Statistik und ihre Anwendungen

Andreas Handl

Multivariate Analysemethoden

Theorie und Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-PLUS

Springer

Andreas Handl

Multivariate Verfahren

Theorie und Praxis multivariater Verfahren unter besonderer

Berücksichtigung von S-PLUS

SPIN Springer's internal project number, if known

Monograph – Mathematics –

20th September 2002

Springer
Berlin Heidelberg New York
Barcelona Hong Kong
London Milan Paris
Tokyo

Für Claudia und Fabian

Vorwort

In den letzten 20 Jahren hat die starke Verbreitung von leistungsfähigen Rechnern unter anderem dazu geführt, dass riesige Datenmengen gesammelt werden, in denen sowohl unter den Objekten als auch den Merkmalen Strukturen gesucht werden. Geeignete Werkzeuge hierzu bieten multivariate Verfahren. Außerdem erhöhte sich durch die Verbreitung der Computer auch die Verfügbarkeit leistungsfähiger Programme zur Analyse multivariater Daten. Statistische Programmpakete wir `SAS`, `SPSS` und `BMDP` laufen auch auf PCs. Daneben wurde eine Reihe von Umgebungen zur Datenanalyse wie `S-PLUS`, `R` und `GAUSS` geschaffen, die nicht nur eine Vielzahl von Funktionen zur Verfügung stellen, sondern in denen auch neue Verfahren schnell implementiert werden können.

Dieses Buch gibt eine Einführung in die Analyse multivariater Daten, die die eben beschriebenen Aspekte berücksichtigt. Jedes Verfahren wird zunächst anhand eines realen Problems motiviert. Darauf aufbauend wird ausführlich die Zielsetzung des Verfahrens herausgearbeitet. Es folgt eine detaillierte Entwicklung der Theorie. Praktische Aspekte runden die Darstellung des Verfahrens ab. An allen Stellen wird die Vorgehensweise anhand realer Datensätze veranschaulicht. Abschließend wird beschrieben, wie das Verfahren in `S-PLUS` durchzuführen ist beziehungsweise wie `S-PLUS` entsprechend erweitert werden kann, wenn das Verfahren nicht implementiert ist.

Das Buch wendet sich zum einen an Studierende des Fachs Statistik im Hauptstudium, die die multivariaten Verfahren sowie deren Durchführung beziehungsweise Implementierung in `S-PLUS` kennenlernen möchten. Es richtet sich zum anderen aber auch an Personen in Wissenschaft und Praxis, die im Rahmen von Diplomarbeiten, Dissertationen und Projekten Datenanalyse betreiben und hierbei multivariate Verfahren unter Zuhilfenahme von `S-PLUS` anwenden möchten. Dabei sind grundsätzlich die Ausführungen so gehalten und die Beispiele derart gewählt, dass sie für die Anwender unterschiedlichster Fachrichtungen interessant sind.

Einige Grundlagen wie Maximum-Likelihood und Testtheorie werden vorausgesetzt. Diese werden zum Beispiel in Schlittgen (2000) und Fahrmeir et al. (2001) dargelegt. Andere grundlegende Aspekte werden aber auch in diesem Buch entwickelt. So findet man in Kapitel 2 einen großen Teil der univariaten Datenanalyse und in Kapitel 3 einige Aspekte von univariaten Zu-

fallsvariablen. Die im Buch benötigte Theorie mehrdimensionaler Zufallsvariablen wird in Kapitel 3 detailliert herausgearbeitet. Um diese und weitere Kapitel verstehen zu können, benötigt man Kenntnisse aus der Linearen Algebra. Deshalb werden im Anhang A.1 die zentralen Begriffe und Zusammenhänge der Linearen Algebra beschrieben und exemplarisch verdeutlicht. Außerdem ist Literatur angegeben, in der die Beweise und Zusammenhänge ausführlich betrachtet werden.

Es ist unmöglich, alle multivariaten Verfahren in einem Buch darzustellen. Ich habe die Verfahren so ausgewählt, dass ein Überblick über die breiten Anwendungsmöglichkeiten multivariater Verfahren gegeben wird. Dabei versuche ich die Verfahren so darzustellen, dass anschließend die Spezialliteratur zu jedem der Gebiete gelesen werden kann. Das Buch besteht aus 4 Teilen. Im ersten Teil werden die Grundlagen gelegt, während in den anderen Teilen unterschiedliche Anwendungsaspekte berücksichtigt werden. Bei einem hochdimensionalen Datensatz kann man an den Objekten oder den Merkmalen interessiert sein. Im zweiten Teil werden deshalb Verfahren vorgestellt, die dazu dienen, die Objekte in einem Raum niedriger Dimension darzustellen. Außerdem wird die Procrustes-Analyse beschrieben, die einen Vergleich unterschiedlicher Konfigurationen erlaubt. Der dritte Teil beschäftigt sich mit Abhängigkeitsstrukturen zwischen Variablen. Hier ist das Modell der bedingten Unabhängigkeit von großer Bedeutung. Im letzten Teil des Buches werden Daten mit Gruppenstruktur betrachtet. Am Ende fast aller Kapitel sind Aufgaben zu finden. Die Lösungen zu den Aufgaben sowie die im Buch verwendeten Datensätze und `S-PLUS`-Funktionen sind auf der Internet-Seite des Springer-Verlages zu finden.

In diesem Buch spielt der Einsatz des Rechners bei der Datenanalyse eine wichtige Rolle. Programmpakete entwickeln sich sehr schnell, sodass das heute Geschriebene oft schon morgen veraltet ist. Um dies zu vermeiden, beschränke ich mich auf den Kern von `S-PLUS`, wie er schon in der Version 3 vorhanden war. Den Output habe ich mit Version 4.5 erstellt. Ich stelle also alles im Befehlsmodus dar. Dies hat aus meiner Sicht einige Vorteile. Zum einen lernt man so, wie man das System schnell um eigene Funktionen erweitern kann. Zum anderen kann man die Funktionen in nahezu allen Fällen auch in `R` ausführen, das man sich kostenlos im Internet unter `http://cran.r-project.org/` herunterladen kann. Informationen zum Bezug von S-Plus für Studenten findet man im Internet unter `http://elms03.e-academy.com/splus/`. Das Buch enthält keine getrennte Einführung in `S-PLUS`. Vielmehr werden im Kapitel 2.3 anhand der elementaren Datenbehandlung die ersten Schritte in `S-PLUS` gezeigt. Dieses Konzept hat sich in Lehrveranstaltungen als erfolgreich erwiesen. Nachdem man dieses Kapitel durchgearbeitet hat, sollte man sich dann Kapitel A.3 widmen, in dem gezeigt wird, wie man die Matrizenrechnung in `S-PLUS` umsetzt. Bei der Erstellung eigener Funktionen benötigt man diese Kenntnisse. Ansonsten bietet es sich an, einen Blick in die Lehrbuchliteratur zu werfen.

Hier sind Süselbeck (1993), Krause & Olson (2000) und Venables & Ripley (1999) zu empfehlen.

Das Buch ist aus Skripten entstanden, die ich seit Mitte der Achtziger Jahre zu Vorlesungen an der Freien Universität Berlin und der Universität Bielefeld angefertigt habe. Ich danke an erster Stelle Herrn Prof. Dr. Herbert Büning von der Freien Universität Berlin, der mich ermutigt und unterstützt hat, aus meinem Skript ein Lehrbuch zu erstellen. Er hat Teile des Manuskripts gelesen und korrigiert und mir sehr viele wertvolle Hinweise gegeben. Dankbar bin ich auch Herrn Dipl.-Volkswirt Wolfgang Lemke von der Universität Bielefeld, der die Kapitel über Regressionsanalyse und insbesondere Faktorenanalyse durch seine klugen Fragen und Anmerkungen bereichert hat. Ebenfalls danken möchte ich Herrn Dr. Stefan Niermann, der das Skript schon seit einigen Jahren in seinen Lehrveranstaltungen an der Universität Hannover verwendet und einer kritischen Würdigung unterzogen hat.

Herrn Andreas Schleicher von der OECD in Paris danke ich für die Genehmigung, die Daten der PISA-Studie zu verwenden. Herrn Prof. Dr. Wolfgang Härdle von der Humboldt-Universität zu Berlin und Herrn Prof. Dr. Holger Dette von der Ruhr-Universität Bochum danke ich, dass sie das Buch in ihre Reihe aufgenommen haben. Vom Springer-Verlag erhielt ich jede nur denkbare Hilfe bei der Erstellung der druckreifen Version. Herr Holzwarth vom Springer-Verlag fand für jedes meiner LATEX-Probleme sofort eine Lösung und Frau Kehl gab mir viele wichtige Hinweise in Bezug auf das Layout.

Abschließend möchte ich an Herrn Professor Dr. Bernd Streitberg erinnern, der ein großartiger Lehrer war. Er konnte schwierige Zusammenhänge einfach veranschaulichen und verstand es, Studenten und Mitarbeiter für die Datenanalyse zu begeistern. Auch ihm habe ich sehr viel zu verdanken.

Bielefeld, im Juni 2002 Andreas Handl

Table of Contents

Part I

Grundlagen

1 Beispiele multivariater Datensätze

Ausgangspunkt jeder statistischen Analyse ist ein Problem. Um dieses Problem zu lösen, werden entweder Daten erhoben oder es wird auf vorhandene Datenbestände zurückgegriffen. Für jedes der n Objekte liegen die Ausprägungen von p Merkmalen vor. Das Objekt kann natürlich auch eine Person sein. Ist p gleich 1, so spricht man von univariater, ansonsten von multivariater Datenanalyse. Bei der Analyse der Daten kann man entweder explorativ oder konfirmatorisch vorgehen. Im ersten Fall sucht man gezielt nach Strukturen, während man im zweiten Fall von einer Hypothese oder mehreren Hypothesen ausgeht, die man überprüfen will. Die Fragestellung kann sich dabei auf die Objekte oder die Merkmale beziehen. Wir werden in diesem Buch sowohl explorative als auch konfirmatorische multivariate Verfahren beschreiben. Ein wichtiges Anliegen dieses Buches ist es, die Problemstellung und Vorgehensweise multivariater Verfahren anhand von realen Datensätzen zu illustrieren. Im ersten Kapitel wollen wir uns darauf einstimmen und uns bereits vorab einige der verwendeten Datensätze sowie die sich daraus ergebenden Fragestellungen ansehen.

