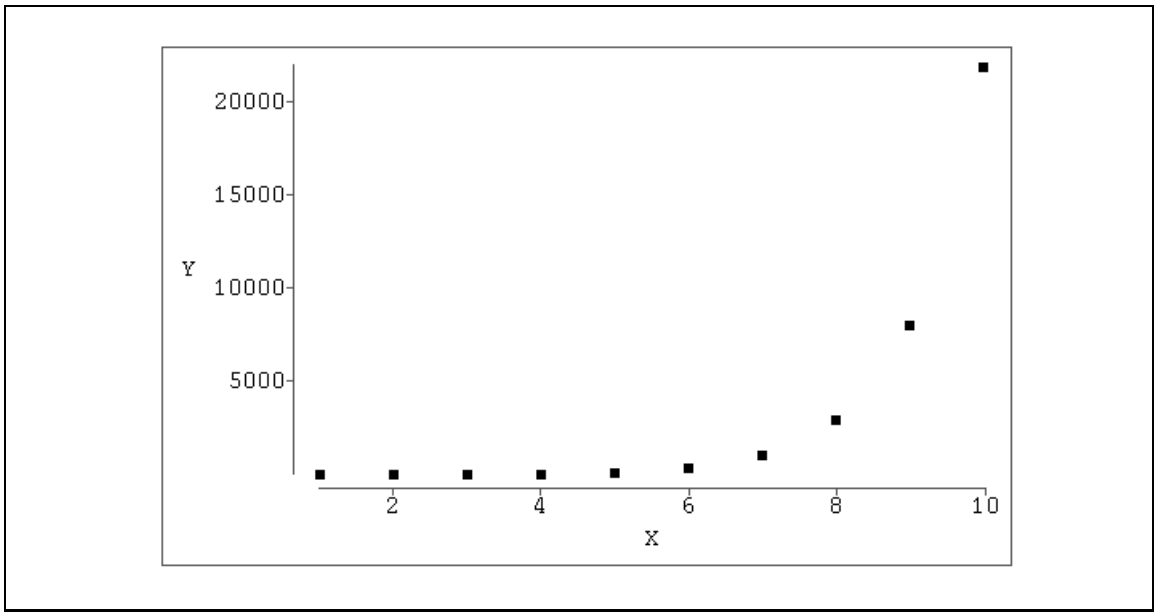

Abbildung 5.3: $Y = \exp(X) + \varepsilon$

In Abbildung 5.3 ist nun schließlich die Beziehung $Y = \exp(X) + \varepsilon, \varepsilon \sim \mathbb{U}(0,1)$ modelliert worden. Diese Abhängigkeit ist von $\rho_n(X, Y)$ nur noch schlecht zu erfassen, der berechnete Wert ergibt sich zu $\rho_n(X, Y) = 0.7169$.

Sicherlich wäre es jedoch von Interesse, in allen drei Fällen auf die Abhängigkeit von X und Y aufmerksam zu werden. Setzt man nun wie oben beschrieben anstelle der tatsächlichen Simulationswerte die zugehörigen Ränge ein, ergibt sich in allen drei Fällen derselbe scatter plot (siehe Abbildung 5.4):

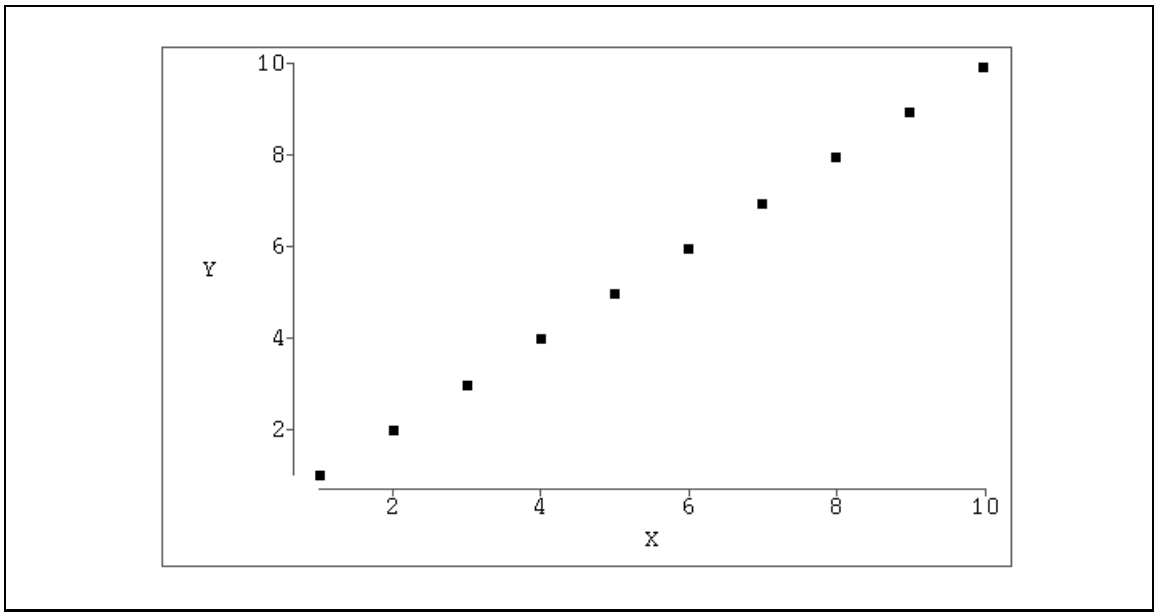

Abbildung 5.4: Streubild der Ränge zu Abbildungen 5.1 bis 5.3

Die „Punktewolke“ entspricht der ersten Winkelhalbierenden, da mit wachsenden Werten für X auch in Y-Richtung immer größere Messwerte beobachtet wurden. Dabei spielt es keine Rolle, in

welcher Form das Wachstum in Y-Richtung mit dem in X-Richtung gekoppelt ist. Der Spearmansche Rangkorrelationskoeffizient, hier mit $\rho_n^S(X,Y)$ bezeichnet, errechnet sich damit in allen drei Modellbeispielen zu $\rho_n^S(X,Y) = 1$.

5.2 Theorie

Bemerkung 5.2.1
Um eine Vereinfachung der Notation zu erreichen, wird in diesem Abschnitt die empirische Varianz $Var_n(X)$ einer Zufallsgröße X auf der Basis von n Realisierungen $(x_1, \ldots, x_n)$ definiert als

$$Var_n(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \text{ mit } \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Es wird also - abweichend von der Festsetzung in Kapitel 2 - der Normierungsfaktor $\frac{1}{n}$ anstelle von $\frac{1}{n-1}$ verwendet.

Definition 5.2.1 (Rang)
Der Rang r_k einer Beobachtung x_k aus einer Stichprobe $(x_1, \ldots, x_n)$ vom Umfang n mit $x_i \neq x_j \ \forall i \neq j$ ist definiert als die Position von x_k in der zugehörigen geordneten Stichprobe, also

$$r_k = \sum_{j=1}^{n} I_{\{x_j \leq x_k\}}.$$

Satz 5.2.1 (Arithmetisches Mittel und empirische Varianz eines Rangvektors)
Sei $R = (r_1, \ldots, r_n)$ der Vektor der Ränge der Beobachtungen $(x_1, \ldots, x_n)$ einer Stichprobe vom Umfang n mit $x_i \neq x_j \ \forall i \neq j$. Dann gilt:

$$\begin{aligned} \bar{r}_n &= \frac{n+1}{2} \\ Var_n(R) &= \frac{n^2-1}{12} \end{aligned}$$

Beweis 5.2.1

$$\bar{r}_n = \ \tfrac{1}{n}\textstyle\sum_{i=1}^{n} i = \tfrac{1}{n}\tfrac{n\cdot(n+1)}{2} = \tfrac{n+1}{2}.$$

$$\begin{aligned} Var_n(R) &= \frac{1}{n}\sum_{i=1}^{n}(i - \bar{r}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(i - \frac{n+1}{2})^2 \\ &= \frac{1}{n}[\sum_{i=1}^{n} i^2 - \sum_{i=1}^{n}(n+1)\cdot i + \sum_{i=1}^{n}(\frac{n+1}{2})^2] \\ &= \frac{1}{n}[\frac{n\cdot(n+1)\cdot(2n+1)}{6} - \frac{n\cdot(n+1)^2}{2} + \frac{n\cdot(n+1)^2}{4}] \\ &= \frac{1}{n}[\frac{n\cdot(n+1)\cdot(2n+1)}{6} - \frac{n\cdot(n+1)^2}{4}] \\ &= \frac{1}{n}[\frac{2n\cdot(n+1)\cdot(2n+1) - 3n\cdot(n+1)^2}{12}] \\ &= \frac{1}{n}[\frac{4n^3+2n^2+4n^2+2n-3n^3-6n^2-3n}{12}] \\ &= \frac{1}{n}(\frac{n\cdot(n^2-1)}{12}) = \frac{n^2-1}{12}. \end{aligned}$$

Die Definition des Ranges in Definition 5.2.1 setzt voraus, dass alle Messwerte unterschiedlich sind. Da dies jedoch nicht immer der Fall ist, sind Überlegungen nötig, wie man mehrfache („verbundene“) Werte (englisch: „ties“) behandelt. Zur Vereinfachung der Darstellung nehmen wir an, dass für die der Größe nach geordneten Beobachtungen $x_{i:n}$ der betrachteten Stichprobe $x_{1:n} = x_{2:n} = \ldots = x_{d_1:n}$ gilt. Dann ist es sicherlich nicht sinnvoll, alle Ränge von 1 bis d_1 zu vergeben, denn damit würde die Information der Gleichheit von $x_{1:n}, \ldots, x_{d_1:n}$ verlorengehen. Ebenso wenig sollte der Rang 1 d_1-mal gewählt werden, denn dies würde einen sehr großen Abstand der Werte $x_{1:n}, \ldots, x_{d_1:n}$ von $x_{d_1+1:n}$ erzeugen, der in Wirklichkeit nicht gegeben sein muss. Ein analoges Argument gilt selbstverständlich für die Festsetzung, dass alle Werte $x_{1:n}, \ldots, x_{d_1:n}$ den Rang d_1 erhalten.

Vielmehr wird man dazu übergehen, den Beobachtungen $x_{1:n}, \ldots, x_{d_1:n}$ den mittleren Rang $\frac{d_1+1}{2}$ zuzuweisen. Allgemein geschieht die Zuordnung der Ränge im Falle des Vorhandenseins von Ties wie folgt:

Definition 5.2.2 (Mittlerer Rang)
Gegeben sei eine Stichprobe $(x_1, \ldots, x_n)$ vom Umfang n mit p unterschiedlichen Werten v_i, also $|\{x_i : i = 1, \ldots, n\}| = p$. Der Wert $v_i, i = 1, \ldots, p$ mit $v_1 < v_2 < \ldots < v_p$ trete dabei d_i-mal auf. Der mittlere Rang r_k^* einer Beobachtung x_k aus einer solchen Stichprobe ist definiert als:

$$r_k^* = \sum_{j=1}^{i-1} d_j + \frac{1}{2} \cdot (d_i + 1) \;\; \forall k : x_k = v_i; \;\; i = 1, \ldots, p$$

Satz 5.2.2 (Arithmetisches Mittel des Vektors der mittleren Ränge)
Es gilt:

$$\frac{1}{n} \sum_{k=1}^{n} r_k^* = \frac{n+1}{2}.$$

Beweis 5.2.2
Zunächst seien k_1 und k_2 zwei natürliche Zahlen mit $k_2 > k_1$. Dann gilt:

$$\begin{aligned}
(k_2 - k_1) \cdot [k_1 + \frac{1}{2} \cdot (k_2 - k_1 + 1)] &= \frac{(k_2 - k_1) \cdot (2k_1 + k_2 - k_1 + 1)}{2} \\
&= \frac{(k_2 - k_1) \cdot (k_1 + k_2 + 1)}{2} \\
&= \frac{k_1 k_2 + k_2(k_2+1) - k_1(k_1+1) - k_1 k_2}{2} \\
&= \frac{k_2(k_2+1)}{2} - \frac{k_1(k_1+1)}{2} \\
&= \sum_{i=1}^{k_2} i - \sum_{i=1}^{k_1} i \\
&= \sum_{i=k_1+1}^{k_2} i
\end{aligned}$$

Wählen wir nun speziell $k_1 = \sum_{j=1}^{i-1} d_j$ und $k_2 = \sum_{j=1}^{i} d_j$, dann gilt offensichtlich $k_2 - k_1 = d_i$

und damit

$$\begin{aligned}
\Rightarrow \sum_{k=1}^{n} r_k^* &= \sum_{i=1}^{p} \left[d_i \cdot (\sum_{j=1}^{i-1} d_j + \frac{1}{2} \cdot (d_i + 1)) \right] \\
&= \sum_{i=1}^{p} \sum_{l=\sum_{j=1}^{i-1} d_j + 1}^{\sum_{j=1}^{i} d_j} l \\
&= \sum_{i=1}^{n} i = \frac{n \cdot (n+1)}{2}.
\end{aligned}$$

Hieraus folgt, dass das arithmetische Mittel der (mittleren) Ränge einer Stichprobe vom Umfang n konstant, also insbesondere invariant bezüglich des Vorhandenseins beliebig gearteter Ties ist. Für die empirische Varianz gilt diese Aussage hingegen nicht. Es ist zwar möglich, eine geschlossene Formel anzugeben, diese beinhaltet jedoch einen Korrekturterm, in welchen die Informationen über die vorhandenen verbundenen Werte eingehen:

Satz 5.2.3 (Empirische Varianz des Vektors der mittleren Ränge)
Es gilt:

$$\frac{1}{n} \sum_{k=1}^{n} (r_k^* - (\frac{n+1}{2}))^2 = \frac{n^2 - 1}{12} - \frac{\sum_{i=1}^{p} d_i (d_i^2 - 1)}{12n}.$$

Eine Herleitung dieser Formel findet sich unter anderem in [8] auf S. 329f. Hier wird sie im Folgenden durch elementares Nachrechnen bewiesen.

Beweis 5.2.3
Zunächst treffen wir zur Vereinfachung der Schreibweise folgende Festsetzungen:

$$\begin{aligned}
s_i &:= \sum_{j=1}^{i} d_j \\
a_i &:= s_{i-1} + \frac{1}{2} \cdot (d_i + 1).
\end{aligned}$$

$$
\begin{aligned}
\Rightarrow \quad & \frac{1}{n}\sum_{k=1}^{n}(r_k^* - \frac{n+1}{2})^2 \\
= \quad & \frac{1}{n}\sum_{i=1}^{p} d_i \cdot (a_i - \frac{n+1}{2})^2 \\
= \quad & \frac{1}{n}\sum_{i=1}^{p}\sum_{l=s_{i-1}+1}^{s_i} (a_i - \frac{n+1}{2})^2 \\
= \quad & \frac{1}{n}\sum_{i=1}^{p}\sum_{l=s_{i-1}+1}^{s_i} ((l - \frac{n+1}{2}) - (l - a_i))^2 \\
= \quad & \frac{1}{n}\sum_{i=1}^{p}\sum_{l=s_{i-1}+1}^{s_i} \left[(l - \frac{n+1}{2})^2 + (l-a_i)^2 - 2\cdot(l - \frac{n+1}{2})\cdot(l-a_i)\right] \\
= \quad & \frac{1}{n}\sum_{k=1}^{n}(k - \frac{n+1}{2})^2 \\
& + \frac{1}{n}\sum_{i=1}^{p}\sum_{l=s_{i-1}+1}^{s_i} \left[(l-a_i)^2 - 2\cdot(l - a_i + a_i - \frac{n+1}{2})\cdot(l-a_i)\right] \\
= \quad & \frac{n^2-1}{12} + \frac{1}{n}\sum_{i=1}^{p}\sum_{l=s_{i-1}+1}^{s_i} \left[-(l-a_i)^2 - 2\cdot(a_i - \frac{n+1}{2})\cdot(l-a_i)\right] \\
= \quad & \frac{n^2-1}{12} - \frac{1}{n}\sum_{i=1}^{p}\sum_{l=s_{i-1}+1}^{s_i} (l-a_i)^2 \\
& - \frac{2}{n}\sum_{i=1}^{p}\left[(a_i - \frac{n+1}{2})\cdot\sum_{l=s_{i-1}+1}^{s_i}(l-a_i)\right] \\
\underset{l' := l - s_{i-1}}{=} \quad & \frac{n^2-1}{12} - \frac{1}{n}\sum_{i=1}^{p}\sum_{l'=1}^{d_i}(l' - \frac{d_i+1}{2})^2 \\
& - \frac{2}{n}\sum_{i=1}^{p}\left[(a_i - \frac{n+1}{2})\cdot\sum_{l'=1}^{d_i}(l' - \frac{d_i+1}{2})\right] \\
= \quad & \frac{n^2-1}{12} - \frac{1}{n}\sum_{i=1}^{p}\frac{d_i\cdot(d_i^2-1)}{12} \\
& - \frac{2}{n}\sum_{i=1}^{p}\left[(a_i - \frac{n+1}{2})\cdot(\frac{d_i\cdot(d_i+1)}{2} - \frac{d_i\cdot(d_i+1)}{2})\right] \\
= \quad & \frac{n^2-1}{12} - \frac{\sum_{i=1}^{p} d_i\cdot(d_i^2-1)}{12n}.
\end{aligned}
$$

Um nun die gegebene Aufgabe, die Rangkorrelation zweier Stichproben $(x_1, \ldots, x_n)$ sowie $(y_1, \ldots, y_n)$ mit zugehörigen mittleren Rängen $(r_1^*, r_2^*, \ldots, r_n^*)$ bzw. $(s_1^*, s_2^*, \ldots, s_n^*)$ zu berechnen, die überdies jeweils Ties an unterschiedlichen Stellen aufweisen können, kann man sich der Formel für die empirische Varianz der Differenz zweier Zufallsgrößen bedienen:

$$
\begin{aligned}
Var_n(X-Y) &= Var_n(X) + Var_n(Y) - 2Kov_n(X,Y) \\
\Rightarrow Kov_n(X,Y) &= \frac{1}{2}(Var_n(X) + Var_n(Y) - Var_n(X-Y))
\end{aligned}
$$

Die (empirische) Varianz von $Z := R^* - S^*$ ist aber in dem Fall, dass die Realisierungen von R^* und S^* die Vektoren der (mittleren) Ränge sind, einfach zu berechnen, denn es gilt $\bar{r^*}_n = \bar{s^*}_n \equiv \frac{n+1}{2}$ und damit $\bar{z}_n = \bar{r^*}_n - \bar{s^*}_n = 0$.

$$\begin{aligned} \Rightarrow Var_n(Z) &= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z}_n)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} z_i^2 - \bar{z}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(r_i^* - s_i^*)^2 \end{aligned}$$

Setzt man dieses Ergebnis in die Formel für die Kovarianz bezüglich der Rangvektoren ein, so ergibt sich mit $D^* := \sum_{i=1}^{n}(r_i^* - s_i^*)^2$:

$$Kov_n(R^*, S^*) = \frac{1}{2}(Var_n(R^*) + Var_n(S^*) - \frac{2D^*}{n}).$$

und schließlich definiert sich der **Rangkorrelationskoeffizient $\rho_{\mathbf{n}}^{\mathbf{S}}(\mathbf{X}, \mathbf{Y})$ nach Spearman** als Produktmomentkorrelation der zu (X, Y) gehörenden Ranggrößen wie folgt:

Definition 5.2.3 (Rangkorrelationskoeffizient nach Spearman)
Voraussetzungen:

1. *Gegeben sind Realisierungen $(x_1, \ldots, x_n)$ sowie $(y_1, \ldots, y_n)$ zweier Zufallsvariablen X und Y.*
2. *Die Realisierungen von X weisen p unterschiedliche Werte v_i auf, also $|\{x_i : i = 1, \ldots, n\}| = p$. Der Wert $v_i, i = 1, \ldots, p$ mit $v_1 < v_2 < \ldots < v_p$ tritt dabei d_i-mal auf.*
3. *Die Realisierungen von Y weisen q unterschiedliche Werte w_i auf, also $|\{y_i : i = 1, \ldots, n\}| = q$. Der Wert $w_i, i = 1, \ldots, q$ mit $w_1 < w_2 < \ldots < w_q$ tritt dabei e_i-mal auf.*
4. *$R^* = (r_1^*, r_2^*, \ldots, r_n^*)$ bezeichnet den Vektor der mittleren Ränge von X.*
5. *$S^* = (s_1^*, s_2^*, \ldots, s_n^*)$ bezeichnet den Vektor der mittleren Ränge von Y.*

Dann heißt die wie folgt definierte Größe $\rho_n^S(X, Y)$ der
Spearmansche Rangkorrelationskoeffizient von X und Y:

$$\rho_n^S(X, Y) = \rho_n(R^*, S^*) = \frac{\frac{1}{2}(Var_n(R^*) + Var_n(S^*) - \frac{2D^*}{n})}{\sqrt{Var_n(R^*)} \cdot \sqrt{Var_n(S^*)}}.$$

Dabei gilt:

$$\begin{aligned} r_k^* &= \sum_{j=1}^{i-1} d_j + \frac{1}{2} \cdot (d_i + 1) \;\; \forall k : x_k = v_i; \;\; i = 1, \ldots, p \\ s_k^* &= \sum_{j=1}^{i-1} e_j + \frac{1}{2} \cdot (e_i + 1) \;\; \forall k : y_k = w_i; \;\; i = 1, \ldots, q \\ Var_n(R^*) &= \frac{n^2 - 1}{12} - \frac{\sum_{i=1}^{p} d_i(d_i^2 - 1)}{12n} \\ Var_n(S^*) &= \frac{n^2 - 1}{12} - \frac{\sum_{i=1}^{q} e_i(e_i^2 - 1)}{12n} \\ D^* &= \sum_{i=1}^{n}(r_i^* - s_i^*)^2. \end{aligned}$$

Bemerkung 5.2.2
In vielen Lehrbüchern wird direkt eine geschlossene Formel für $\rho_n^S(X,Y)$ angegeben, so z.B. auch in der Beschreibung der NAG Fortran Routine zur Berechnung des Spearmanschen Rangkorrelationskoeffizienten. Diese ist selbstverständlich äquivalent zu der hier hergeleiteten. Die hier gewählte Vorgehensweise, $\rho_n^S(X,Y)$ über die Momente der Ranggrößen einzuführen hat jedoch den Vorteil, dass die Analogie zur Produktmomentkorrelation deutlich wird.

5.3 Anwendung

Neben der Berechnung der Produktmomentkorrelation bietet die im Rahmen dieser Diplomarbeit entstandene Software auch die Möglichkeit, den Spearmanschen Rangkorrelationskoeffizienten für je ein Variablenpaar zu berechnen. Dies bietet sich insbesondere in dem Fall an, dass eine Ausreißererkennung gemäß der in Abschnitt 3.3.3 vorgestellten Methode des Erkennens „echter“ Ausreißer vorgenommen werden soll. Hierbei werden die kritischen Werte für die Mahalanobisdistanz, die zu dem Urteil „Datenpunkt ist Ausreißer“ führen, groß und deswegen reduziert sich die Anzahl der als Ausreißer markierten und damit von der weiteren Analyse ausgenommenen Datenpunkte bei Verwendung des Korrelationskoeffizienten nach Pearson z.T. erheblich.
Dies kann in Einzelfällen dazu führen, dass signifikante Korrelationen unerkannt bleiben bzw. nicht durch Elimination von Ausreißern aufgedeckt werden. Zudem wird im Falle der Verwendung kleiner kritischer Werte für die Mahalanobisdistanz bei einem nichtlinearen Zusammenhang zwischen zwei betrachteten Variablen ein Großteil der Randbereiche der Daten durch die Ausreißereliminaton entfernt und es erfolgt so eine Konzentration auf die Kernbereiche des Datenmaterials, in denen in der Regel dann eine stark ausgeprägte lineare Beziehung besteht. Werden hingegen große kritische Werte verwendet, so ergibt sich dieser Effekt nicht und es kann passieren, dass signifikante Zusammenhänge verloren gehen.

Betrachtet man in diesen Fällen hingegen die Rangkorrelation nach Spearman, so geht anstelle der absoluten Distanz der Messwerte vom Datenzentrum nur deren Ordnung in die Berechnung ein und die Gefahr, Ausreißer zu „vorsichtig“ zu entfernen, ist damit deutlich geringer und nicht-lineare Abhängigkeiten werden linearisiert.