Example 1. Ende des Jahres 2001 wurden die Ergebnisse der sogenannten PISA-Studie von Deutsches PISA-Konsortium (Hrsg.) (2001) veröffentlicht. In dieser Studie wurden die Merkmale `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` getestet. In Tabelle 1.1 sind die Mittelwerte der Punkte, die von den Schülern in den einzelnen Ländern erreicht wurden, zu finden.

□

Dieser Datensatz beinhaltet ausschließlich quantitative Merkmale. Wir wollen ihn im Kapitel 2 im Rahmen der elementaren Datenanalyse mit einfachen Hilfsmitteln beschreiben. Ziel ist es, die Verteilung jedes Merkmals graphisch darzustellen und durch geeignete Maßzahlen zu strukturieren. Außerdem wollen wir dort auch Abhängigkeitsstrukturen zwischen zwei Merkmalen analysieren. Darüber hinaus findet der Datensatz insbesondere auch Anwendung im Rahmen von Kapitel 5.

Example 2. Bei Studienanfängern am Fachbereich Wirtschaftswissenschaft der FU Berlin wurde im Wintersemester 1988/89 ein Test zur Mittelstufenalgebra mit 26 Aufgaben durchgeführt. Neben dem Merkmal `Geschlecht`

Table 1.1. Mittelwerte der Punkte in den Bereichen Lesekompetenz, Mathematische Grundbildung und Naturwissenschaftliche Grundbildung im Rahmen der PISA-Studie, vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S. 107, 173, 229

Land	Lesekompetenz	Mathematische Grundbildung	Naturwissenschaftliche Grundbildung
Australien	528	533	528
Belgien	507	520	496
Brasilien	396	334	375
Dänemark	497	514	481
Deutschland	484	490	487
Finnland	546	536	538
Frankreich	505	517	500
Griechenland	474	447	461
Großbritannien	523	529	532
Irland	527	503	513
Island	507	514	496
Italien	487	457	478
Japan	522	557	550
Kanada	534	533	529
Korea	525	547	552
Lettland	458	463	460
Liechtenstein	483	514	476
Luxemburg	441	446	443
Mexiko	422	387	422
Neuseeland	529	537	528
Norwegen	505	499	500
Österreich	507	515	519
Polen	479	470	483
Portugal	470	454	459
Russland	462	478	460
Schweden	516	510	512
Schweiz	494	529	496
Spanien	493	476	491
Tschechien	492	498	511
Ungarn	480	488	496
USA	504	493	499

mit den Ausprägungsmöglichkeiten `w` und `m` wurde noch eine Reihe weiterer Merkmale erhoben. Die Studenten wurden gefragt, ob sie den Leistungskurs Mathematik besucht haben und ob sie im Jahr 1988 das Abitur gemacht haben. Diese Merkmale bezeichnen wir mit `MatheLK` und `Abitur88`. Bei beiden Merkmalen gibt es die Ausprägungsmöglichkeiten `j` und `n`. Außerdem sollten sie ihre Abiturnote in Mathematik angeben. Dieses Merkmal bezeichnen wir mit `MatheNote`. Das Merkmal `Punkte` gibt die Anzahl der im Test richtig gelösten Aufgaben an. Die Daten sind in Tabelle 1.2 zu finden.

Table 1.2. Ergebnisse von Studienanfängern bei einem Mathematik-Test

Geschlecht	MatheLK	MatheNote	Abitur88	Punkte
m	n	3	n	8
m	n	4	n	7
m	n	4	n	4
m	n	4	n	2
m	n	3	n	7
w	n	3	n	6
w	n	4	j	3
w	n	3	j	7
w	n	4	j	14
m	j	3	n	19
m	j	3	n	15
m	j	2	n	17
m	j	3	n	10
w	j	3	n	22
w	j	2	n	23
w	j	2	n	15
m	j	1	j	21
w	j	2	j	10
w	j	2	j	12
w	j	4	j	17

□

Dieser Datensatz enthält auch qualitative Merkmale. Diese wollen wir ebenfalls im Kapitel 2 geeignet darstellen. Außerdem hat der Datensatz wesentliche Bedeutung im Rahmen des Kapitels 12.

Example 3. Im Wintersemester `1996/97` wurden an der Fakultät für Wirtschaftswissenschaften der Universität Bielefeld 265 Erstsemesterstudenten in der Statistik I Vorlesung befragt. Neben dem Merkmal `Geschlecht` mit den Ausprägungsmöglichkeiten `w` und `m` wurden die Merkmale `Gewicht`, `Alter` und `Größe` erhoben. Außerdem wurden die Studenten gefragt, ob sie rauchen und ob sie ein Auto besitzen. Diese Merkmale bezeichnen wir mit `Raucher` und `Auto`. Auf einer Notenskala von 1 bis 5 sollten sie angeben, wie ihnen Cola schmeckt. Das Merkmal bezeichnen wir mit `Cola`. Als letztes wurde noch gefragt, ob die Studenten den Leistungskurs Mathematik besucht haben. Dieses Merkmal bezeichnen wir mit `MatheLK`. Tabelle 1.3 gibt die Ergebnisse von 5 Studenten wieder.

□

Ziel einer multivariaten Analyse dieses Datensatzes wird es sein, Ähnlichkeiten zwischen den Studenten festzustellen. Wir wollen uns mit solchen Ähnlichkeits- und Distanzmaßen im Kapitel 4 beschäftigen.

Table 1.3. Ergebnis der Befragung von 5 Erstsemesterstudenten

Geschlecht	Alter	Größe	Gewicht	Raucher	Auto	Cola	MatheLK
m	23	171	60	n	j	2	j
m	21	187	75	n	j	1	n
w	20	180	65	n	n	3	j
w	20	165	55	j	n	2	j
m	23	193	81	n	n	3	n

Example 4. Von den Studienanfängern des Wintersemesters 1995/96 an der Fakultät für Wirtschaftswissenschaften der Universität Bielefeld haben 17 Studenten alle 16 Klausuren nach vier Semestern im ersten Anlauf bestanden. Tabelle 1.4 zeigt die Durchschnittsnoten dieser Studenten in den vier Bereichen `Mathematik`, `BWL`, `VWL` und `Methoden`.

Table 1.4. Noten von Studenten in Fächern

Mathematik	BWL	VWL	Methoden	Mathematik	BWL	VWL	Methoden
1.325	1.000	1.825	1.750	2.500	3.250	3.075	2.250
2.000	1.250	2.675	1.750	1.675	2.500	2.675	1.250
3.000	3.250	3.000	2.750	2.075	1.750	1.900	1.500
1.075	2.000	1.675	1.000	1.750	2.000	1.150	1.250
3.425	2.000	3.250	2.750	2.500	2.250	2.425	2.500
1.900	2.000	2.400	2.750	1.675	2.750	2.000	1.250
3.325	2.500	3.000	2.000	3.675	3.000	3.325	2.500
3.000	2.750	3.075	2.250	1.250	1.500	1.150	1.000
2.075	1.250	2.000	2.250				

□

Es handelt sich hierbei um einen hochdimensionalen Datensatz. Ziel wird es deshalb sein, die Einzelnoten der Studenten zu einer Gesamtnote zusammenzufassen, also die Objekte in einem Raum niedriger Dimension darzustellen. Im Rahmen der Hauptkomponentenanalyse in Kapitel 5 werden wir sehen, dass es für diesen Zweck neben der Mittelwertbildung andere Verfahren gibt, die wesentlichen Merkmalen ein größeres Gewicht beimessen.

Example 5. In Tabelle 1.5 sind die Luftlinienentfernungen zwischen deutschen Städten in Kilometern angegeben. Für die Namen der Städte wurden die Autokennzeichen verwendet.

□

Anhand dieses Datensatzes werden wir mit Hilfe der mehrdimensionalen Skalierung in Kapitel 6 zeigen, wie sich aus den Distanzen Konfigurationen gewinnen lassen. Im Beispiel wäre das eine Landkarte Deutschlands.

Table 1.5. Luftlinienentfernungen in Kilometern zwischen deutschen Städten

	HH	B	K	F	M
HH	0	250	361	406	614
B	250	0	475	432	503
K	361	475	0	152	456
F	406	432	152	0	305
M	614	503	456	305	0

Example 6. In der Süddeutschen Zeitung vom 18.12.2001 wurden Reisezeiten verglichen, die mit unterschiedlichen Verkehrsmitteln innerhalb Deutschlands benötigt werden. Dabei wurde die Reisezeit von Innenstadt zu Innenstadt betrachtet. Dies hat zur Konsequenz, dass bei der Bahn 40 Minuten zur reinen Reisezeit addiert wurden. Die Tabellen 1.6 und 1.7 zeigen die benötigten Reisezeiten für Pkws und die Bahn zwischen ausgewählten Städten.

Table 1.6. Reisezeiten (in Minuten) mit dem Pkw zwischen deutschen Städten

	HH	B	K	F	M
HH	0	192	271	314	454
B	192	0	381	365	386
K	271	381	0	134	295
F	314	365	134	0	251
M	454	386	295	251	0

Table 1.7. Reisezeiten (in Minuten) mit der Bahn zwischen deutschen Städten

	HH	B	K	F	M
HH	0	184	247	254	409
B	184	0	297	263	433
K	247	297	0	175	385
F	254	263	175	0	257
M	409	433	385	257	0

□

Auch in diesem Beispiel lassen sich Konfigurationen gewinnen. Für jedes Verkehrsmittel lässt sich eine zweidimensionale Darstellung der Städte ermitteln. Das Problem besteht allerdings in der Vergleichbarkeit der Darstellungen, da die Konfigurationen verschoben, gedreht, gestaucht oder gestreckt werden können, ohne dass sich die Verhältnisse der Distanzen ändern. Mit der

Procrustes-Analyse in Kapitel 7 wollen wir deshalb ein Verfahren darstellen, mit dessen Hilfe Konfigurationen ähnlich gemacht werden können.

Example 7. In der Süddeutschen Zeitung wurden Ende Juli 1999 im Anzeigenteil 33 VW-Golf 3 angeboten. In Tabelle 1.8 sind deren Merkmale `Alter` in Jahren, `Gefahrene Kilometer` (in tausend) und `Angebotspreis` (in DM) zu finden.