Kapitel 6

Hauptkomponentenanalyse

6.1 Motivation

„The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first *few* retain most of the variation present in *all* of the original variables.“ [9]

Dieses Zitat aus der Einleitung zu dem Buch von I. T. Jolliffe beschreibt die Ziele der Hauptkomponentenanalyse (englisch: „principal component analysis“). Der m-dimensionale Datenraum, der durch die gemeinsam (multivariat) zu analysierenden Zufallsvariablen $X_1, X_2, \ldots, X_m$ aufgespannt wird, soll durch einen p-dimensionalen Datenraum mit $p \ll m$ ersetzt werden. Hierbei soll jedoch ein möglichst großer Anteil der Information, die die Originaldaten liefern, erhalten bleiben. Deswegen macht man sich bei dem Konzept der Hauptkomponentenanalyse die Abhängigkeiten (Korrelationen) zwischen den Originalvariablen zu Nutze und bestimmt p geeignete Linearkombinationen der m ursprünglichen Variablen mit möglichst großer Varianz, die dann unkorreliert sind; geometrisch ausgedrückt bedeutet dies, dass die Streuung (Variation) in den Originaldaten in orthogonale Streurichtungen zerlegt wird, wobei die Streuintensität entlang dieser Richtungen immer weiter abnimmt. Bei diesem Vorgehen wird davon ausgegangen, dass die Gesamtstreuung in $p \ll m$ Richtungen hinreichend gut erfasst wird, also in den verbleibenden $(m-p)$ Richtungen eine nahezu konstante lineare Beziehung zwischen den Originalvariablen besteht. Die durch die oben genannten Linearkombinationen entstehenden, neuen p Zufallsvariablen $C_1, C_2, \ldots, C_p$ heißen die ersten p Hauptkomponenten der zu Grunde liegenden m-dimensionalen Verteilung.

6.2 Theoretischer Hintergrund

Um die in 6.1 postulierten Eigenschaften der Hauptkomponenten mathematisch zu fassen, bezeichnen nun $v_i, i = 1, \ldots, p$ mit $v_i \in \mathbb{R}^m \quad \forall\ i = 1, \ldots, p$ die Vektoren, die die Koeffizienten der zu den ersten p Hauptkomponenten gehörenden Linearkombinationen enthalten. Diese v_i (und damit die Hauptkomponenten $C_1, C_2, \ldots, C_p$) werden durch folgende (Optimalitäts-) Forderungen festgelegt:

P1: $v_i^t \cdot v_j = \delta_{i,j} \ \ \forall i, j = 1, \ldots, p$

P2: $(v_1, \ldots, v_k) = \arg(\max_{A \in \mathbb{R}^{m \times k}} (tr(A^t \cdot K \cdot A))) \ \ \forall k = 1, \ldots, p$

mit $K \in \mathbb{R}^{m \times m}$ Varianz- / Kovarianzmatrix der Originalvariablen $X_1, \ldots, X_m$

Die erste Forderung entspricht der Orthogonalität der entstehenden Streurichtungen sowie einer Normierung der Vektoren v_i bezüglich der L_2-Norm und die zweite Forderung formalisiert die sukzessive Bestimmung der Richtung mit maximaler Streuung, denn $K^* := A^t \cdot K \cdot A$ ist gerade die Varianz- / Kovarianzmatrix der Linearkombination $A^t \cdot (X_1, \ldots, X_m)^t$ (Basiswechsel im k-dimensionalen Teilraum).

Um die v_i zu berechnen, setzen wir zunächst $p = 1$. Damit vereinfachen sich P1 und P2 zu

$$v_1 = \arg(\max_{a \in \mathbb{R}^m}(a^t \cdot K \cdot a)) \text{ mit } ||a||_2 = 1.$$

Wendet man das Verfahren der Lagrange-Multiplikatoren zur Lösung von Maximierungsproblemen unter Nebenbedingungen an (siehe hierzu z.B. [10], Kapitel 3.5), so ergibt sich hier als Lagrange-Funktion der Ausdruck

$$a^t \cdot K \cdot a - \lambda_1 \cdot (a^t \cdot a - 1),\ \lambda_1 \in \mathbb{R}.$$

Partielles Differenzieren nach den Komponenten von a führt auf das Gleichungssystem

$$\begin{aligned} K \cdot a - \lambda_1 \cdot a &= 0 \\ \Longleftrightarrow (K - \lambda_1 \cdot E) \cdot a &= 0 \end{aligned}$$

Dieses letzte Gleichungssystem entspricht gerade dem Eigenwertproblem der Varianz- / Kovarianzmatrix K, so dass sich v_1 als ein Eigenvektor von K ergibt. Um nun zu entscheiden, welcher Eigenvektor das Maximierungsproblem löst, beachte man, dass

$$v_1^t \cdot K \cdot v_1 = v_1^t \cdot \lambda_1 \cdot v_1 = \lambda_1 \cdot v_1^t \cdot v_1 = \lambda_1$$

gilt. Also muss λ_1 der größte Eigenwert von K und v_1 der zugehörige Eigenvektor sein. Ferner gilt

$$Var(C_1) = Var(v_1^t \cdot (X_1, \ldots, X_m)^t) = v_1^t \cdot K \cdot v_1 = \lambda_1.$$

Es ergibt sich nun aufgrund der Orthonormalitätsforderung an die v_i, dass sich die Maximierung der Spur von K^* entkoppelt in p Maximierungsprobleme obiger Form (siehe dazu auch [9], Seiten 9f.), so dass sich letztendlich das folgende Ergebnis festhalten lässt:

1. v_i ist Eigenvektor zum Eigenwert λ_i von K mit $\lambda_1 > \lambda_2 > \ldots > \lambda_m$.
2. $Var(C_i) = \lambda_i,\ \ i = 1, \ldots, m.$

Bemerkung 6.2.1
Führt man eine Hauptkomponentenanalyse für standardisierte Zufallsgrößen durch, so geht die Varianz- / Kovarianzmatrix K in die zugehörige Korrelationsmatrix R über. Es gilt dann $Var(X_i) = 1\ \ \forall i = 1, \ldots, m$ und damit $tr(R) = m$. Will man also bei der Hauptkomponentenanalyse eine implizite Standardisierung bzw. Maßstabsvereinheitlichung der zu analysierenden Variablen erzielen, so kann anstelle der Varianz- / Kovarianzmatrix die Korrelationsmatrix zur Bestimmung der Eigenwerte und Eigenvektoren zur Definition der Hauptkomponenten herangezogen werden. Die so berechneten $(\tilde{\lambda_i}, \tilde{v_i})$ weichen im Allgemeinen von denen ab, die sich im Falle der Verwendung der Varianz- / Kovarianzmatrix ergeben.

Die im Rahmen dieser Diplomarbeit entstandene Implementierung lässt sowohl die Verwendung von Varianz- / Kovarianzmatrizen als auch die Verwendung von Korrelationsmatrizen bei der Hauptkomponentenanalyse zu, wobei in der konkreten Berechnung natürlich wieder zu den entsprechenden empirischen Größen K_n bzw. R_n übergegangen wird.

6.3 Beispiel

Beispiel 6.3.1 (Nationenwertung bei Olympischen Laufdisziplinen)
In [11] werden als einführendes Beispiel für die Hauptkomponentenanalyse die Zeiten für sieben Laufdisziplinen der Damen bei den Olympischen Spielen 1984 in Los Angeles betrachtet. Genauer handelt es sich hierbei um die Zeiten über die Distanzen 100 Meter, 200 Meter, 400 Meter, 800 Meter, 1500 Meter, 3000 Meter sowie den Marathonlauf. Diese Daten liegen für 55 Länder vor und werden in Anhang B wiedergegeben.

Will man die Nationen nun entsprechend ihrer Gesamtleistung bei allen sieben Disziplinen (die mit Zufallsvariablen identifiziert werden) ordnen, so ergibt sich zunächst einmal das Problem, dass die sieben Variablen unterschiedliche Maßstäbe aufweisen. Während die Kurzstreckenzeiten in Sekunden angegeben werden, liegen die Daten für die Mittel- und Langstrecken in Minuten vor. Nun könnte so vorgegangen werden, alle Ergebnisse in eine gemeinsame Zeiteinheit (z.B. Sekunden) umzurechnen und dann das arithmetische Mittel der sieben Werte pro Nation als Ordnungsgrundlage zu verwenden.

Wie jedoch leicht zu ersehen ist, differiert die Variation der Zeiten stark in Abhängigkeit von der Laufstrecke. So sind die Abstände beim Marathonlauf selbstverständlich wesentlich größer als beim 100 Meter-Lauf, so dass dieses Vorgehen nicht sinnvoll erscheint. Es bietet sich in diesem Fall sicherlich an, die Variablen zunächst auf Mittelwert 0 und Varianz 1 zu standardisieren, um sie einheitenfrei und damit untereinander vergleichbar zu machen.

Aber selbst nachdem diese Standardisierung durchgeführt worden ist, kann man sich die Frage stellen, ob es sinnvoll ist, das arithmetische Mittel der Werte für die Ordnung zu verwenden (also allen sieben Variablen das gleiche Gewicht $\frac{1}{7}$ beizumessen), oder ob sich nicht eine Linearkombination finden lässt, die mehr Information aus den Ausgangsdaten zieht, also in dem konkreten Beispiel die Unterschiede zwischen den Nationen besser wiedergibt.

Da dies aber gerade die in diesem Kapitel erläuterte Aufgabenstellung der Hauptkomponentenanalyse darstellt, wird in [11] eine Ordnung der Nationen gemäß ihrer Einträge zur ersten Hauptkomponente vorgeschlagen, wobei diese sich aus der Lösung des Eigenwertproblems der Korrelationsmatrix der sieben durch die Laufzeiten induzierten Variablen (aus Gründen der Maßstabsvereinheitlichung wird die Korrelationsmatrix verwendet) ergibt. Wie die in [11] mit dem integrierten Statistik-Softwaresystem SAS durchgeführten Berechnungen zeigen, lässt sich damit ein Anteil von 83% der Variation in den Ausgangsdaten erklären.

6.4 Anwendung

In praktischen Anwendungsfällen der Hauptkomponentenanalyse besteht die grundlegende Fragestellung offensichtlich darin, wie groß der Parameter p zu wählen ist. Dies wird davon abhängen, welchen Informationsverlust man hinzunehmen bereit ist. Ein häufig gemachter Vorschlag dazu besteht darin, p so zu wählen, dass im Falle der Verwendung von Korrelationsmatrizen $p = \min\{1 \leq i \leq m : \lambda_i < 1\}$ gilt. Anschaulich bedeutet diese Wahl von p, dass die einzubeziehenden Hauptkomponenten jeweils den Informationsgehalt mindestens einer Originalvariablen haben. In einigen Situationen kann diese Wahl von p jedoch dazu führen, dass einige Variablen komplett unberücksichtigt bleiben und deswegen empfiehlt I.T. Jolliffe in [9] die Wahl von $p = \min\{1 \leq i \leq m : \lambda_i < 0.7\}$, wenn möglichst sichergestellt werden soll, dass alle Variablen anteilig in den Hauptkomponenten Berücksichtigung finden.

Ein anderer Ansatz besteht darin, den Anteil der durch die Hauptkomponenten $C_1, \ldots, C_p$ erklärten Streuung, also $\sum_{i=1}^{p} \lambda_i$ mit $\lambda_1 > \ldots > \lambda_p$ Eigenwerte der Korrelationsmatrix $R \in \mathbb{R}^{m \times m}$ von $(X_1, \ldots, X_m)$ und ihr Verhältnis zu $tr(R) = m$ (es werden hier wie in Bemerkung 6.2.1 angegeben standardisierte Variablen X_i betrachtet) zum Entscheidungskriterium für die Größe von p zu machen. In dem Fall, dass sich einzelne Variablen annähernd mit Hauptkomponenten identifizieren lassen, also nur sehr gering mit den restlichen Zufallsgrößen korrelieren, wird es jedoch auch hier dazu kommen, dass diese Variablen durch die ersten Hauptkomponenten fast gar nicht erfasst werden. Da dies eventuell unerwünscht ist, ist zusätzlich eine Information darüber nötig, wie groß im Fall der am schlechtesten erfassten Variable deren „Restvarianz“ ist. Die Restvarianz bezeichnet den Anteil der Varianz der Variablen X_i, der nicht durch die Hauptkomponenten $C_1, \ldots C_p$ erklärt wird, also:

$$Var_{Rest,p}(X_i) = \min_{(a_1,\ldots,a_p)\in\mathbb{R}^p} Var(X_i - \sum_{j=1}^{p} a_j \cdot C_j).$$

Es ergibt sich:

$$\begin{aligned}
& Var(X_i - \sum_{j=1}^{p} a_j \cdot C_j) \\
= \;& Var(X_i) - 2 \cdot \sum_{j=1}^{p} a_j \cdot Kov(X_i, C_j) + \sum_{j=1}^{p}\sum_{l=1}^{p} a_j \cdot a_l \cdot Kov(C_j, C_l) \\
= \;& Var(X_i) + \sum_{j=1}^{p} a_j^2 \cdot Var(C_j) - 2 \cdot \sum_{j=1}^{p} a_j \cdot \sqrt{Var(C_j)} \cdot \frac{Kov(X_i, C_j)}{\sqrt{Var(C_j)}} \\
& + \sum_{j=1}^{p} \frac{Kov(X_i, C_j)^2}{Var(C_j)} - \sum_{j=1}^{p} \frac{Kov(X_i, C_j)^2}{Var(C_j)} \qquad (Kov(C_j, C_l) = 0 \; \forall j \neq l) \\
= \;& Var(X_i) + \sum_{j=1}^{p} (a_j \cdot \sqrt{Var(C_j)} - \frac{Kov(X_i, C_j)}{\sqrt{Var(C_j)}})^2 - \sum_{j=1}^{p} \frac{Kov(X_i, C_j)^2}{Var(C_j)} \\
= \;& Var(X_i) \cdot \left(1 - \sum_{j=1}^{p} \rho^2(X_i, C_j)\right) + \sum_{j=1}^{p} Var(C_j) \cdot (a_j - \frac{Kov(X_i, C_j)}{Var(C_j)})^2.
\end{aligned}$$

Aus dieser Rechnung folgt, dass die die obige Varianz minimierenden Koeffizienten a_j gerade die **Regressionskoeffizienten**

$$a_j = \frac{Kov(X_i, C_j)}{Var(C_j)}, \; \forall j = 1, \ldots, p$$

sind und die gesuchte Restvarianz damit von der Form

$$Var_{Rest,p}(X_i) = Var(X_i) \cdot \left(1 - \sum_{j=1}^{p} \rho^2(X_i, C_j)\right)$$

ist.

Um schließlich noch $\rho(X_i, C_j)$ zu berechnen, bezeichne $V = (v_{i,j}), \; i = 1, \ldots, m;$
$j = 1, \ldots, m; V \in \mathbb{R}^{m \times m}$ die Matrix der im L_2-Sinne normierten Eigenvektoren v_j von K, die nach der Größe der zugehörigen Eigenwerte absteigend geordnet sind, so dass sich die Hauptkomponenten $C_1, \ldots, C_m$ als $C_j = v_j^t \cdot (X_1, \ldots, X_m)^t \;\; \forall j = 1, \ldots, m$ ergeben. Dann gilt:

$$
\begin{aligned}
K &= V \cdot diag(\lambda_i) \cdot V^t \\
\Rightarrow Kov(X_i, X_k) &= \sum_{j=1}^{m} v_{i,j} \cdot \lambda_j \cdot v_{k,j} \\
Kov(X_i, C_j) &= Kov(X_i, \sum_{k=1}^{m} v_{k,j} \cdot X_k) \\
&= \sum_{k=1}^{m} Kov(X_i, X_k) \cdot v_{k,j} \\
&= \sum_{k=1}^{m} v_{k,j} \cdot (\sum_{l=1}^{m} v_{i,l} \cdot \lambda_l \cdot v_{k,l}) \\
&= \sum_{l=1}^{m} v_{i,l} \cdot \lambda_l \cdot (\sum_{k=1}^{m} v_{k,j} v_{k,l}) \\
&= \sum_{l=1}^{m} v_{i,l} \cdot \lambda_l \cdot \delta_{l,j} \\
&= v_{i,j} \cdot \lambda_j
\end{aligned}
$$

und damit (unter Beachtung von $Var(C_j) = \lambda_j$):

$$\rho(X_i, C_j) = \frac{v_{i,j} \cdot \sqrt{\lambda_j}}{\sqrt{Var(X_i)}}.$$

Um also dem Benutzer eine möglichst gute Entscheidungshilfe für die Wahl von p an die Hand zu geben, wurde in der erfolgten Implementierung eine Tabelle der folgenden Form ausgegeben:

i	λ_i	$\lambda_{i-1} - \lambda_i$	λ_i/m	$\sum_{j=1}^{i} \lambda_j/m$	$\max_{k \in \{1,\ldots,m\}} Var_{Rest,i}(X_k)$
1	λ_1	-	λ_1/m	λ_1/m	$\max_{k \in \{1,\ldots,m\}} Var_{Rest,1}(X_k)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	λ_p	$\lambda_{p-1} - \lambda_p$	λ_p/m	$\sum_{j=1}^{p} \lambda_j/m$	$\max_{k \in \{1,\ldots,m\}} Var_{Rest,p}(X_k)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	λ_m	$\lambda_{m-1} - \lambda_m$	λ_m/m	1	0

Für die Wahl von p können insbesondere die letzten beiden Spalten dieser Tabelle zu Rate gezogen werden. Überschreitet der Wert in der vorletzten bzw. unterschreitet der Wert in der letzten Spalte in einer Zeile p einen Schwellenwert, so ist dieses p als die Anzahl zu berücksichtigender Hauptkomponenten zu wählen.

6.5 Ausreißererkennung mit Hauptkomponenten

Wie im Falle der Betrachtung der Originalvariablen kann es auch beim Übergang zu den Hauptkomponenten von Interesse sein, Beobachtungen zu erkennen, die in dem (jetzt durch die Hauptkomponenten aufgespannten) Datenraum auffällig aus dem Gros der Daten herausfallen. In der vorliegenden Arbeit sind drei Ausprägungen solcher „Auffälligkeiten“ untersucht worden:

- Es kann Datenpunkte geben, die bezüglich der Verteilung einer bestimmten Hauptkomponente signifikant weit vom Datenzentrum entfernt liegen.

- Ein in Koordinaten des durch die Hauptkomponenten $(C_1, \ldots, C_p)$ gegebenen Koordinatensystems ausgedrückter Beobachtungsvektor kann einen signifikant großen Abstand zum Datenzentrum im $\mathbb{R}^p$ haben.
- Im Allgemeinen wird wie in Abschnitt 6.1 beschrieben davon ausgegangen, dass aus m Komponenten bestehenden Beobachtungen in dem durch die ersten p Hauptkomponenten gegebenen p-dimensionalen Teil-Datenraum hinreichend gut beschrieben werden. Es kann jedoch Beobachtungen geben, für welche diese Annahme nicht zutrifft.

Datenpunkte, bei welchen eine solche Situation vorliegt, werden hier wieder als Ausreißer bezeichnet. Um diese Ausreißer nun erkennen zu können, sind Teststatistiken mit zugehörigen Verlässlichkeitsschranken für die obigen Problemstellungen nötig. In den folgenden Abschnitten wird das in dieser Arbeit gewählte Vorgehen hierzu erläutert.

Dazu gelten vorab folgende Festsetzungen:

1. Die Ausgangsdaten bestehen aus den Beobachtungsvektoren $x_i = (x_{i,1}, \ldots, x_{i,m})^t,\ i = 1, \ldots, n$. Ein solcher Beobachtungsvektor enthält die Messwerte aus den Laborexperimenten, die mit einer chemischen Struktur durchgeführt wurden. Zusammengefasst werden diese Beobachtungsvektoren in der $(n \times m)$-Datenmatrix $X = (x_{i,j}),\ i = 1, \ldots, n;\ j = 1, \ldots, m$, wobei n die Anzahl der Beobachtungen (Strukturen) und m die Anzahl der Variablen (Experimente) bezeichnet. Die Zeilen von X sind also die transponierten Beobachtungsvektoren x_i. Zudem sind die Spaltenvektoren von X - die jeweils ein Laborexperiment widerspiegeln - auf Mittelwert 0 und Varianz 1 standardisiert.
2. Mit $V = (v_{i,j}),\ i = 1, \ldots, m;\ j = 1, \ldots, m; V \in \mathbb{R}^{m \times m}$ wird die Matrix der im L_2-Sinne normierten Eigenvektoren v_j der Korrelationsmatrix der Originalvariablen bezeichnet. Diese sind nach der Größe der zugehörigen Eigenwerte absteigend geordnet, so dass sich die Hauptkomponenten $C_1, \ldots, C_m$ als $C_j = v_j^t \cdot (X_1, \ldots, X_m)^t\ \forall j = 1, \ldots, m$ ergeben.
3. Die Matrix der auf $C_1, \ldots, C_m$ bezogenen Koordinaten der Beobachtungen (im Folgenden auch als Matrix der „Scores“ der Beobachtungen bezeichnet) wird notiert mit $Y = X \cdot V;\ Y \in \mathbb{R}^{n \times m}$. Auch im Falle der Scores liegen die Beobachtungen also wieder als Zeilenvektoren vor.

6.5.1 Eindimensionale Ausreißer

Bei der Bestimmung von Ausreißern bezüglich einer speziellen Hauptkomponente C_j wurden zunächst die Laborexperimente bzw. die durch sie induzierten Zufallsvariablen als normalverteilt vorausgesetzt. Diese Annahme wurde in Ermangelung von Informationen über die tatsächlichen Verteilungen getroffen und muss daher im Einzelfall nicht zwingend korrekt sein. Die k-te Spalte von $X, X^{(k)}, k = 1, \ldots, m$ wird also als Vektor von n unabhängigen und identisch $\mathbb{N}(0, 1)$-verteilten Zufallsgrößen interpretiert. Damit ergibt sich die mit einer Hauptkomponente C_j assoziierte Spalte $Y^{(j)}, j = 1, \ldots, m$ der Score-Matrix Y ebenfalls als ein Vektor von n normalverteilten Größen, denn sie entsteht durch eine Linearkombination von $\mathbb{N}(0, 1)$-verteilten Variablen. Zudem sind die einzelnen Komponenten von $Y^{(j)}$ wieder unabhängig, denn es gehen in die Linearkombination nur Werte der jeweils gleichen Beobachtung ein. Wie in Abschnitt 6.2 hergeleitet, ergibt sich die Varianz der Größen in einer Spalte $Y^{(j)}$ als Wert des zugehörigen Eigenwertes λ_j der zu Grunde liegenden Korrelationsmatrix.

$$\Rightarrow y_{i,j} \text{ i.i.d.} \sim \mathbb{N}(0, \lambda_j)\ \ \forall i = 1, \ldots, n;\ j \in \{1, \ldots, m\}.$$

$$\Rightarrow z_{i,j} := \frac{y_{i,j}}{\sqrt{\lambda_j}} \text{ i.i.d.} \sim \mathbb{N}(0, 1)\ \ \forall i = 1, \ldots, n;\ j \in \{1, \ldots, m\}.$$

Damit lassen sich die normierten Scores $z_{i,j}$ als Teststatistiken zur Ausreißererkennung verwenden. In praxi ist jedoch wiederum zu beachten, dass anstelle der exakten Korrelationsmatrizen die ent-

sprechenden Schätzungen verwendet werden müssen und die hier gemachten Verteilungsaussagen dann nur noch asymptotisch gelten. Wird nun analog zu der Vorgehensweise in Abschnitt 3.3.3 wieder gefordert, dass nur in $\alpha \cdot 100\%$ der Fälle Beobachtungen als Ausreißer deklariert werden, wenn keine tatsächlichen Ausreißer vorliegen, so ergibt sich asymptotisch:

$$\begin{array}{lll}
 & \mathbb{P}(\max\limits_{1\le i\le n}|z_{i,j}|>c) & \stackrel{!}{=} \alpha \\
\iff & \mathbb{P}(|z_{i,j}|\le c\ \forall i=1,\dots,n) & = 1-\alpha \\
\iff & \mathbb{P}(-c\le z_{i,j}\le c\ \forall i=1,\dots,n) & = 1-\alpha \\
\underset{\text{Unabhängigkeit}}{\iff} & \prod_{i=1}^{n}\mathbb{P}(-c\le z_{i,j}\le c) & = 1-\alpha \\
\underset{\text{Verteilungsidentität}}{\iff} & (\mathbb{P}(-c\le Z\le c))^n & = 1-\alpha \text{ mit } Z\sim \mathbb{t}(0,1) \\
\iff & (\Phi(c)-\Phi(-c))^n & = 1-\alpha \\
\iff & (\Phi(c)-(1-\Phi(c)))^n & = 1-\alpha \\
\iff & 2\cdot\Phi(c)-1 & = \sqrt[n]{(1-\alpha)} \\
\iff & \Phi(c) & = \frac{1}{2}\cdot(1+\sqrt[n]{(1-\alpha)}) \\
\iff & c & = \Phi^{-1}(\frac{1+\sqrt[n]{(1-\alpha)}}{2}).
\end{array}$$

Überschreitet $|z_{i,j}| = |y_{i,j}/\sqrt{\lambda_j}|$ für eine Beobachtung i in Spalte j der Score-Matrix Y also das Quantil $c = \Phi^{-1}(\frac{1+\sqrt[n]{(1-\alpha)}}{2})$ der Standard-Normalverteilung, so wird der zugehörige Datenpunkt als eindimensionaler Ausreißer zur Hauptkomponente C_j angesehen.