Table 1.8. Alter, Gefahrene Kilometer und Angebotspreis von 33 VW-Golf 3

Alter	Gefahrene Kilometer	Angebotspreis	Alter	Gefahrene Kilometer	Angebotspreis
2	15	21800	5	78	15900
2	66	18800	5	55	16900
2	29	20500	5	106	14800
3	40	18900	5	30	15500
3	68	21200	5	27	16500
3	37	16800	5	83	14900
3	60	17500	5	75	12400
3	26	23800	6	53	12800
3	58	16800	6	70	14900
4	96	14500	6	94	12900
4	60	19900	6	86	12800
4	69	15900	6	70	13500
4	44	17900	7	121	10950
4	37	19500	7	78	12900
4	46	16000	7	104	10800
5	70	16500	7	95	11600
5	90	15800			

□

In den bisherigen Beispielen standen die Objekte im Mittelpunkt des Interesses. Im Beispiel 7 interessieren uns die Abhängigkeitsstrukturen zwischen Merkmalen. Wir wollen wissen, inwieweit der Angebotspreis vom Alter und dem Kilometerstand des PKW abhängt. In Kapitel 8 werden wir Abhängigkeitsstrukturen durch ein Regressionsmodell beschreiben.

Example 8. Bödeker & Franke (2001) beschäftigen sich in Ihrer Diplomarbeit mit den Möglichkeiten und Grenzen von Virtual-Reality-Technologien auf industriellen Anwendermärkten. Hierbei führten sie eine Befragung bei Unternehmen durch, in der sie unter anderem den Nutzen ermittelten, den Unternehmen von einem Virtual-Reality-System erwarten. Auf einer Skala von 1 bis 5 sollte dabei angegeben werden, wie wichtig die Merkmale `Veranschaulichung von Fehlfunktionen`, `Ermittlung von Kundenanforderungen`, `Angebotserstellung`, `Qualitätsverbesserung`, `Kostenre-`

`duktion` und `Entwicklungszeitverkürzung` sind. 508 Unternehmen bewerteten alle sechs Aspekte. In Tabelle 1.9 sind die Korrelationen zwischen den Merkmalen zu finden. Dabei wird `Veranschaulichung von Fehlfunktionen` durch `Fehler`, `Ermittlung von Kundenanforderungen` durch `Kunden`, `Angebotserstellung` durch `Angebot`, `Qualitätsverbesserung` durch `Qualität`, `Entwicklungszeitverkürzung` durch `Zeit` und `Kostenreduktion` durch `Kosten` abgekürzt.

Table 1.9. Korrelationen zwischen Merkmalen

	Fehler	Kunden	Angebot	Qualität	Zeit	Kosten
Fehler	1.000	0.223	0.133	0.625	0.506	0.500
Kunden	0.223	1.000	0.544	0.365	0.320	0.361
Angebot	0.133	0.544	1.000	0.248	0.179	0.288
Qualität	0.625	0.365	0.248	1.000	0.624	0.630
Zeit	0.506	0.32	0.179	0.624	1.000	0.625
Kosten	0.500	0.361	0.288	0.630	0.625	1.000

□

Bei vielen Anwendungen werden die Merkmale gleich behandelt. Im Zusammmenhang mit dem Beispiel 8 lassen sich Korrelationen zwischen mehreren Merkmalen durch sogenannte Faktoren erklären. Es fällt auf, dass alle Korrelationen positiv sind. Außerdem gibt es Gruppen von Merkmalen, zwischen denen hohe Korrelationen existieren, während die Korrelationen mit den anderen Merkmalen niedrig sind. Diese Struktur der Korrelationen soll durch unbeobachtbare Variablen, die Faktoren genannt werden, erklärt werden. Hiermit werden wir uns im Rahmen der Faktorenanalyse in Kapitel 9 beschäftigen.

Example 9. Bei einer Befragung von Studienanfängern wurden die Merkmale `Geschlecht`, `Studienfach` mit den Ausprägungen `BWL` und `VWL` und `Wahlverhalten` mit den Ausprägungen `CDU` und `SPD` erhoben. Die Kontingenztabelle mit den absoluten Häufigkeiten für die Merkmale `Wahlverhalten` und `Studienfach` bei den Frauen ist in Tabelle 1.10, bei den Männern in Tabelle 1.11 zu finden.

Table 1.10. Studienfach und Wahlverhalten bei den Studentinnen

Studienfach \ Wahlverhalten	CDU	SPD
BWL	4	12
VWL	2	2

Table 1.11. Studienfach und Wahlverhalten bei den Studenten

Studienfach \ Wahlverhalten	CDU	SPD
BWL	46	24
VWL	4	6

□

Wie bereits im Beispiel 7 interessieren uns auch im Beispiel 9 die Abhängigkeitsstrukturen zwischen den Merkmalen. Allerdings handelt es sich im Beispiel 9 nicht um quantitative, sondern um qualitative Merkmale. Welche Abhängigkeitsstrukturen zwischen mehreren qualitativen Merkmalen bestehen können und wie man sie durch geeignete Modelle beschreiben kann, werden wir im Kapitel 10 über loglineare Modelle sehen.

Example 10. Im Rahmen der im Beispiel 1 auf Seite 3 beschriebenen PISA-Studie wurde auch der Zeitaufwand der Schüler für Hausaufgaben erhoben (vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S.417). Dort wird unterschieden zwischen sehr geringem, geringem, mittlerem, großem und sehr großem Aufwand. Wir fassen die Länder mit sehr geringem und geringem Aufwand und die Länder mit großem und sehr großem Aufwand zusammen. Wir wollen vergleichen, ob sich die Verteilung des Merkmals `Mathematische Grundbildung` in den drei Gruppen unterscheidet. Wir sind aber auch daran interessiert, ob sich die drei Merkmale `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` in den drei Gruppen unterscheiden. □

In diesem Beispiel liegen die Daten in Form von Gruppen vor, wobei die Gruppen bekannt sind. Die Gruppenunterschiede werden wir mit einer einfaktoriellen Varianzanalyse im Kapitel 11 bestimmen.

Example 11. Lasch & Edel (1994) betrachten 127 Zweigstellen eines Kreditinstituts in Baden-Württemberg und bilden auf der Basis einer Vielzahl von Merkmalen 9 Gruppen von Zweigstellen. In Tabelle 1.12 sind die Merkmale `Einwohnerzahl` und jährliche `Gesamtkosten` in tausend DM für 20 dieser Zweigstellen zusammengestellt.

Unter diesen Zweigstellen gibt es zwei Typen. Die ersten 14 Zweigstellen haben einen hohen Marktanteil und ein überdurchschnittliches Darlehens- und Kreditgeschäft. Die restlichen 6 Zweigstellen sind technisch gut ausgestattet, besitzen ein überdurchschnittliches Einlage- und Kreditgeschäft und eine hohe Mitarbeiterzahl. Es soll nun eine Entscheidungsregel angegeben werden, mit der man auf der Basis der Werte der Merkmale `Einwohnerzahl` und `Gesamtkosten` eine neue Zweigstelle einer der beiden Gruppen zuordnen kann. □

Table 1.12. Eigenschaften von 20 Zweigstellen eines Kreditinstituts in Baden-Württemberg

Filiale	Einwohner	Gesamtkosten	Filiale	Einwohner	Gesamtkosten
1	1642	478.2	11	3504	413.8
2	2418	247.3	12	5431	379.7
3	1417	223.6	13	3523	400.5
4	2761	505.6	14	5471	404.1
5	3991	399.3	15	7172	499.4
6	2500	276.0	16	9419	674.9
7	6261	542.5	17	8780	468.6
8	3260	308.9	18	5070	601.5
9	2516	453.6	19	8780	578.8
10	4451	430.2	20	8630	641.5

Auch hier liegen bekannte Gruppen vor. Die im Beispiel 11 angesprochene Entscheidungsregel werden wir im Kapitel 12 mit Hilfe der Verfahren der Diskriminanzanalyse ermitteln.

Example 12. Brühl & Kahn (2001) betrachten in Ihrer Diplomarbeit unter anderem sechs Regionen Deutschlands und bestimmen für jede der Regionen eine Reihe von Merkmalen. Das Merkmal `Bev` ist die absolute Bevölkerungszahl (in tausend Einwohner) der Region, während das Merkmal `BevOZ` die Bevölkerungszahl (in tausend Einwohner) im Oberzentrum und das Merkmal `BevUmland` die Bevölkerungsdichte (in Einwohner je Quadratkilometer) im Umland angibt. Das Merkmal `Luft` gibt die durchschnittliche Flugzeit zu allen 41 europäischen Agglomerationsräumen in Minuten an. Das Merkmal `PKW` gibt die durchschnittliche PKW-Fahrzeit zu den nächsten drei Agglomerationsräumen in Minuten an. Das Merkmal `IC` gibt die PKW-Fahrzeit zum nächsten IC-Systemhalt des Kernnetzes in Minuten an. Tabelle 1.13 zeigt die Ausprägungen der Merkmale in den sechs Regionen.

Table 1.13. Merkmale von 6 Regionen in Deutschland

	Bev	BevOZ	Luft	PKW	IC	BevUmland
Münster	1524.8	265.4	272	79	24	223.5
Bielefeld	1596.9	323.6	285	87	23	333.9
Duisburg/Essen	2299.7	610.3	241	45	9	632.1
Bonn	864.1	303.9	220	53	11	484.7
Rhein-Main	2669.9	645.5	202	61	15	438.6
Düsseldorf	2985.2	571.2	226	45	16	1103.9

□

Im Unterschied zu den Beispielen 10 und 11 wollen wir hier die Gruppen erst noch bilden. Es sollen Gruppen von Regionen so gebildet werden, dass die Regionen in einer Gruppe ähnlich sind, während die Gruppen sich unterscheiden. Möglich ist das mit dem Verfahren der Clusteranalyse, die Gegenstand des Kapitels 13 ist.