6.5.2 *p*-dimensionale Ausreißer

Betrachtet man Ausreißer in dem Raum, der durch die ersten p Hauptkomponenten aufgespannt wird, so bietet es sich aus Gründen der Maßstabsvereinheitlichung wieder an, die normierten Scores $z_{i,j}$ für das Testproblem zu verwenden. Berechnet man den quadrierten p-dimensionalen euklidischen Abstand

$$d_p^2(z_i,0) = \sum_{k=1}^{p}(\frac{y_{i,k}}{\sqrt{\lambda_k}})^2$$

der normierten Scores einer Beobachtung i zu ihrem Mittelwert $0\in\mathbb{R}^p$, so ist dieser Ausdruck eine Summe von Quadraten asymptotisch standardnormalverteilter Zufallsgrößen und somit lässt sich die asymptotische Verteilung von $d_p^2(z_i,0)$ als χ^2-Verteilung mit p Freiheitsgraden (kurz: χ_p^2-Verteilung) angeben. Damit ergibt sich zur Bestimmung des kritischen Wertes für $d_p^2(z_i,0)$ analog zur obigen Berechnung, wenn wieder mit Konfidenzwahrscheinlichkeit α nur „echte“ Ausreißer erkannt werden sollen:

$$\begin{array}{lll}
 & \mathbb{P}(\max\limits_{1\le i\le n} d_p^2(z_i,0)>c) & \stackrel{!}{=} \alpha \\
\iff & \mathbb{P}(d_p^2(z_i,0)\le c\ \forall i=1,\dots,n) & = 1-\alpha \\
\underset{\text{Unabhängigkeit}}{\iff} & \prod_{i=1}^{n}\mathbb{P}(d_p^2(z_i,0)\le c) & = 1-\alpha \\
\underset{\text{Verteilungsidentität}}{\iff} & (\mathbb{P}(Z\le c))^n & = 1-\alpha \text{ mit } Z\sim\chi_p^2 \\
\iff & (F_{\chi_p^2}(c))^n & = 1-\alpha \\
\iff & F_{\chi_p^2}(c) & = \sqrt[n]{(1-\alpha)} \\
\iff & c & = \chi^2_{p;\,\sqrt[n]{(1-\alpha)}}.
\end{array}$$

Als kritischer Wert für $d_p^2(z_i,0)$ wird also hier das $\sqrt[n]{(1-\alpha)}$-Quantil der χ^2-Verteilung mit p Freiheitsgraden gewählt und Punkten, deren quadrierter p-dimensionaler euklidischer Abstand vom Nullpunkt diesen Wert c überschreitet, das Attribut „Ausreißer“ zugeordnet.

6.5.3 Schlecht modellierte Beobachtungen

Als letztes werden in diesem Abschnitt die Beobachtungen daraufhin untersucht, inwiefern sie gegebenenfalls durch p Hauptkomponenten schlecht modelliert werden. Hierzu beachte man zunächst, dass die Matrix V orthogonal und daher invertierbar mit $V^{-1} = V^t$ ist. Damit ist es möglich, die Originalvariablen, also die Beobachtungsmatrix X, als Funktion der Scores zu schreiben:

$$Y = X \cdot V \underset{\text{V orthogonal}}{\iff} X = Y \cdot V^t \text{ mit } V \in \mathbb{R}^{m \times m} \text{ und } Y \in \mathbb{R}^{n \times m}.$$

Verwendet man nun anstelle der vollständigen Matrizen $V \in \mathbb{R}^{m \times m}$ und $Y \in \mathbb{R}^{n \times m}$ nur jeweils deren ersten p Spalten, so geht die rechte Seite der obigen Äquivalenz in die Gleichung

$$\hat{X} = \tilde{Y} \cdot \tilde{V}^t \text{ mit } \tilde{V} \in \mathbb{R}^{m \times p} \text{ und } \tilde{Y} \in \mathbb{R}^{n \times p}$$

über, wobei $\hat{X}$ die Matrix der Projektion von X auf den durch die ersten p Hauptkomponenten $(C_1, \ldots, C_p)$ aufgespannten p-dimensionalen Teilraum des $\mathbb{R}^m$ ist.
Will man nun messen, wie gut ein Beobachtungsvektor $x_i, i = 1, \ldots, n$ durch das p-dimensionale Modell beschrieben wird, so bietet sich die Verwendung der Länge des Residuenvektors $x_i - \hat{x}_i$ an. Deswegen wird in [12] in Abschnitt 2.7.2 die dort als „Q-Statistik“ bezeichnete Größe

$$Q_i = (x_i - \hat{x}_i)^t \cdot (x_i - \hat{x}_i) \text{ mit } x_i \in \mathbb{R}^m \text{ Beobachtungsvektor}$$

zur Behandlung der erwähnten Fragestellung eingeführt. Aus Berechnungssicht ist es günstig, dass sich Q_i auch schreiben lässt als

$$Q_i = \sum_{k=p+1}^{m} y_{i,k}^2,$$

wobei y_i den Score-Vektor zum Beobachtungsvektor x_i bezeichnet.

Die Verteilung der so definierten Teststatistik Q_i lässt sich nur mühsam herleiten. In die Berechnung gehen im Wesentlichen die Summen der ersten drei Potenzen der $(m - p)$ kleinsten Eigenwerte der Korrelationsmatrix ein. In [12] wird das Ergebnis dieser Berechnung angegeben und es finden sich weitere Literaturhinweise für dieses Problem. Es ergibt sich, dass eine geschickte Transformation von Q_i zu einer möglichst guten Approximation der Standardnormalverteilung führt, so dass entsprechend tranformierte Quantile der $\mathcal{N}(0, 1)$-Verteilungsfunktion als Schranken für Q_i verwendet werden können. Die genaue Gestalt dieser Quantile kann dem Anhang C entnommen werden.

Will man nun also die Beobachtungsvektoren $x_i, i = 1, \ldots, n$ daraufhin untersuchen, ob sie signifikant schlecht durch das p-dimensionale lineare Modell der ersten p Hauptkomponenten beschrieben werden, so gilt es zu überprüfen, ob

$$Q_i = (x_i - \hat{x}_i)^t \cdot (x_i - \hat{x}_i) = \sum_{k=p+1}^{m} y_{i,k}^2$$

das zugehörige, transformierte $\mathcal{N}(0, 1)$-Quantil überschreitet.

Kapitel 7

Variablenselektion

7.1 Motivation

In dem vorangegangenen Kapitel wurde mit der Hauptkomponentenanalyse eine Methode vorgestellt, die Dimensionalität des zu untersuchenden Datenraumes zu reduzieren und dabei nur einen möglichst geringen Informationsverlust hinnehmen zu müssen. Dies ist zum Beispiel im Hinblick auf die Berechnungs- und Speicherplatzkomplexität von auf das Datenmaterial anzuwendenden Algorithmen günstig, denn gegebenenfalls ist es möglich, diese nur auf dem durch einige Hauptkomponenten gegebenen, niederdimensionalen Datenraum auszuführen und die gelieferten Ergebnisse dann eventuell so zu transformieren, dass wieder Informationen über die Ausgangsdaten gewonnen werden.

Verfolgt man jedoch das Ziel, den Datenraum für einen Analysten überschaubar zu machen oder ist man daran interessiert, einige Variablen ganz aus der Betrachtung auszuklammern (also gegebenenfalls einige Laborexperimente gar nicht mehr durchführen zu müssen), so ergeben sich bei der Verwendung von Hauptkomponenten für diese Aufgaben zwei grundlegende Schwierigkeiten:

- Hauptkomponenten sind als Linearkombinationen der Originalvariablen wesentlich schlechter interpretierbar als die Ausgangsdaten, welche unmittelbar die experimentell erhobenen Meßwerte wiedergeben.
- In die Hauptkomponenten gehen in der Regel alle Variablen mit einem festgelegten Gewichtsfaktor ein, es kann also kein Experiment fortgelassen werden.

Aus diesen Gründen ist es oftmals günstiger, anstelle von p Variablenkombinationen mit maximalem Informationsgehalt lieber p der m Originalvariablen selbst auszuwählen, die am repräsentativsten für das gesamte Datenmaterial sind, um die Dimension des zu betrachtenden Datenraums dadurch zu verringern. Dieses Vorgehen der *Variablenselektion* wird in diesem Kapitel behandelt und es werden Algorithmen erläutert, die verschiedene Auswahlkriterien für die zu selektierenden Variablen implementieren, wobei in der vorliegenden Diplomarbeit zwei generelle Vorgehensweisen umgesetzt worden sind:

1. Verfahren, die sich der Ergebnisse der Hauptkomponentenanalyse bedienen.
2. Das Verfahren der „Haupt-Variablen“ („principal variables“) nach McCabe.

7.2 Auf Hauptkomponenten basierende Verfahren

Die von I.T. Jolliffe in [13] publizierten und in dieser Diplomarbeit implementierten, auf Hauptkomponenten basierenden Verfahren zur Variablenselektion arbeiten derart, dass ausgewählte Spalten der in Kapitel 6.4 definierten Matrix V, welche die Eigenvektoren der der Hauptkomponenten-

analyse zu Grunde gelegten Varianz- / Kovarianzmatrix bzw. Korrelationsmatrix enthält, auf ihre Maximaleinträge hin untersucht werden. Weist ein Spaltenvektor y_j von Y seinen größten Eintrag in Zeile i auf, wobei $1 \leq i, j \leq m$ gilt, so wird

(a) die Variable X_i ausgewählt, falls j einen kleinen Wert hat und X_i demnach mit einem großen Gewicht in die zugehörige Hauptkomponente C_j mit $j \ll m$ eingeht (*Selektion*).

(b) die Variable X_i als zu vernachlässigen angesehen und aus der Menge der zu untersuchenden Variablen entfernt, falls j nahe bei m liegt (*Elimination*).

Damit nun genau p auszuwählende Variablen bestimmt werden können, wird das Verfahren der Selektion oder der Elimination iterativ wiederholt, bis dass genau p Variablen selektiert bzw. genau $m-p$ Variablen eliminiert worden sind. Dabei ist es zum einen möglich, nach jedem Selektions- bzw. Eliminationsschritt eine neue Hauptkomponentenanalyse für die verbleibenden Variablen durchzuführen oder die Ergebnisse einer einzigen Hauptkomponentenanalyse für den gesamten Selektions- bzw. Eliminationsvorgang zu nutzen.

Entscheidet man sich dazu, die p Variablen über das Verfahren der Selektion zu bestimmen und zudem eine neue Hauptkomponentenanalyse für jeden Selektionsschritt zu tätigen, so gilt es zu beachten, dass die Variablen in der Regel untereinander korreliert sind. Ist jedoch die Variable X_i einmal ausgewählt worden, so soll für die weiteren Auswahlschritte ihr Einfluß auf die restlichen noch verbliebenen Variablen ausgeschaltet werden. Deswegen gilt es zu untersuchen, wie sich die **partielle Varianz- / Kovarianzmatrix** $K_{m-1,k}$ berechnet. $K_{m-1,k}$ beinhaltet hierbei die Varianzen/Kovarianzen der $m-1$ verbleibenden Variablen, wenn man die Variable X_k aus der Menge der Variablen $(X_1, \ldots, X_m)$ entfernt und ihren Einfluß auf die restlichen Variablen „herausrechnet".

Nehmen wir zur Lösung dieses Problems an, es soll die partielle Kovarianz zweier Zufallsvariablen X und Y berechnet werden, wobei der Einfluß einer dritten Zufallsvariable Z, die sowohl mit X als auch mit Y korreliert ist, ausgeschaltet werden soll. Dazu setzen wir fest:

$$\begin{aligned} X_Z &:= a_{x,z} + b_{x,z} \cdot Z \\ Y_Z &:= a_{y,z} + b_{y,z} \cdot Z \end{aligned}$$

mit den **Regressionskoeffizienten**

$$\begin{aligned} b_{x,z} &= \frac{Kov(X,Z)}{Var(Z)} \\ b_{y,z} &= \frac{Kov(Y,Z)}{Var(Z)}. \end{aligned}$$

X_Z bzw. Y_Z bezeichnen also hierbei den Anteil der Variablen X bzw. Y, der durch ihren linearen Bezug zu Z gegeben ist.

$$\begin{aligned} &\Rightarrow \quad Kov((X - X_Z), (Y - Y_Z)) \\ &= \quad Kov((X - (a_{x,z} + b_{x,z} \cdot Z)), (Y - (a_{y,z} + b_{y,z} \cdot Z))) \\ &= \quad Kov(X,Y) - Kov(X, a_{y,z} + b_{y,z} \cdot Z) - Kov(Y, a_{x,z} + b_{x,z} \cdot Z) \\ &\quad + Kov(a_{x,z} + b_{x,z} \cdot Z, a_{y,z} + b_{y,z} \cdot Z) \\ &= \quad Kov(X,Y) - b_{y,z} \cdot Kov(X,Z) - b_{x,z} \cdot Kov(Y,Z) + b_{x,z} \cdot b_{y,z} \cdot Var(Z) \end{aligned}$$

Beachtet man die Definition von $b_{x,z}$ und $b_{y,z}$, so vereinfacht sich der Ausdruck weiter:

$$\begin{aligned} & Kov(X,Y) - b_{y,z} \cdot Kov(X,Z) - b_{x,z} \cdot Kov(Y,Z) + b_{x,z} \cdot b_{y,z} \cdot Var(Z) \\ = \; & Kov(X,Y) - 2 \cdot \frac{Kov(X,Z) \cdot Kov(Y,Z)}{Var(Z)} + \frac{Kov(X,Z) \cdot Kov(Y,Z)}{Var(Z)} \\ = \; & Kov(X,Y) - \frac{Kov(X,Z) \cdot Kov(Y,Z)}{Var(Z)}. \end{aligned}$$

Es gilt also für die partielle Varianz- / Kovarianzmatrix:

$$K_{m-1,k} = K'_{m,k} - \left(\frac{1}{Var(X_k)} \cdot K^{(k)} \cdot K^{(k)^t} \right)'_{m,k}$$

mit

$K_{m-1,k}$	partielle Varianz- / Kovarianzmatrix ohne Einfluß von X_k
$K'_{m,k}$	Streichungsmatrix, die aus K durch Streichung von Zeile k und Spalte k hervorgeht.
$K^{(k)}$	k-te Spalte der Matrix K.

Aus algorithmischer Sicht ist hierbei natürlich zu beachten, dass die Indizes der verbleibenden Variablen verschoben werden, wenn $k \neq m$ gilt. In den folgenden Abschnitten wird nun kurz für die vier auf Hauptkomponenten basierenden Verfahren, die im Rahmen dieser Diplomarbeit umgesetzt worden sind, das jeweils zugehörige Rekursionsschema angegeben.

7.2.1 Selektion mit *p* Hauptkomponentenanalysen

Zunächst wird eine Hauptkomponentenanalyse der Originalvariablen $(X_1, \ldots, X_m)$ auf Basis der Varianz- / Kovarianzmatrix K bzw. (standardisierte Größen) der Korrelationsmatrix R durchgeführt, um die Matrix $V = (v_{i,j})$ der Eigenvektoren zu erhalten. Zudem wird die Anzahl p auszuwählender Variablen vorgegeben. Die selektierten Variablen $(X_{i_1}, \ldots, X_{i_p})$ werden nun dadurch festgelegt, dass zunächst

$$i_1 = \arg(\max_{1 \leq i \leq m} (v_{i,1}))$$

gewählt wird. Die Variable X_{i_1} ist hiermit ausgewählt und wie oben beschrieben soll ihr Einfluß auf die restlichen Variablen für den nächsten Selektionsschritt ausgeschaltet werden. Es wird also die partielle Varianz- / Kovarianzmatrix K_{m-1,i_1} der verbleibenden $m-1$ Variablen gebildet und auf ihrer Basis eine erneute Hauptkomponentenanalyse durchgeführt. Nun wird wieder der Maximaleintrag des zum größten Eigenwert gehörenden Eigenvektors von K_{m-1,i_1} gesucht, die entsprechende Zeile mit i_2 bezeichnet und die Variable X_{i_2} dem Satz ausgewählter Variablen hinzugefügt. Dies wiederholt sich sodann, bis dass p so selektierte Variablen feststehen.

7.2.2 Selektion mit genau einer Hauptkomponentenanalyse

Will man den nicht unerheblichen Berechnungsaufwand, der sich durch die p zu tätigenden Hauptkomponentenanalysen der in 7.2.1 vorgestellten Methode ergibt, vermeiden, so kann die Selektion von p Variablen auch auf der Basis einer einzigen Hauptkomponentenanalyse geschehen. Zunächst ergibt sich die Matrix $V = (v_{i,j})$ der Eigenvektoren analog zu dem Vorgehen in 7.2.1. Dann werden

jedoch alle Spaltenvektoren $v_j,\ j = 1, \ldots, p$ von V auf ihren Maximaleintrag hin untersucht und die Indizes $i_1, \ldots, i_p$ der selektierten Variablen ergeben sich zu:

$$\begin{aligned} i_1 &= \arg(\max_{1 \le i \le m} (v_{i,1})) \\ i_2 &= \arg(\max_{1 \le i \le m : i \neq i_1} (v_{i,2})) \\ &\vdots \quad\quad \vdots \\ i_p &= \arg(\max_{1 \le i \le m : i \notin \{i_1, \ldots, i_{p-1}\}} (v_{i,p})). \end{aligned}$$

Hier ist also die Berechnung der partiellen Varianz- / Kovarianzmatrizen vom Algorithmus her nicht gefordert. Für die Darstellung der Ergebnisse (vgl. Kapitel 7.4) sind die Restvarianzen hingegen doch erforderlich, so dass auch hier die entsprechenden Größen berechnet wurden.

7.2.3 Elimination mit *(m-p)* Hauptkomponentenanalysen

Diesem Verfahren liegt die umgekehrte Argumentationsrichtung gegenüber dem Verfahren in 7.2.1 zu Grunde. Es werden nicht die Variablen, die stark in die ersten Hauptkomponenten eingehen, ausgewählt, sondern vielmehr diejenigen, die sich am besten mit den letzten Hauptkomponenten identifizieren lassen, eliminiert. Hier werden also Indizes $(i_1, \ldots, i_{m-p})$ bestimmt, so dass die Variablen $(X_{i_1}, \ldots, X_{i_{m-p}})$ vernachlässigt und die verbleibenden Variablen mit einem Index i^*, für den $i^* \notin \{i_1, \ldots, i_{m-p}\}$ gilt, demnach als ausgewählt angesehen werden. Es wird also (mit den bisher in diesem Kapitel gewählten Bezeichnungen) i_1 hier als

$$i_1 = \arg(\max_{1 \le i \le m} (v_{i,m}))$$

festgelegt. Danach wird die Varianz- / Kovarianzmatrix K'_{m,i_1} der verbleibenden $m-1$ Variablen, die sich aus der ursprünglichen Varianz- / Kovarianzmatrix K durch Streichung von Zeile i_1 und Spalte i_1 ergibt, für eine neue Hauptkomponentenanalyse benutzt und es werden iterativ die restlichen $m-p-1$ Indizes bestimmt, deren zugehörige Variablen eliminiert werden sollen.

Es ist anzumerken, dass in dem Regelfall $p \ll m$ dieses Verfahren den bei weitem größten Rechenaufwand aller hier diskutierten Methoden verursacht.

7.2.4 Elimination mit genau einer Hauptkomponentenanalyse

Das vierte angewendete Verfahren verwirklicht die Grundidee des vorangegangenen, wobei jedoch auf eine erneute Hauptkomponentenanalyse pro Eliminationsschritt verzichtet wird. Die Indizes $(i_1, \ldots, i_{m-p})$ berechnen sich damit in Analogie zu der in 7.2.2 gewählten Vorgehensweise zu:

$$\begin{aligned} i_1 &= \arg(\max_{1 \le i \le m} (v_{i,m})) \\ i_2 &= \arg(\max_{1 \le i \le m : i \neq i_1} (v_{i,m-1})) \\ &\vdots \quad\quad \vdots \\ i_{m-p} &= \arg(\max_{1 \le i \le m : i \notin \{i_1, \ldots, i_{m-p-1}\}} (v_{i,p+1})). \end{aligned}$$

7.3 Verfahren der „Principal variables“

Das von McCabe in [14] beschriebene Verfahren der „Haupt-Variablen“ (in Anlehnung an den Ausdruck „Hauptkomponenten“) stützt sich nicht auf die Ergebnisse einer Hauptkomponentenanalyse, sondern versucht, die Optimalitätseigenschaft der Hauptkomponenten direkt auf einzelne Variablen zu übertragen. Diese Optimalitätseigenschaft besteht (vgl. hierzu Kapitel 6.2) darin, dass die erste Hauptkomponente C_1 eine möglichst große Varianz besitzt und damit die Summe der Restvarianzen minimiert. Wie in Kapitel 6.4 (Definition des Begriffs „Restvarianz“) beschrieben, läßt sich der Anteil der durch eine Hauptkomponente C_j erklärten Varianz einer Variablen X_i ausdrücken durch $Var(X_i) \cdot \rho^2(X_i, C_j)$. Setzt man nun anstelle von C_j eine Originalvariable X_j ein und fordert, dass die Summe der Restvarianzen der Variablen $\{X_i : i \neq j\}$ minimiert oder invers formuliert die Summe der durch X_j erklärten Varianz der restlichen Variablen maximiert wird, so führt dies also auf die folgende Maximierungsaufgabe:

$$\begin{aligned} j_1 &= \arg(\max_{1\leq j\leq m}(\sum_{i=1}^{m} Var(X_i)\cdot\rho^2(X_i,X_j))) \\ &= \arg(\max_{1\leq j\leq m}(\sum_{i=1}^{m} \frac{Kov^2(X_i,X_j)}{Var(X_j)})) \\ &= \arg(\max_{1\leq j\leq m}(\frac{1}{Var(X_j)}\cdot\sum_{i=1}^{m} Kov^2(X_i,X_j))) \end{aligned}$$

Die zu dem so bestimmten Index j_1 gehörende Variable X_{j_1} wird in diesem Verfahren als erste ausgewählt. Anschließend wird ihr Einfluß auf die restlichen Variablen wieder „herausgerechnet“, indem die Quadrate der Einträge der partiellen Varianz- / Kovarianzmatrix K_{m-1,j_1} in der obigen Maximierungsaufgabe betrachtet werden und so bestimmen sich iterativ die übrigen Indizes $(j_2, \ldots, j_p)$ für die Variablenselektion. In der praktischen Verwendung bietet sich dieses Verfahren aufgrund seiner einfachen Implementierbarkeit und geringen Ressourceninanspruchnahme an. Da zudem im Falle der hier untersuchten Daten aus der pharmazeutischen Forschung die so erzielten Ergebnisse qualitativ nicht hinter denen zurückgeblieben sind, welche sich bei Anwendung der in Kapitel 7.2 vorgestellten, aufwändigeren Methoden ergaben, hat sich hier das Verfahren von McCabe als überlegen im Kosten- / Nutzenvergleich erwiesen.

7.4 Anwendung

In Analogie zu der Darstellung der Ergebnisse der Hauptkomponentenanalyse wurden auch im Falle der Variablenselektion die Resultate tabellarisch in der folgenden Form wiedergegeben:

j	i_j	η_j	$\eta_{j-1} - \eta_j$	η_j/m	$\sum_{k=1}^{j} \eta_k/m$	$\max_{k\in\{1,\ldots,m\}} Var_{Rest,j}(X_k)$
1	i_1	η_1	-	η_1/m	η_1/m	$\max_{k\in\{1,\ldots,m\}} Var_{Rest,1}(X_k)$
2	i_2	η_2	$\eta_1 - \eta_2$	η_2/m	$(\eta_1 + \eta_2)/m$	$\max_{k\in\{1,\ldots,m\}} Var_{Rest,2}(X_k)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
p	i_p	η_p	$\eta_{p-1} - \eta_p$	η_p/m	$\sum_{k=1}^{p} \eta_k/m$	$\max_{k\in\{1,\ldots,m\}} Var_{Rest,p}(X_k)$

mit

j	Iteration
i_j	Nummer der in Iteration j ausgewählten Variable

η_j	durch X_{i_j} erklärte Varianz
$Var_{Rest,j}(X_k)$	Restvarianz von Variable X_k nach Iteration j.