2 Elementare Behandlung der Daten

2.1 Beschreibung und Darstellung univariater Datensätze

Wie wir an den Beispielen in Kapitel 1 gesehen haben, werden im Rahmen der multivariaten Analyse an jedem von n Objekten p Merkmale erhoben. Die Werte dieser Merkmale werden in der *Datenmatrix* $\mathbf{X}$ zusammengefasst, wobei alle Werte numerisch kodiert werden:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}.$$

Diese Datenmatrix besteht aus n Zeilen und p Spalten. Dabei ist x_{ij} der Wert des j-ten Merkmals beim i-ten Objekt. In der i-ten Zeile der Datenmatrix $\mathbf{X}$ stehen also die Werte der p Merkmale beim i-ten Objekt. In der j-ten Spalte der Datenmatrix $\mathbf{X}$ stehen die Werte des j-ten Merkmals bei allen Objekten. Oft werden die Werte der einzelnen Merkmale beim i-ten Objekt benötigt. Man fasst diese in einem Vektor $\mathbf{x}_i$ zusammen:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}. \tag{2.1}$$

Example 13. Im Beispiel 2 auf Seite 3 wurden 5 Merkmale bei 20 Studenten erhoben. Also ist $n = 20$ und $p = 5$. Wir müssen die Merkmale `Geschlecht`, `MatheLK` und `Abitur88` kodieren. Beim Merkmal `Geschlecht` weisen wir der Ausprägung `w` die 1 und der Ausprägung `m` die 0 zu. Bei den beiden anderen Merkmalen ordnen wir der Ausprägung `j` eine 1 und der Ausprägung `n` eine 0 zu. Die Datenmatrix sieht also folgendermaßen aus:

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 3 & 0 & 8 \\ 0 & 0 & 4 & 0 & 7 \\ 0 & 0 & 4 & 0 & 4 \\ 0 & 0 & 4 & 0 & 2 \\ 0 & 0 & 3 & 0 & 7 \\ 1 & 0 & 3 & 0 & 6 \\ 1 & 0 & 4 & 1 & 3 \\ 1 & 0 & 3 & 1 & 7 \\ 1 & 0 & 4 & 1 & 14 \\ 0 & 1 & 3 & 0 & 19 \\ 0 & 1 & 3 & 0 & 15 \\ 0 & 1 & 2 & 0 & 17 \\ 0 & 1 & 3 & 0 & 10 \\ 1 & 1 & 3 & 0 & 22 \\ 1 & 1 & 2 & 0 & 23 \\ 1 & 1 & 2 & 0 & 15 \\ 0 & 1 & 1 & 1 & 21 \\ 1 & 1 & 2 & 1 & 10 \\ 1 & 1 & 2 & 1 & 12 \\ 1 & 1 & 4 & 1 & 17 \end{pmatrix}. \tag{2.2}$$

□

Vor einer multivariaten Analyse wird man sich die Eigenschaften der Verteilungen der einzelnen Merkmale ansehen. Aus diesem Grunde beschäftigen wir uns zunächst mit der *univariaten Analyse*. Wir betrachten also die Werte in einer Spalte der Datenmatrix $\mathbf{X}$. Bei der Beschreibung und Darstellung der Merkmale werden wir in Abhängigkeit vom Merkmal unterschiedlich vorgehen. Man unterscheidet *qualitative* und *quantitative* Merkmale. Bei qualitativen Merkmalen sind die einzelnen Merkmalsausprägungen *Kategorien*, wobei jeder Merkmalsträger zu genau einer Kategorie gehört. Kann man die Ausprägungen eines qualitativen Merkmals nicht anordnen, so ist das Merkmal *nominalskaliert*. Kann man die Kategorien anordnen, so spricht man von einem *ordinalskalierten* Merkmal. Quantitative Merkmale zeichnen sich dadurch aus, dass die Merkmalsausprägungen Zahlen sind, mit denen man rechnen kann. hmcounterend. (fortgesetzt)

Example 13. Die Merkmale `Geschlecht`, `MatheLK`, `MatheNote` und `Abitur88` sind qualitative Merkmale, wobei die Merkmale `Geschlecht`, `MatheLK` und `Abitur88` nominalskaliert sind, während das Merkmal `MatheNote` ordinalskaliert ist. Das Merkmal `Punkte` ist quantitativ. □

2.1.1 Beschreibung und Darstellung qualitativer Merkmale

Wir gehen aus von den Ausprägungen $x_1, \ldots, x_n$ eines Merkmals bei n Objekten. Man spricht von der *Urliste*. Man nennt x_i auch die i-te Beobachtung. hmcounterend. (fortgesetzt)

Example 13. Wir wollen uns das Merkmal `MatheLK` näher ansehen. Hier sind die Werte der 20 Studenten:

```
0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1.
```

Konkret gilt $x_1 = 0$. □

Die Analyse eines qualitativen Merkmals mit den Merkmalsausprägungen $A_1, \ldots, A_k$ beginnt mit dem Zählen. Man bestimmt die *absolute Häufigkeit* n_i der i-ten Merkmalsausprägung A_i. hmcounterend. (fortgesetzt)

Example 13. Von den 20 Studenten haben 11 den Mathematik-Leistungskurs besucht, während 9 ihn nicht besucht haben. Die Merkmalsausprägung A_1 sei die 0 und die Merkmalsausprägung A_2 die 1. Es gilt also $n_1 = 9$ und $n_2 = 11$. □

Ob eine absolute Häufigkeit groß oder klein ist, hängt von der Anzahl n der untersuchten Objekte ab. Wir beziehen die absolute Häufigkeit n_i auf n und erhalten die *relative Häufigkeit* h_i mit

$$h_i = \frac{n_i}{n}.$$

hmcounterend. (fortgesetzt)

Example 13. Es gilt $h_1 = 0.45$ und $h_2 = 0.55$. □

Absolute und relative Häufigkeiten stellt man in einer *Häufigkeitstabelle* zusammen. Tabelle 2.1 zeigt den allgemeinen Aufbau einer Häufigkeitstabelle.

Table 2.1. Allgemeiner Aufbau der Häufigkeitstabelle eines qualitativen Merkmals

Merkmals-ausprägung	absolute Häufigkeit	relative Häufigkeit
A_1	n_1	h_1
$\vdots$	$\vdots$	$\vdots$
A_k	n_k	h_k

hmcounterend. (fortgesetzt)

Example 13. Für die 20 Studenten erhalten wir in Tabelle 2.2 die Häufigkeitstabelle. □

Table 2.2. Häufigkeitstabelle des Merkmals MatheLK

Merkmals-ausprägung	absolute Häufigkeit	relative Häufigkeit
0	9	0.45
1	11	0.55

Die Informationen in einer Häufigkeitstabelle werden in einem *Stabdiagramm* graphisch dargestellt. Hierbei stehen in einem kartesischen Koordinatensystem auf der Abszisse die Merkmalsausprägungen und auf der Ordinate die relativen Häufigkeiten. Über jeder Merkmalsausprägung wird eine senkrechte Linie abgetragen, deren Länge der relativen Häufigkeit der Merkmalsausprägung entspricht. hmcounterend. (fortgesetzt)

Example 13. In Abbildung 2.1 ist das Stabdiagramm des Merkmals `MatheLK` zu finden. Um es leichter interpretieren zu können, haben wir bei der Achsenbeschriftung die Merkmalsausprägungen `n` und `j` gewählt. Wir erkennen an der Graphik auf einen Blick, dass die relativen Häufigkeiten der beiden Merkmalsausprägungen sich kaum unterscheiden. □

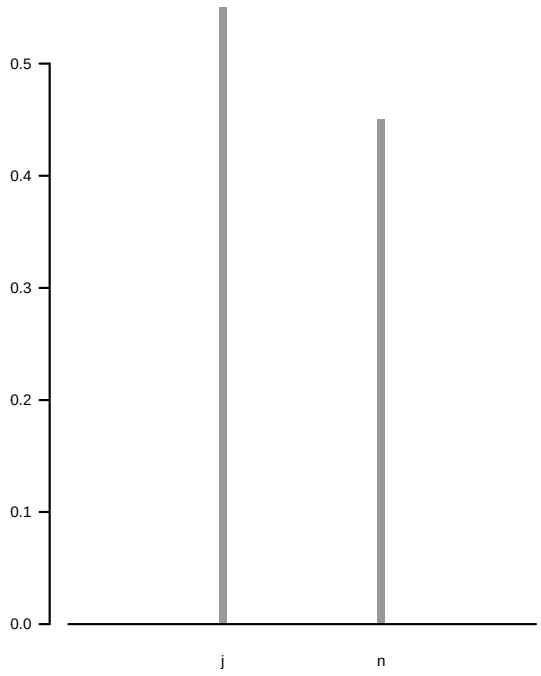


Fig. 2.1. Stabdiagramm des Merkmals MatheLK

2.1.2 Beschreibung und Darstellung quantitativer Merkmale

Example 14. Im Beispiel 1 auf Seite 3 sind alle Merkmale quantitativ. Sehen wir uns das Merkmal `Mathematische Grundbildung` an. Die Urliste sieht folgendermaßen aus:

```
533 520 334 514 490 536 517 447 529 503 514 457 557
533 547 463 514 446 387 537 499 515 470 454 478 510
529 476 498 488 493 .
```

□

Die Urliste ist sehr unübersichtlich. Ordnen wir die Werte der Größe nach, so können wir bereits Struktur erkennen. Man bezeichnet die i-t kleinste Beobachtung mit $x_{(i)}$. Der *geordnete Datensatz* ist somit $x_{(1)}, \ldots, x_{(n)}$. hmcounterend. (fortgesetzt)

Example 14. Der geordnete Datensatz ist

```
334 387 446 447 454 457 463 470
476 478 488 490 493 498 499
503
510 514 514 514 515 517 520
529 529 533 533 536 537 547 557 .
```

□

Bei so vielen unterschiedlichen Werten ist es nicht sinnvoll, ein Stabdiagramm zu erstellen. Es werden Klassen gebildet. Eine Beobachtung gehört zu einer Klasse, wenn sie größer als die Untergrenze, aber kleiner oder gleich der Obergrenze dieser Klasse ist. hmcounterend. (fortgesetzt)

Example 14. Wir wählen 6 äquidistante Klassen so, dass die Untergrenze der ersten Klasse gleich 300 und die Obergrenze der letzten Klasse gleich 600 ist. Die Untergrenze der 4-ten Klasse ist 450 und die Obergrenze 500. Zur 4-ten Klasse gehören die Beobachtungen 454, 457, 463, 470, 476, 478, 488, 490, 493, 498, und 499. □

Die Häufigkeitsverteilung der Klassen wird in einem *Histogramm* dargestellt. Dabei trägt man die Klassen auf der Abszisse ab und zeichnet über jeder Klasse ein Rechteck, dessen Höhe gleich der relativen Häufigkeit der Klasse dividiert durch die Klassenbreite ist. Hierdurch ist die Fläche unter dem Histogramm gleich 1. hmcounterend. (fortgesetzt)

Example 14. Abbildung 2.2 zeigt das Histogramm des Merkmals `Mathematische Grundbildung`. Das Histogramm deutet auf eine *rechtssteile* Verteilung hin. Man bezeichnet diese auch als *linksschief*. □

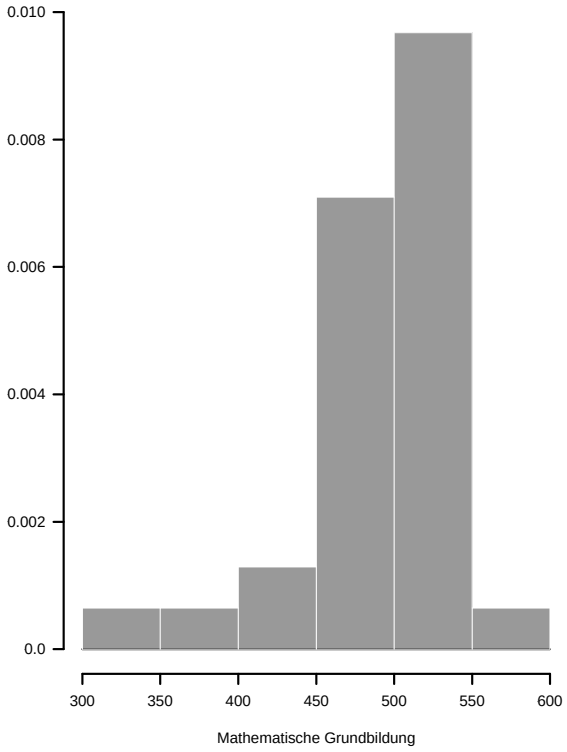


Fig. 2.2. Histogramm des Merkmals Mathematische Grundbildung

Wir wollen noch eine andere Art der Darstellung eines quantitativen univariaten Merkmals betrachten. Tukey (1977) hat vorgeschlagen, einen Datensatz durch folgende 5 Zahlen zusammenzufassen:

das *Minimum*	$x_{(1)}$,
das *untere Quartil*	$x_{0.25}$,
der *Median*	$x_{0.5}$,
das *obere Quartil*	$x_{0.75}$,
das *Maximum*	$x_{(n)}$.

Zunächst bestimmt man das Minimum $x_{(1)}$ und das Maximum $x_{(n)}$. hmcounterend. (fortgesetzt)

Example 14. Es gilt $x_{(1)} = 334$ und $x_{(n)} = 557$. □

Durch Minimum und Maximum kennen wir den Bereich, in dem die Werte liegen. Außerdem können wir mit Hilfe dieser beiden Zahlen eine einfache

Maßzahl für die *Streuung* bestimmen. Die Differenz aus Maximum und Minimum nennt man die *Spannweite* R. Es gilt also

$$R = x_{(n)} - x_{(1)}. \tag{2.3}$$

hmcounterend. (fortgesetzt)

Example 14. Die Spannweite beträgt 223. □

Eine Maßzahl für die *Lage* des Datensatzes ist der Median $x_{(0.5)}$. Dieser ist die Zahl, die den geordneten Datensatz in zwei gleiche Teile teilt. Ist der Stichprobenumfang ungerade, dann ist der Median die Beobachtung in der Mitte des geordneten Datensatzes. Ist der Stichprobenumfang gerade, so ist der Median der Mittelwert der beiden mittleren Beobachtungen im geordneten Datensatz. Formal kann man den Median folgendermaßen definieren:

$$x_{0.5} = \begin{cases} x_{(0.5(n+1))} & \text{falls } n \text{ ungerade ist} \\ 0.5(x_{(0.5n)} + x_{(1+0.5n)}) & \text{falls } n \text{ gerade ist.} \end{cases} \tag{2.4}$$

hmcounterend. (fortgesetzt)

Example 14. Der Stichprobenumfang ist gleich 31. Der Median ist somit die Beobachtung an der 16-ten Stelle des geordneten Datensatzes. Der Wert des Medians beträgt somit 503. □

Neben dem Minimum, Maximum und Median betrachtet Tukey (1977) noch das untere Quartil $x_{0.25}$ und das obere Quartil $x_{0.75}$. 25 Prozent der Beobachtungen sind kleiner oder gleich dem unteren Quartil $x_{0.25}$ und 75 Prozent der Beobachtungen sind kleiner oder gleich dem oberen Quartil $x_{0.75}$. Das untere Quartil teilt die untere Hälfte des geordneten Datensatzes in zwei gleich große Hälften, während das obere Quartil die obere Hälfte des geordneten Datensatzes in zwei gleich große Hälften teilt. Somit ist das untere Quartil der Median der unteren Hälfte des geordneten Datensatzes, während das obere Quartil der Median der oberen Hälfte des geordneten Datensatzes ist. Ist der Stichprobenumfang gerade, so ist die untere und obere Hälfte des geordneten Datensatzes eindeutig definiert. Bei einem ungeraden Stichprobenumfang gehört der Median sowohl zur oberen als auch zur unteren Hälfte des geordneten Datensatzes. hmcounterend. (fortgesetzt)

Example 14. Das untere Quartil ist der Mittelwert aus $x_{(8)} = 470$ und $x_{(9)} = 476$ und beträgt somit 473, während das obere Quartil der Mittelwert aus $x_{(23)} = 520$ und $x_{(24)} = 529$ ist. Es beträgt 524.5. Die 5 Zahlen sind somit

$$\begin{aligned} x_{(1)} &= 334, \\ x_{0.25} &= 473, \\ x_{0.5} &= 503, \\ x_{0.75} &= 524.5, \\ x_{(n)} &= 557. \end{aligned}$$

□

Tukey (1977) hat vorgeschlagen, die 5 Zahlen in einem sogenannten *Boxplot* graphisch darzustellen. Beim Boxplot wird ein Kasten vom unteren Quartil bis zum oberen Quartil gezeichnet. Außerdem wird der Median als Linie in den Kasten eingezeichnet. Von den Rändern des Kastens bis zu den Extremen werden Linien gezeichnet, die an sogenannten *Zäunen* enden. Um Ausreißer zu markieren, wird der letzte Schritt modifiziert: Sind Punkte mehr als das 1.5-fache der Kastenbreite von den Quartilen entfernt, so wird die Linie nur bis zum 1.5-fachen der Kastenbreite gezeichnet. Alle Punkte, die außerhalb liegen, werden markiert. hmcounterend. (fortgesetzt)

Example 14. Abbildung 2.3 zeigt den Boxplot des Merkmals `Mathematische Grundbildung`. Der Boxplot deutet auch auf eine linksschiefe Verteilung hin. Außerdem ist ein Ausreißer gut zu erkennen.

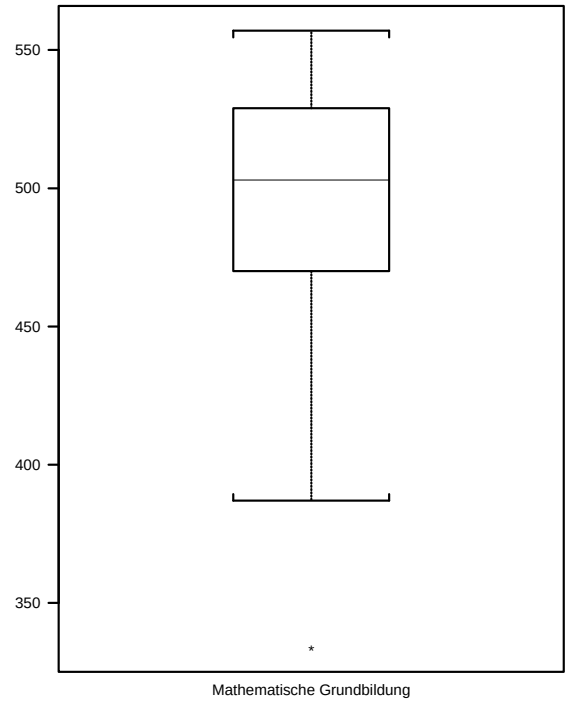


Fig. 2.3. Boxplot des Merkmals Mathematische Grundbildung im Rahmen der PISA-Studie

□

Ein wichtiger Aspekt der Verteilung eines quantitativen Merkmals ist die Lage. Wir haben bisher den Median als eine Maßzahl zur Beschreibung der Lage kennengelernt. Neben dem Median ist der *Mittelwert* $\bar{x}$ die wichtigste Maßzahl zur Beschreibung der Lage. Dieser ist folgendermaßen definiert:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

hmcounterend. (fortgesetzt)

Example 14. Es gilt $\bar{x} = 493.16$. Der Mittelwert ist kleiner als der Median. Dies ist bei einer linksschiefen Verteilung der Fall. □

Transformieren wir alle Beobachtungen x_i linear zu $y_i = b + a\,x_i$, so gilt

$$\bar{y} = b + a\,\bar{x}. \tag{2.5}$$

Dies sieht man folgendermaßen:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(b + a\,x_i) = \frac{1}{n}\sum_{i=1}^{n} b + \frac{1}{n}\sum_{i=1}^{n} a\,x_i$$

$$= \frac{1}{n}\,n\,b + a\,\frac{1}{n}\sum_{i=1}^{n} x_i = b + a\,\bar{x}.$$

Bei einer symmetrischen Verteilung sind Mittelwert und Median identisch. Der Mittelwert ist ausreißerempfindlich. Eine Beobachtung, die stark von den anderen Beobachtungen abweicht, hat einen großen Einfluss auf den Mittelwert. Man sagt auch, dass der Mittelwert nicht *robust* ist. Da Ausreißer einen starken Einfluss auf den Mittelwert haben, liegt es nahe, einen Anteil α auf beiden Seiten der geordneten Stichprobe zu entfernen und den Mittelwert der restlichen Beobachtungen zu bestimmen. Man spricht in diesem Fall von einem *getrimmten Mittelwert* $\bar{x}_\alpha$. Formal kann man diesen so beschreiben:

$$\bar{x}_\alpha = \frac{1}{n - 2\,\lfloor n\,\alpha \rfloor}\sum_{i=1+\lfloor n\alpha \rfloor}^{n-\lfloor n\alpha \rfloor} x_{(i)}. \tag{2.6}$$

Dabei ist $\lfloor c \rfloor$ der ganzzahlige Teil der positiven reellen Zahl c. Typische Werte für α sind 0.05 und 0.10. hmcounterend. (fortgesetzt)

Example 14. Für $n = 31$ gilt $\lfloor n\,0.05 \rfloor = 1$ und $\lfloor n\,0.1 \rfloor = 3$. Für das Merkmal `Mathematische Grundbildung` erhalten wir $\bar{x}_{0.05} = 496.45$ und $\bar{x}_{0.10} = 499.2$. □

Den Median kann man als getrimmten Mittelwert mit $\alpha = 0.5$ auffassen. Je höher der Wert von α ist, umso mehr Beobachtungen können vom Rest der Beobachtungen abweichen, ohne dass dies den getrimmten Mittelwert beeinflusst.