7.5 Ergebnisse

Abschließend werden in der folgenden Tabelle anhand eines authentischen Beispieldatensatzes aus der pharmazeutischen Forschung mit $m = 100$ Variablen die Ergebnisse der implementierten Variablenselektionsverfahren gegenüber gestellt:

Kapitel	p	η_p	$\sum_{k=1}^{p} \eta_k/m$	$\max_{k\in\{1,\ldots,m\}} Var_{Rest,p}(X_k)$
7.2.1	22	0.761	0.903	0.638
7.2.1	45	0.085	0.995	0.046
7.2.2	24	0.727	0.905	0.507
7.2.2	66	0.0084	0.999	0.0497
7.2.3	29	0.881	0.9004	0.328
7.2.3	46	0.3496	0.991	0.0387
7.2.4	25	1.049	0.901	0.6939
7.2.4	53	0.1101	0.9984	0.0169
7.3	21	0.7963	0.9007	0.5983
7.3	44	0.0993	0.9932	0.0469

In der jeweils ersten Zeile zu jedem Verfahren wurde p so gewählt, dass der kumulierte Anteil erklärter Streuung über 0.9 liegt und die zweite Zeile gehört zu dem Wert von p, ab welchem jede Restvarianz den Schwellenwert von 5 % unterschreitet.

Anhang A

Datenmaterial zu Beispiel 2.3.2

Name	Jahre	Treffer
Aldrete, Mike	1.000000	54.000000
Allanson, Andy	1.000000	66.000000
Almon, Bill	13.000000	43.000000
Anderson, Dave	4.000000	53.000000
Armas, Tony	11.000000	112.000000
Ashby, Alan	14.000000	81.000000
Backman, Wally	7.000000	124.000000
Baines, Harold	7.000000	169.000000
Baker, Dusty	19.000000	58.000000
Balboni, Steve	6.000000	117.000000
Bando, Chris	6.000000	68.000000
Barfield, Jesse	6.000000	170.000000
Barrett, Marty	5.000000	179.000000
Bass, Kevin	5.000000	184.000000
Baylor, Don	17.000000	139.000000
Beane, Billy	3.000000	39.000000
Bell, Buddy	15.000000	158.000000
Bell, George	5.000000	198.000000
Belliard, Rafael	5.000000	72.000000
Beniquez, Juan	15.000000	103.000000
Bernazard, Tony	8.000000	169.000000
Biancalana, Buddy	5.000000	46.000000
Bilardello, Dann	4.000000	37.000000
Bochte, Bruce	12.000000	104.000000
Bochy, Bruce	8.000000	32.000000
Boggs, Wade	5.000000	207.000000
Bonds, Barry	1.000000	92.000000
Bonilla, Bobby	1.000000	109.000000
Bonilla, Juan	5.000000	69.000000
Boone, Bob	15.000000	98.000000
Boston, Daryl	3.000000	53.000000
Bradley, Phil	4.000000	163.000000
Bradley, Scott	3.000000	66.000000
Braggs, Glenn	1.000000	51.000000
Bream, Sid	4.000000	140.000000
Brenly, Bob	6.000000	116.000000
Brett, George	14.000000	128.000000
Brock, Greg	5.000000	76.000000
Brookens, Tom	8.000000	76.000000
Brooks, Hubie	7.000000	104.000000
Brown, Chris	3.000000	132.000000
Brown, Mike	4.000000	53.000000
Brunansky, Tom	6.000000	152.000000
Buckner, Bill	18.000000	168.000000

Buechele, Steve	2.000000	112.000000
Burleson, Rick	12.000000	77.000000
Bush, Randy	5.000000	96.000000
Butler, Brett	6.000000	163.000000
Cabell, Enos	15.000000	71.000000
Cangelosi, John	2.000000	103.000000
Canseco, Jose	2.000000	144.000000
Carter, Gary	13.000000	125.000000
Carter, Joe	4.000000	200.000000
Castillo, Carmen	5.000000	57.000000
Cerone, Rick	12.000000	56.000000
Cey, Ron	16.000000	70.000000
Clark, Jack	12.000000	55.000000
Clark, Will	1.000000	117.000000
Coleman, Vince	2.000000	139.000000
Coles, Darnell	4.000000	142.000000
Collins, Dave	12.000000	113.000000
Concepcion, Dave	17.000000	81.000000
Cooper, Cecil	16.000000	140.000000
Cruz, Jose	17.000000	133.000000
Cruz, Julio	10.000000	45.000000
Daniels, Kal	1.000000	58.000000
Daulton, Darren	3.000000	31.000000
Davis, Alan	3.000000	130.000000
Davis, Chili	6.000000	146.000000
Davis, Eric	3.000000	115.000000
Davis, Glenn	3.000000	152.000000
Davis, Jody	6.000000	132.000000
Davis, Mike	7.000000	131.000000
Dawson, Andre	11.000000	141.000000
DeCinces, Doug	14.000000	131.000000
Deer, Rob	3.000000	108.000000
Dempsey, Rick	18.000000	68.000000
Dernier, Bob	7.000000	73.000000
Diaz, Bo	10.000000	129.000000
Diaz, Mike	2.000000	56.000000
Doran, Bill	5.000000	152.000000
Downing, Brian	14.000000	137.000000
Duncan, Mariano	2.000000	93.000000
Dunston, Shawon	2.000000	145.000000
Durham, Leon	7.000000	127.000000
Dwyer, Jim	14.000000	39.000000
Dykstra, Len	2.000000	127.000000
Easler, Mike	13.000000	148.000000
Esasky, Nick	4.000000	76.000000
Evans, Darrell	18.000000	122.000000
Evans, Dwight	15.000000	137.000000
Fernandez, Tony	4.000000	213.000000
Fisk, Carlton	17.000000	101.000000
Fitzgerald, Mike	4.000000	59.000000
Flannery, Tim	8.000000	103.000000
Fletcher, Scott	6.000000	159.000000
Foley, Tom	4.000000	70.000000
Ford, Curt	2.000000	53.000000
Foster, George	18.000000	64.000000
Franco, Julio	5.000000	183.000000
Gaetti, Gary	6.000000	171.000000
Gagne, Greg	4.000000	118.000000
Galarraga, Andres	2.000000	87.000000
Gantner, Jim	11.000000	136.000000
Garcia, Damaso	9.000000	119.000000
Garner, Phil	14.000000	83.000000
Garvey, Steve	18.000000	142.000000

Gedman, Rich	7.000000	119.000000
Gibson, Kirk	8.000000	118.000000
Gladden, Dan	4.000000	97.000000
Grich, Bobby	17.000000	84.000000
Griffey, Ken	14.000000	150.000000
Griffin, Alfredo	11.000000	169.000000
Grubb, Johnny	15.000000	70.000000
Guillen, Ozzie	2.000000	137.000000
Gwynn, Tony	5.000000	211.000000
Hairston, Jerry	11.000000	61.000000
Hall, Mel	6.000000	131.000000
Harper, Terry	7.000000	68.000000
Harrah, Toby	17.000000	63.000000
Hassey, Ron	9.000000	110.000000
Hatcher, Billy	3.000000	108.000000
Hatcher, Mickey	8.000000	88.000000
Hayes, Von	6.000000	186.000000
Heath, Mike	9.000000	65.000000
Heep, Danny	8.000000	55.000000
Henderson, Dave	6.000000	103.000000
Henderson, Rickey	8.000000	160.000000
Hendrick, George	16.000000	77.000000
Hernandez, Keith	13.000000	171.000000
Herndon, Larry	12.000000	70.000000
Herr, Tommy	8.000000	141.000000
Hill, Donnie	4.000000	96.000000
Horner, Bob	9.000000	141.000000
Howell, Jack	2.000000	41.000000
Hrbek, Kent	6.000000	147.000000
Hubbard, Glenn	9.000000	94.000000
Hulett, Tim	4.000000	120.000000
Incaviglia, Pete	1.000000	135.000000
Iorg, Garth	8.000000	85.000000
Jackson, Reggie	20.000000	101.000000
Jacoby, Brook	5.000000	168.000000
Jeltz, Steve	4.000000	96.000000
Johnson, Cliff	15.000000	84.000000
Johnson, Howard	5.000000	54.000000
Jones, Ruppert	11.000000	90.000000
Joyner, Wally	1.000000	172.000000
Kearney, Bob	7.000000	49.000000
Kennedy, Terry	9.000000	114.000000
Kingery, Mike	1.000000	54.000000
Kingman, Dave	16.000000	118.000000
Kittle, Ron	5.000000	82.000000
Knight, Ray	11.000000	145.000000
Krenchicki, Wayne	8.000000	53.000000
Kruk, John	1.000000	86.000000
Kutcher, Randy	1.000000	44.000000
LaValliere, Mike	3.000000	71.000000
Lacy, Lee	15.000000	141.000000
Landreaux, Ken	10.000000	74.000000
Landrum, Tito	7.000000	43.000000
Lansford, Carney	9.000000	168.000000
Laudner, Tim	6.000000	47.000000
Law, Rudy	7.000000	80.000000
Law, Vance	7.000000	81.000000
Leach, Rick	6.000000	76.000000
Lemon, Chet	12.000000	101.000000
Leonard, Jeffrey	10.000000	95.000000
Lombardozzi, Steve	2.000000	103.000000
Lopes, Davey	15.000000	70.000000
Lynn, Fred	13.000000	114.000000

Madlock, Bill	14.000000	106.000000
Maldonado, Candy	6.000000	102.000000
Manning, Rick	12.000000	52.000000
Marshall, Mike	6.000000	77.000000
Martinez, Carmelo	4.000000	58.000000
Matthews, Gary	15.000000	96.000000
Mattingly, Don	5.000000	238.000000
Matuszek, Len	6.000000	52.000000
McDowell, Oddibe	2.000000	152.000000
McGee, Willie	5.000000	127.000000
McRae, Hal	18.000000	70.000000
McReynolds, Kevin	4.000000	161.000000
Meacham, Bobby	4.000000	36.000000
Melvin, Bob	2.000000	60.000000
Milner, Eddie	7.000000	110.000000
Mitchell, Kevin	2.000000	91.000000
Molitor, Paul	9.000000	123.000000
Moore, Charlie	14.000000	61.000000
Moreland, Keith	9.000000	159.000000
Moreno, Omar	12.000000	84.000000
Morrison, Jim	10.000000	147.000000
Moseby, Lloyd	7.000000	149.000000
Moses, John	5.000000	102.000000
Motley, Darryl	5.000000	46.000000
Mulliniks, Rance	10.000000	90.000000
Mumphrey, Jerry	13.000000	94.000000
Murphy, Dale	11.000000	163.000000
Murphy, Dwayne	9.000000	83.000000
Murray, Eddie	10.000000	151.000000
Nettles, Graig	20.000000	77.000000
Newman, Al	2.000000	37.000000
O'Brien, Pete	5.000000	160.000000
O'Malley, Tom	5.000000	46.000000
Oberkfell, Ken	10.000000	136.000000
Oester, Ron	9.000000	135.000000
Oglivie, Ben	16.000000	98.000000
Orsulak, Joe	4.000000	100.000000
Orta, Jorge	15.000000	93.000000
Owen, Spike	4.000000	122.000000
Paciorek, Tom	17.000000	61.000000
Pagliarulo, Mike	3.000000	120.000000
Parker, Dave	14.000000	174.000000
Parrish, Lance	10.000000	84.000000
Parrish, Larry	13.000000	128.000000
Pasqua, Dan	2.000000	82.000000
Pena, Tony	7.000000	147.000000
Pendleton, Terry	3.000000	138.000000
Perez, Tony	23.000000	51.000000
Pettis, Gary	5.000000	139.000000
Phelps, Ken	7.000000	85.000000
Phillips, Tony	5.000000	113.000000
Porter, Darrell	16.000000	41.000000
Presley, Jim	3.000000	163.000000
Puckett, Kirby	3.000000	223.000000
Puhl, Terry	10.000000	42.000000
Quinones, Rey	1.000000	68.000000
Quirk, Jamie	12.000000	47.000000
Raines, Tim	8.000000	194.000000
Ramirez, Rafael	7.000000	119.000000
Randolph, Willie	12.000000	136.000000
Ray, Johnny	6.000000	174.000000
Rayford, Floyd	6.000000	37.000000
Redus, Gary	5.000000	84.000000