Neben der Lage ist die Streuung von größtem Interesse. Wir haben bereits die Spannweite R als Maß für die Streuung kennengelernt. Ein anderes Maß für die Streuung ist die *Stichprobenvarianz*. Diese ist definiert durch

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.7}$$

hmcounterend. (fortgesetzt)

Example 14. Es gilt $s_x^2 = 2192.873$. □

Die Stichprobenvarianz besitzt nicht die gleiche Maßeinheit wie die Beobachtungen. Zieht man aus der Stichprobenvarianz die Quadratwurzel, so erhält man eine Maßzahl, die die gleiche Dimension wie die Beobachtungen besitzt. Diese heißt *Standardabweichung* s_x. hmcounterend. (fortgesetzt)

Example 14. Es gilt $s_x = 46.83$. □

Transformieren wir alle Beobachtungen x_i linear zu $y_i = b + a\,x_i$, so gilt

$$s_y^2 = a^2 s_x^2. \tag{2.8}$$

Dies sieht man mit Gleichung (2.5) folgendermaßen:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(b + a\,x_i - b - a\,\bar{x})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(a\,(x_i - \bar{x}))^2 = a^2\,\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = a^2 s_x^2.$$

2.2 Beschreibung und Darstellung multivariater Datensätze

Bisher haben wir nur ein einzelnes Merkmal analysiert. Nun wollen wir mehrere Merkmale gemeinsam betrachten, um zum Beispiel Abhängigkeitsstrukturen zwischen den Merkmalen aufzudecken. Wir gehen davon aus, dass an jedem von n Objekten p Merkmale erhoben wurden. Wir wollen zeigen, wie man Informationen in Datenmatrizen einfach darstellen kann. Dabei wollen wir wieder zwischen qualitativen und quantitativen Merkmalen unterscheiden.

2.2.1 Beschreibung und Darstellung von Datenmatrizen quantitativer Merkmale

Example 15. Wir betrachten den Datensatz im Beispiel 1 auf Seite 3 und stellen die Daten in einer Datenmatrix zusammen. In der ersten Spalte stehen die Werte des Merkmals `Lesekompetenz`, in der zweiten Spalte die Werte des Merkmals `Mathematische Grundbildung` und in der letzten Spalte die Werte des Merkmals `Naturwissenschaftliche Grundbildung`:

$$\mathbf{X} = \begin{pmatrix} 528 & 533 & 528 \\ 507 & 520 & 496 \\ 396 & 334 & 375 \\ 497 & 514 & 481 \\ 484 & 490 & 487 \\ \vdots & \vdots & \vdots \\ 492 & 498 & 511 \\ 480 & 488 & 496 \\ 504 & 493 & 499 \end{pmatrix}.$$

In der fünften Zeile der Matrix $\mathbf{X}$ stehen die Merkmalsausprägungen von Deutschland:

$$\mathbf{x}_5 = \begin{pmatrix} 484 \\ 490 \\ 487 \end{pmatrix}.$$

□

Wir wollen nun das Konzept des Mittelwerts auf mehrere Merkmale übertragen. Dies ist ganz einfach. Wir bestimmen den Mittelwert jedes Merkmals und fassen diese Mittelwerte zum Vektor der Mittelwerte zusammen. Wir bezeichnen den Mittelwert des j-ten Merkmals mit $\bar{x}_j$. Es gilt also

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}\,.$$

Für den Vektor $\bar{\mathbf{x}}$ der Mittelwerte gilt also

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}.$$

hmcounterend. (fortgesetzt)

Example 15. Es gilt

$$\bar{\mathbf{x}} = \begin{pmatrix} 493.45 \\ 493.16 \\ 492.61 \end{pmatrix}. \tag{2.9}$$

Wir sehen, dass im Bereich `Lesekompetenz` im Durchschnitt am meisten Punkte erreicht wurden, während die Leistungen im Bereich `Naturwissenschaftliche Grundbildung` im Mittel am schlechtesten waren. □

Mit den Beobachtungsvektoren $\mathbf{x}_i$, $i = 1, \ldots, n$, aus Gleichung (2.1) können wir den Vektor $\bar{\mathbf{x}}$ der Mittelwerte auch bestimmen durch

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i\,.$$

Dies sieht man folgendermaßen:

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} x_{ip} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} x_{i1} \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i\,.$$

Manche der multivariaten Verfahren, die wir betrachten werden, gehen davon aus, dass die Merkmale *zentriert* sind. Wir zentrieren die Werte des i-ten Merkmals, indem wir von jedem Wert x_{ij} den Mittelwert $\bar{x}_j$ subtrahieren:

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_j \,.$$

Der Mittelwert eines zentrierten Merkmals ist gleich 0. Dies sieht man folgendermaßen:

$$\bar{\tilde{x}}_j = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) = \frac{1}{n} \sum_{i=1}^{n} x_{ij} - \frac{1}{n} \sum_{i=1}^{n} \bar{x}_j = \bar{x}_j - \frac{1}{n}\, n\, \bar{x}_j = 0 \,.$$

Die *zentrierte Datenmatrix* ist

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \dots & x_{np} - \bar{x}_p \end{pmatrix} . \tag{2.10}$$

hmcounterend. (fortgesetzt)

Example 15. Es gilt

$$\tilde{\mathbf{X}} = \begin{pmatrix} 34.55 & 39.84 & 35.39 \\ 13.55 & 26.84 & 3.39 \\ -97.45 & -159.16 & -117.61 \\ 3.55 & 20.84 & -11.61 \\ -9.45 & -3.16 & -5.61 \\ \vdots & \vdots & \vdots \\ -1.45 & 4.84 & 18.39 \\ -13.45 & -5.16 & 3.39 \\ 10.55 & -0.16 & 6.39 \end{pmatrix} .$$

An der zentrierten Datenmatrix kann man sofort erkennen, wie sich jedes Land vom Mittelwert unterscheidet. Wir sehen, dass Deutschland als fünftes Land in der Matrix in allen Bereichen unter dem Durchschnitt liegt, während Australien als erstes Land in der Matrix in allen Bereichen über dem Durchschnitt liegt. □

Wir wollen uns nun noch anschauen, wie man die zentrierte Datenmatrix durch eine einfache Multiplikation mit einer anderen Matrix gewinnen kann. Diese Matrix werden wir im Folgenden öfter verwenden. Sei

$$\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\, \mathbf{1}\mathbf{1}' \,. \tag{2.11}$$

Dabei ist $\mathbf{I}_n$ die Einheitsmatrix und $\mathbf{1}$ der Einservektor. Es gilt

$$\tilde{\mathbf{X}} = \mathbf{M}\mathbf{X}\,. \tag{2.12}$$

Um Gleichung (2.12) zu zeigen, formen wir sie um:

$$\mathbf{M}\mathbf{X} = (\mathbf{I}_n - \frac{1}{n}\,\mathbf{1}\mathbf{1}')\mathbf{X} = \mathbf{X} - \frac{1}{n}\,\mathbf{1}\mathbf{1}'\mathbf{X}\,.$$

Wir betrachten zunächst $\frac{1}{n}\,\mathbf{1}\mathbf{1}'\mathbf{X}$. Da $\mathbf{1}$ der summierende Vektor ist, gilt

$$\mathbf{1}'\mathbf{X} = (\sum_{i=1}^{n} x_{i1}, \ldots, \sum_{i=1}^{n} x_{ip})\,.$$

Da $\frac{1}{n}$ ein Skalar ist, gilt

$$\frac{1}{n}\,\mathbf{1}\mathbf{1}'\mathbf{X} = \mathbf{1}\,\frac{1}{n}\,\mathbf{1}'\mathbf{X}\,.$$

Es gilt

$$\frac{1}{n}\,\mathbf{1}'\mathbf{X} = (\bar{x}_1, \ldots, \bar{x}_p)\,.$$

Somit folgt

$$\frac{1}{n}\,\mathbf{1}\mathbf{1}'\mathbf{X} = \mathbf{1}\,\frac{1}{n}\,\mathbf{1}'\mathbf{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} (\bar{x}_1, \ldots, \bar{x}_p) = \begin{pmatrix} \bar{x}_1 & \ldots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \ldots & \bar{x}_p \end{pmatrix}.$$

Also gilt

$$\mathbf{M}\mathbf{X} = \mathbf{X} - \frac{1}{n}\,\mathbf{1}\mathbf{1}'\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} - \begin{pmatrix} \bar{x}_1 & \ldots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \ldots & \bar{x}_p \end{pmatrix}$$

$$= \begin{pmatrix} x_{11} - \bar{x}_1 & \ldots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \ldots & x_{np} - \bar{x}_p \end{pmatrix} = \tilde{\mathbf{X}}.$$

Man nennt $\mathbf{M}$ auch die *Zentrierungsmatrix*. Sie ist symmetrisch. Es gilt nämlich

$$\mathbf{M}' = (\mathbf{I}_n - \frac{1}{n}\,\mathbf{1}\mathbf{1}')' = \mathbf{I}_n' - (\frac{1}{n}\,\mathbf{1}\mathbf{1}')' = \mathbf{I}_n - \frac{1}{n}\,\mathbf{1}\mathbf{1}' = \mathbf{M}.$$

Multipliziert man die Datenmatrix also von rechts mit der Matrix

$$\mathbf{M} = \mathbf{I}_p - \frac{1}{p}\,\mathbf{1}\mathbf{1}',$$

so werden die Zeilen zentriert.