Reed, Jeff	3.000000	39.000000
Reynolds, Craig	12.000000	78.000000
Reynolds, Harold	4.000000	99.000000
Reynolds, R.J.	4.000000	108.000000
Rice, Jim	13.000000	200.000000
Riles, Ernest	2.000000	132.000000
Ripken, Cal	6.000000	177.000000
Rivera, Luis	1.000000	34.000000
Roberts, Bip	1.000000	61.000000
Robidoux, Billy Jo	2.000000	41.000000
Roenicke, Ron	6.000000	68.000000
Romero, Ed	8.000000	49.000000
Rose, Pete	24.000000	52.000000
Royster, Jerry	14.000000	66.000000
Russell, Bill	18.000000	54.000000
Russell, John	3.000000	76.000000
Salas, Mark	3.000000	60.000000
Salazar, Argenis	3.000000	73.000000
Sample, Billy	9.000000	57.000000
Samuel, Juan	4.000000	157.000000
Sandberg, Ryne	6.000000	178.000000
Santana, Rafael	4.000000	86.000000
Sax, Steve	6.000000	210.000000
Schmidt, Mike	15.000000	160.000000
Schofield, Dick	4.000000	114.000000
Schroeder, Bill	4.000000	46.000000
Schu, Rick	3.000000	57.000000
Scioscia, Mike	7.000000	94.000000
Sheets, Larry	3.000000	92.000000
Shelby, John	6.000000	92.000000
Sheridan, Pat	5.000000	56.000000
Sierra, Ruben	1.000000	101.000000
Simmons, Ted	19.000000	32.000000
Skinner, Joel	4.000000	73.000000
Slaught, Don	5.000000	83.000000
Smalley, Roy	12.000000	113.000000
Smith, Lonnie	9.000000	146.000000
Smith, Ozzie	9.000000	144.000000
Snyder, Cory	1.000000	113.000000
Speier, Chris	16.000000	44.000000
Spilman, Harry	9.000000	39.000000
Stillwell, Kurt	1.000000	64.000000
Stone, Jeff	4.000000	69.000000
Strawberry, Darryl	4.000000	123.000000
Stubbs, Franklin	3.000000	95.000000
Sundberg, Jim	13.000000	91.000000
Sveum, Dale	1.000000	78.000000
Tabler, Pat	6.000000	154.000000
Tartabull, Danny	3.000000	138.000000
Templeton, Garry	11.000000	126.000000
Tettleton, Mickey	3.000000	43.000000
Teufel, Tim	4.000000	69.000000
Thomas, Andres	2.000000	81.000000
Thomas, Gorman	13.000000	59.000000
Thompson, Milt	3.000000	75.000000
Thompson, Robby	1.000000	149.000000
Thon, Dickie	8.000000	69.000000
Thornton, Andre	13.000000	92.000000
Tolleson, Wayne	6.000000	126.000000
Traber, Jim	2.000000	54.000000
Trammell, Alan	10.000000	159.000000
Trevino, Alex	9.000000	53.000000
Upshaw, Willie	8.000000	144.000000

Uribe, Jose	3.000000	101.000000
Van Slyke, Andy	4.000000	113.000000
Virgil, Ozzie	7.000000	80.000000
Walker, Greg	5.000000	78.000000
Wallach, Tim	7.000000	112.000000
Walling, Denny	12.000000	119.000000
Ward, Gary	8.000000	120.000000
Webster, Mitch	4.000000	167.000000
Whitaker, Lou	10.000000	157.000000
White, Frank	14.000000	154.000000
Whitt, Ernie	10.000000	106.000000
Wiggins, Alan	6.000000	60.000000
Wilfong, Rob	10.000000	63.000000
Wilkerson, Curt	4.000000	56.000000
Willard, Jerry	3.000000	43.000000
Williams, Reggie	2.000000	84.000000
Wilson, Glenn	5.000000	158.000000
Wilson, Mookie	7.000000	110.000000
Wilson, Willie	11.000000	170.000000
Winfield, Dave	14.000000	148.000000
Winningham, Herm	3.000000	40.000000
Wynegar, Butch	11.000000	40.000000
Wynne, Marvell	4.000000	76.000000
Young, Mike	5.000000	93.000000
Youngblood, Joel	11.000000	47.000000
Yount, Robin	13.000000	163.000000

Die empirische Korrelation zwischen X und Y betraegt -0.00803.

Anhang B

Datenmaterial zu Beispiel 6.3.1

```
/* WOMEN TRACK DATA SET: womentrack.dat                                  */

100m     200m     400m     800m    1500m    3000m   Marathon  Nation

11.61    22.94    54.50    2.15    4.43     9.79    178.52    Argentina
11.20    22.35    51.08    1.98    4.13     9.08    152.37    Australia
11.43    23.09    50.62    1.99    4.22     9.34    159.37    Austria
11.41    23.04    52.00    2.00    4.14     8.88    157.85    Belgium
11.46    23.05    53.30    2.16    4.58     9.81    169.98    Bermuda
11.31    23.17    52.80    2.10    4.49     9.77    168.75    Brazil
12.14    24.47    55.00    2.18    4.45     9.51    191.02    Burma
11.00    22.25    50.06    2.00    4.06     8.81    149.45    Canada
12.00    24.52    54.90    2.05    4.23     9.37    171.38    Chile
11.95    24.41    54.97    2.08    4.33     9.31    168.48    China
11.60    24.00    53.26    2.11    4.35     9.46    165.42    Columbia
12.90    27.10    60.40    2.30    4.84    11.10    233.22    Cookis
11.96    24.60    58.25    2.21    4.68    10.43    171.80    Costa
11.09    21.97    47.99    1.89    4.14     8.92    158.85    Czech
11.42    23.52    53.60    2.03    4.18     8.71    151.75    Denmark
11.79    24.05    56.05    2.24    4.74     9.89    203.88    Dominican
11.13    22.39    50.14    2.03    4.10     8.92    154.23    Finland
11.15    22.59    51.73    2.00    4.14     8.98    155.27    France
10.81    21.71    48.16    1.93    3.96     8.75    157.68    GDR
11.01    22.39    49.75    1.95    4.03     8.59    148.53    FRG
11.00    22.13    50.46    1.98    4.03     8.62    149.72    GB&NI
11.79    24.08    54.93    2.07    4.35     9.87    182.20    Greece
11.84    24.54    56.09    2.28    4.86    10.54    215.08    Guatemal
11.45    23.06    51.50    2.01    4.14     8.98    156.37    Hungary
11.95    24.28    53.60    2.10    4.32     9.98    188.03    India
11.85    24.24    55.34    2.22    4.61    10.02    201.28    Indonesia
11.43    23.51    53.24    2.05    4.11     8.89    149.38    Ireland
11.45    23.57    54.90    2.10    4.25     9.37    160.48    Israel
11.29    23.00    52.01    1.96    3.98     8.63    151.82    Italy
11.73    24.00    53.73    2.09    4.35     9.20    150.50    Japan
11.73    23.88    52.70    2.00    4.15     9.20    181.05    Kenya
11.96    24.49    55.70    2.15    4.42     9.62    164.65    Korea
12.25    25.78    51.20    1.97    4.25     9.35    179.17    DPRKorea
12.03    24.96    56.10    2.07    4.38     9.64    174.68    Luxembou
12.23    24.21    55.09    2.19    4.69    10.46    182.17    Malasiya
11.76    25.08    58.10    2.27    4.79    10.90    261.13    Mauritius
11.89    23.62    53.76    2.04    4.25     9.59    158.53    Mexico
11.25    22.81    52.38    1.99    4.06     9.01    152.48    Netherlands
11.55    23.13    51.60    2.02    4.18     8.76    145.48    NZealand
11.58    23.31    53.12    2.03    4.01     8.53    145.48    Norway
```

```
12.25    25.07    56.96    2.24    4.84    10.69    233.00    Guinea
11.76    23.54    54.60    2.19    4.60    10.16    200.37    Philippi
11.13    22.21    49.29    1.95    3.99     8.97    160.82    Poland
11.81    24.22    54.30    2.09    4.16     8.84    151.20    Portugal
11.44    23.46    51.20    1.92    3.96     8.53    165.45    Rumania
12.30    25.00    55.08    2.12    4.52     9.94    182.77    Singapore
11.80    23.98    53.59    2.05    4.14     9.02    162.60    Spain
11.16    22.82    51.79    2.02    4.12     8.84    154.48    Sweden
11.45    23.31    53.11    2.02    4.07     8.77    153.42    Switzerl
11.22    22.62    52.50    2.10    4.38     9.63    177.87    Taipei
11.75    24.46    55.80    2.20    4.72    10.28    168.45    Thailand
11.98    24.44    56.45    2.15    4.37     9.38    201.08    Turkey
10.79    21.83    50.62    1.96    3.95     8.50    142.72    USA
11.06    22.19    49.19    1.89    3.87     8.45    151.22    USSR
12.74    25.85    58.73    2.33    5.81    13.04    306.00    WSamoa

/* Source: IAAF/ATFS Track and Field Statistics Handbook for the 1984   */
/* Los Angeles Olympics. Reproduced by permission of the International  */
/* Amateur Athletics Federation.                                        */
```

Anhang C

Quantile der $\mathcal{N}$(0,1)-Verteilung zu Kapitel 6.5.3

Der in der Folge angegebene kritische Wert Q_α für die Teststatistik Q_i aus Kapitel 6.5.3 wurde der Arbeit von J. Edward Jackson [12] entnommen.

Hierzu bezeichnen $\lambda_j,\ j = 1, \ldots, m$ die der Größe nach absteigend geordneten Eigenwerte der Korrelationsmatrix und $Y = (y_{i,j}),\ i = 1, \ldots, n;\ j = 1, \ldots, m;\ Y \in \mathbb{R}^{n \times m}$ die Score-Matrix der Originalvariablen $(X_1, \ldots, X_m)$.

Ferner wird die Verteilungsfunktion der Standardnormalverteilung mit $\Phi(\cdot)$ bezeichnet.

$$Q_i \quad = \quad \sum_{k=p+1}^{m} y_{i,k}^2 \text{ bei } p \text{ einzubeziehenden Hauptkomponenten}$$

$$\begin{aligned}
\theta_1 &:= \sum_{k=p+1}^{m} \lambda_k \\
\theta_2 &:= \sum_{k=p+1}^{m} \lambda_k^2 \\
\theta_3 &:= \sum_{k=p+1}^{m} \lambda_k^3 \\
h_0 &:= 1 - \frac{2 \cdot \theta_1 \cdot \theta_3}{3 \cdot \theta_2^2} \\
c_\alpha &:= sgn(h_0) \cdot \Phi^{-1}(\sqrt[n]{1-\alpha}) \\
Q_\alpha &:= \theta_1 \cdot \left[\frac{c_\alpha \cdot \sqrt{2 \cdot \theta_2 \cdot h_0^2}}{\theta_1} + \frac{\theta_2 \cdot h_0 (h_0 - 1)}{\theta_1^2} + 1 \right]^{\frac{1}{h_0}}
\end{aligned}$$

$$\Rightarrow \mathbb{P}(Q_i > Q_\alpha) \underset{\text{approx.}}{=} \alpha.$$

Literaturverzeichnis

[1] Die zufälligen Merkmale der Statistiker, in: *Die Hochschullehrer, 2.* S. 65, 1991

[2] COLLIER BOOKS: *The 1987 Baseball Encyclopedia Update.* Macmillan Publishing Company, New York, 1987

[3] PROF. DR. JOACHIM HARTUNG UND DR. BARBARA HEINE: *Statistik-Übungen, Deskriptive Statistik.* 6., unwesentlich veränderte Auflage

[4] BARNETT, V. UND LEWIS, T.: *Outliers in statistical data.* Wiley, 1979

[5] D.C. HOAGLIN, F. MOSTELLER, J.W. TUKEY: *Understanding robust and exploratory data analysis.* Kapitel 4: Transforming data

[6] *Mathematical models in molecular and cellular biology edited by Lee A. Segel.* Cambridge University Press, 1980

[7] P. ÉRDI AND J. TÓTH: *Mathematical models of chemical reactions. Theory and applications of deterministic and stochastic models.* Manchester University Press, 1989

[8] E. L. LEHMANN: *Nonparametrics. Statistical methods based on ranks.* Holden-Day, Inc., San Francisco, 1975

[9] I. T. JOLLIFFE: *Principal component analysis.* Springer series in statistics, Springer-Verlag New York Inc., 1986

[10] GERHARD DIKTA: *Operations Research.* Skript zum Kurs, FH Aachen, Abteilung Jülich, 2002

[11] RAVINDRA KHATTREE, DAYANAND N. NAIK: *Multivariate Data Reduction and Discrimination with SAS Software.* SAS Institute Inc., Cary, North Carolina, 2000

[12] J. EDWARD JACKSON: *A user's guide to principal components.* Wiley series in probability and mathematical statistics, John Wiley & Sons, Inc., 1991

[13] I. T. JOLLIFFE: *Discarding Variables in a Pricipal Component Analysis.* Journal of the Royal Statistical Society, Vol. 21, S. 160 ff., 1972

[14] G. P. MCCABE: *Principal variables.* Technometrics 26, S. 137-144, 1984