Bei der univariaten Datenanalyse haben wir robuste Schätzer wie den Median und den getrimmten Mittelwert betrachtet. Wir wollen nun aufzeigen, wie man diese Konzepte auf den multivariaten Fall übertragen kann. Hierbei werden wir uns aber auf den zweidimensionalen Fall beschränken, da nur in diesem Fall Funktionen in `S-PLUS` existieren. Beginnen wir mit dem Trimmen. Hierzu stellen wir die Werte der beiden Merkmale in einem *Streudiagramm* dar. Die beiden Merkmale bilden die Achsen in einem kartesischen Koordinatensystem. Die Werte jedes Objekts werden als Punkt in dieses Koordinatensystem eingetragen. hmcounterend. (fortgesetzt)

Example 15. Abbildung 2.4 zeigt das Streudiagramm der Merkmale `Lesekompetenz` und `Mathematische Grundbildung`. □

Bei nur einem Merkmal ist das Trimmen eindeutig. Man ordnet die Werte der Größe nach und entfernt jeweils einen Anteil α der extremen Werte auf beiden Seiten der geordneten Stichprobe. Bei zwei Merkmalen gibt es keine

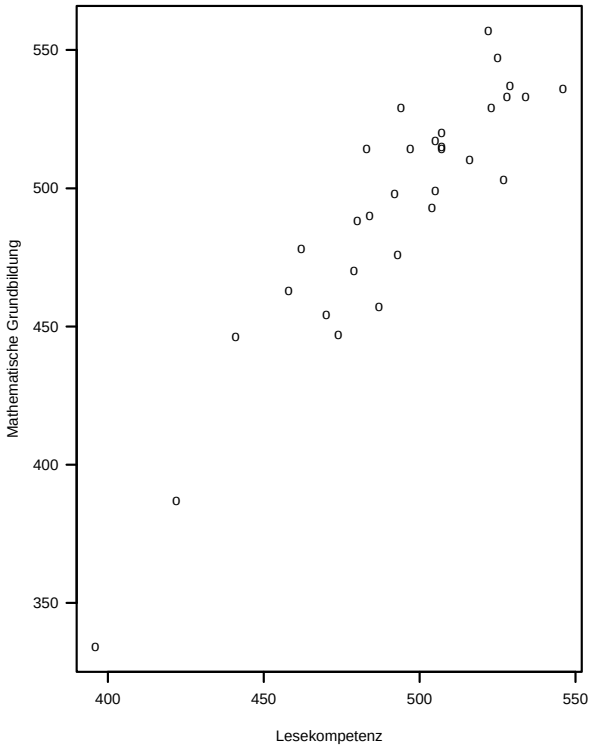


Fig. 2.4. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

natürliche Ordnung. Natürlich kann man jedes der beiden Merkmale getrennt trimmen. Hierbei berücksichtigt man aber nicht, dass beide Merkmale an demselben Objekt erhoben wurden. Es gibt nun eine Reihe von Vorschlägen, wie man im zweidimensionalen Raum trimmen kann. Wir wollen uns einen von diesen anschauen. Man bestimmt hierzu zunächst die *konvexe Hülle* der Menge der Beobachtungen. Die konvexe Hülle ist das kleinste Polygon, in dem entweder jede Beobachtung auf dem Rand oder innerhalb des Polygons liegt. Büning (1991), S.202 veranschaulicht die Konstruktion der konvexen Hülle folgendermaßen:

> Wir können uns die Punkte $\mathbf{x}_1, \ldots, \mathbf{x}_n$ als Nägel auf einem Brett vorstellen, um die ein (großes) elastisches Band gespannt und dann losgelassen wird; das Band kommt in Form eines Polygons zur Ruhe.

hmcounterend. (fortgesetzt)

Example 15. Abbildung 2.5 zeigt die konvexe Hülle der Beobachtungen der Merkmale `Lesekompetenz` und `Mathematische Grundbildung`. Auf der kon-

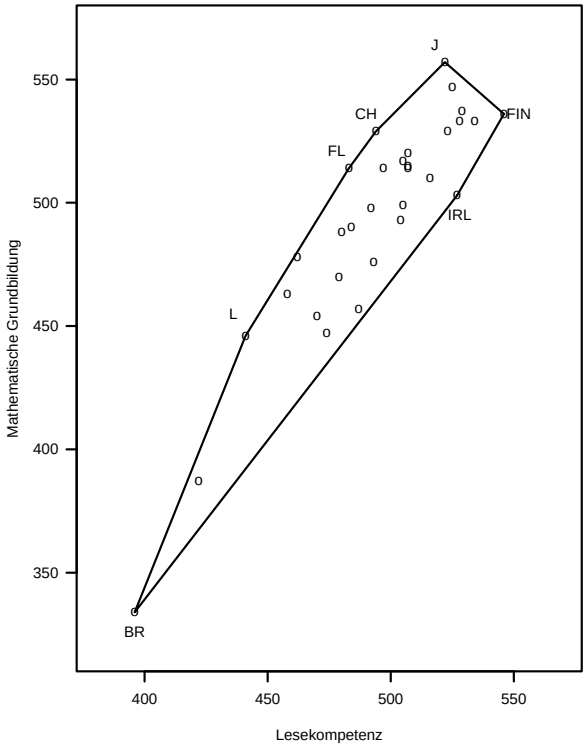


Fig. 2.5. Konvexe Hülle der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

vexen Hülle liegen die Länder `IRL`, `BR`, `L`, `FL`, `CH`, `J` und `FIN`. □

Einen auf der konvexen Hülle basierenden getrimmten Mittelwert erhält man dadurch, dass man alle Beobachtungen auf der konvexen Hülle aus dem Datensatz entfernt und den Mittelwert der restlichen Beobachtungen bestimmt. hmcounterend. (fortgesetzt)

Example 15. Es sind 7 Punkte auf der konvexen Hülle. Somit beträgt der Trimmanteil $7/31 = 0.23$. Der getrimmte Mittelwert beträgt

$$\bar{\mathbf{x}}_{0.23} = \begin{pmatrix} 495.33 \\ 494.54 \end{pmatrix}.$$

Ein Vergleich mit den ersten beiden Komponenten von $\bar{\mathbf{x}}$ in (2.9) zeigt, dass sich dieser nicht stark vom Mittelwert unterscheidet. □

Im Englischen nennt nennt diese Vorgehensweise *Peeling*. Heiler & Michels (1994) verwenden den Begriff *Schälen*. Man kann nun eine konvexe Hülle nach der anderen entfernen, bis nur noch eine übrig bleibt. Liegt innerhalb dieser Hülle noch ein Punkt, so ist dieser der *multivariate Median*. Liegt innerhalb dieser Hülle kein Punkt, so wählt man den Mittelwert der Beobachtungen auf der innersten Hülle als multivariaten Median. Heiler & Michels (1994), S.237 nennen ihn auch *Konvexe-Hüllen-Median*. hmcounterend. (fortgesetzt)

Example 15. Abbildung 2.6 zeigt die innerste Hülle des Datensatzes. Wir

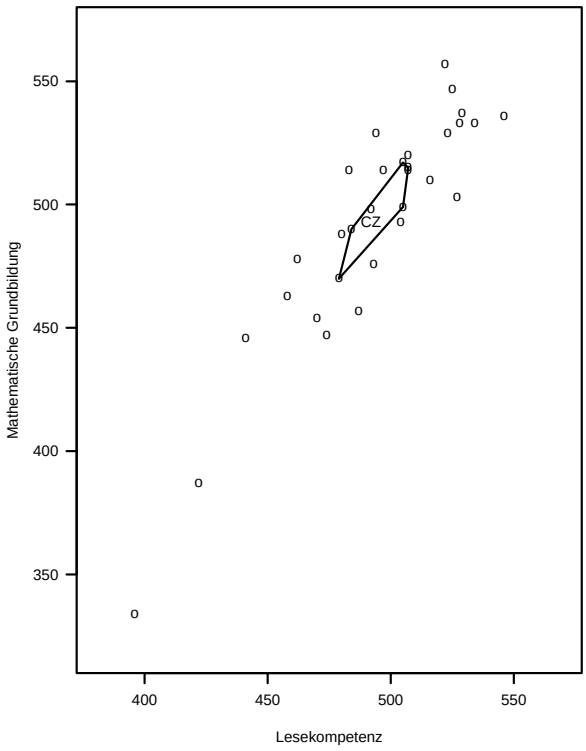


Fig. 2.6. Innerste konvexe Hülle der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

sehen, dass innerhalb dieser konvexen Hülle ein Punkt liegt. Es handelt sich um `CZ`. Also bilden die Werte von Tschechien den multivariaten Median.

Dieser ist gegeben durch

$$\begin{pmatrix} 492 \\ 498 \end{pmatrix}.$$

□

Weitere Ansätze zur Bestimmung eines multivariaten Medians sind bei Büning (1991), Heiler & Michels (1994) und Small (1990) zu finden.

Ein Maß für die Streuung eines univariaten Merkmals ist die Stichprobenvarianz. In Analogie zu (2.7) ist die Stichprobenvarianz des j-ten Merkmals definiert durch

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 . \tag{2.13}$$

Die Standardabweichung des j-ten Merkmals ist $s_j = \sqrt{s_j^2}$. hmcounterend. (fortgesetzt)

Example 15. Die Stichprobenvarianzen der einzelnen Merkmale sind $s_1^2 = 1109.4$, $s_2^2 = 2192.9$ und $s_3^2 = 1419.0$. Wir sehen, dass die Punkte am stärksten im Bereich `Mathematische Grundbildung` und am wenigsten im Bereich `Lesekompetenz` streuen. Die Standardabweichungen der Merkmale sind $s_1 = 33.3$, $s_2 = 46.8$ und $s_3 = 37.7$. □

Wir haben in Gleichung (2.10) die Merkmale zentriert. Dividiert man die Werte eines zentrierten Merkmals noch durch die Standardabweichung dieses Merkmals, so erhält man *standardisierte* Merkmale

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} .$$

Der Mittelwert eines standardisierten Merkmals ist gleich 0. Dies sieht man folgendermaßen:

$$\overline{x_j^*} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^* = \frac{1}{n} \sum_{i=1}^{n} \frac{x_{ij} - \bar{x}_j}{s_j} = \frac{1}{n\, s_j} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) = 0 .$$

Die Stichprobenvarianz der standardisierten Merkmale ist gleich 1. Dies sieht man folgendermaßen:

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_{ij} - \bar{x}_j}{s_j} \right)^2 &= \frac{1}{s_j^2} \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \\ &= \frac{1}{s_j^2}\, s_j^2 = 1. \end{aligned}$$

Die *Matrix der standardisierten Merkmale* ist:

$$\mathbf{X}^* = \begin{pmatrix} \frac{x_{11} - \bar{x}_1}{s_1} & \cdots & \frac{x_{1p} - \bar{x}_p}{s_p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_1}{s_1} & \cdots & \frac{x_{np} - \bar{x}_p}{s_p} \end{pmatrix}. \tag{2.14}$$

hmcounterend. (fortgesetzt)

Example 15. Es gilt

$$\mathbf{X}^* = \begin{pmatrix} 1.037 & 0.851 & 0.939 \\ 0.407 & 0.573 & 0.090 \\ -2.926 & -3.399 & -3.122 \\ 0.107 & 0.445 & -0.308 \\ -0.284 & -0.068 & -0.149 \\ \vdots & \vdots & \vdots \\ -0.044 & 0.103 & 0.488 \\ -0.404 & -0.110 & 0.090 \\ 0.317 & -0.003 & 0.170 \end{pmatrix}.$$

□

Es ist nicht üblich, in Analogie zum Vektor der Mittelwerte einen Vektor der Stichprobenvarianzen zu bilden. Die Stichprobenvarianzen sind Bestandteil der *empirischen Varianz-Kovarianz-Matrix*. Um diese zu erhalten, benötigen wir die *empirische Kovarianz*, die wir nun herleiten wollen. Bisher haben wir uns die Charakteristika jedes einzelnen Merkmals angeschaut. In der multivariaten Analyse sind aber Zusammenhänge zwischen Merkmalen von Interesse. hmcounterend. (fortgesetzt)

Example 15. Schauen wir uns unter diesem Aspekt noch einmal das Streudiagramm der Merkmale `Lesekompetenz` und `Mathematische Grundbildung` in Abbildung 2.4 auf Seite 29 an. Wir sehen, dass Länder, die eine hohe Punktezahl im Bereich `Lesekompetenz` aufweisen, auch im Bereich `Mathematische Grundbildung` eine hohe Punktezahl erreichen. Länder mit einer niedrigen Punktezahl im Bereich `Lesekompetenz` weisen in der Regel auch einen niedrigen Wert im Bereich `Mathematische Grundbildung` auf. Ist ein Land also über dem Durchschnitt in einem Bereich, so ist es in der Regel auch über dem Durchschnitt im anderen Bereich. Dies wird auch am Streudiagramm deutlich, wenn wir die Mittelwerte der beiden Merkmale in diesem berücksichtigen. Hierzu zeichnen wir eine Gerade parallel zur Ordinate in Höhe des Mittelwerts der Punktezahl im Bereich `Lesekompetenz` und eine Gerade parallel zur Abszisse in Höhe des Mittelwerts der Punktezahl im Bereich `Mathematische Grundbildung`. Abbildung 2.7 veranschaulicht dies. Hierdurch erhalten wir 4 Quadranten, die in der Graphik durchnummeriert sind. Im ersten Quadranten

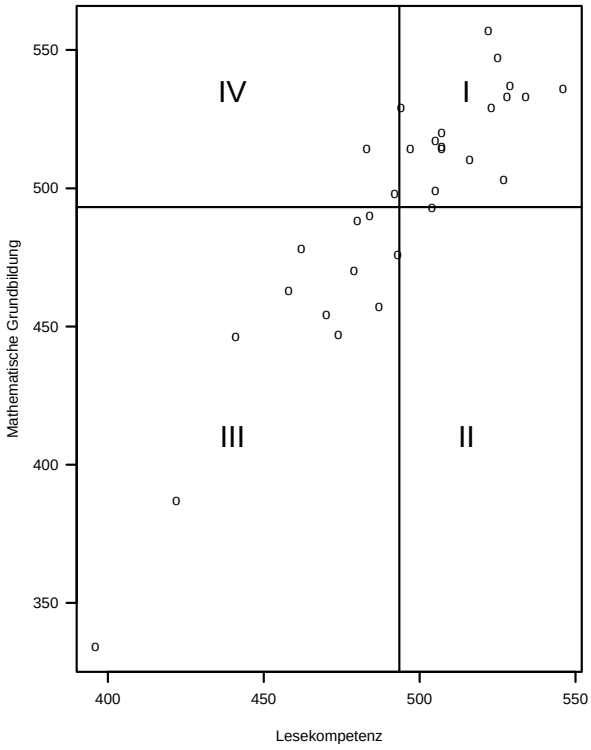


Fig. 2.7. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie, aufgeteilt in 4 Quadranten

sind die Länder, deren Punktezahl in den Bereichen `Lesekompetenz` und `Mathematische Grundbildung` über dem Durchschnitt liegen, während sich im dritten Quadranten die Länder befinden, deren Punktezahl in den Bereichen `Lesekompetenz` und `Mathematische Grundbildung` unter dem Durchschnitt liegen. Im zweiten Quadranten sind die Länder, deren Punktezahl im Bereich `Lesekompetenz` über dem Durchschnitt, im Bereich `Mathematische Grundbildung` hingegen unter dem Durchschnitt liegen, während im vierten Quadranten die Länder liegen, deren Punktezahl im Bereich `Lesekompetenz` unter dem Durchschnitt, im Bereich `Mathematische Grundbildung` hingegen über dem Durchschnitt liegen. Besteht ein positiver Zusammenhang zwischen den beiden Merkmalen, so werden wir die meisten Beobachtungen in den Quadranten I und III erwarten, während wir bei einem negativen Zusammenhang die meisten in den Quadranten II und IV erwarten. Verteilen sich die Punkte

gleichmäßig über die Quadranten, so liegt kein Zusammenhang zwischen den Merkmalen vor. □

Um den im Beispiel veranschaulichten Sachverhalt in eine geeignete Maßzahl für den Zusammenhang zwischen den beiden Merkmalen umzusetzen, gehen wir davon aus, dass das i-te Merkmal auf der Abszisse und das j-te Merkmal auf der Ordinate stehe. Sei x_{ki} die Ausprägung des i-ten Merkmals beim k-ten Objekt und x_{kj} die Ausprägung des j-ten Merkmals beim k-ten Objekt. Dann gilt in den einzelnen Quadranten:

Quadrant I: $x_{ki} > \bar{x}_i \ , \ x_{kj} > \bar{x}_j,$

Quadrant II: $x_{ki} > \bar{x}_i \ , \ x_{kj} < \bar{x}_j,$

Quadrant III: $x_{ki} < \bar{x}_i \ , \ x_{kj} < \bar{x}_j,$

Quadrant IV: $x_{ki} < \bar{x}_i \ , \ x_{kj} > \bar{x}_j.$

Also gilt

Quadrant I: $x_{ki} - \bar{x}_i > 0$, $x_{kj} - \bar{x}_j > 0$,
Quadrant II: $x_{ki} - \bar{x}_i > 0$, $x_{kj} - \bar{x}_j < 0$,
Quadrant III: $x_{ki} - \bar{x}_i < 0$, $x_{kj} - \bar{x}_j < 0$,
Quadrant IV: $x_{ki} - \bar{x}_i < 0$, $x_{kj} - \bar{x}_j > 0$.

Also ist das Produkt $(x_{ki} - \bar{x}_i) \cdot (x_{kj} - \bar{x}_j)$ im ersten und dritten Quadranten positiv, während es im zweiten und vierten Quadranten negativ ist. Dies legt nahe, folgende Maßzahl zu betrachten:

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \,. \tag{2.15}$$

s_{ij} heißt empirische Kovarianz zwischen dem i-ten und j-ten Merkmal. Es gilt

$$s_{jj} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)(x_{kj} - \bar{x}_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)^2 = s_j^2.$$

Bei p Merkmalen $x_1, \ldots, x_p$ bestimmt man zwischen allen Paaren von Merkmalen die Kovarianz und stellt diese Kovarianzen in der empirischen Varianz-Kovarianz-Matrix zusammen:

$$\mathbf{S} = \begin{pmatrix} s_1^2 & \ldots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \ldots & s_p^2 \end{pmatrix}.$$

Wegen $s_{ij} = s_{ji}$ ist die empirische Varianz-Kovarianz-Matrix symmetrisch. hmcounterend. (fortgesetzt)

Example 15. Es gilt

$$\mathbf{S} = \begin{pmatrix} 1109.4 & 1428.3 & 1195.6 \\ 1428.3 & 2192.9 & 1644.0 \\ 1195.6 & 1644.0 & 1419.0 \end{pmatrix}.$$

Wir sehen, dass alle empirischen Kovarianzen positiv sind. Die empirische Kovarianz zwischen den Merkmalen `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` ist am größten. □

Man kann die empirische Varianz-Kovarianz-Matrix auch folgendermaßen bestimmen:

$$\mathbf{S} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}_k - \bar{\mathbf{x}}) \, (\mathbf{x}_k - \bar{\mathbf{x}})'. \tag{2.16}$$

Mit

$$\mathbf{x}_k = \begin{pmatrix} x_{k1} \\ \vdots \\ x_{kp} \end{pmatrix}$$

und

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

gilt

$$(\mathbf{x}_k - \bar{\mathbf{x}})\,(\mathbf{x}_k - \bar{\mathbf{x}})' = \begin{pmatrix} x_{k1} - \bar{x}_1 \\ \vdots \\ x_{kp} - \bar{x}_p \end{pmatrix} \begin{pmatrix} (x_{k1} - \bar{x}_1 & \dots & x_{kp} - \bar{x}_p \end{pmatrix}$$

$$= \begin{pmatrix} (x_{k1} - \bar{x}_1)\,(x_{k1} - \bar{x}_1) & \dots & (x_{k1} - \bar{x}_1)\,(x_{kp} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ (x_{kp} - \bar{x}_p)\,(x_{k1} - \bar{x}_1) & \dots & (x_{kp} - \bar{x}_p)\,(x_{kp} - \bar{x}_p) \end{pmatrix}.$$

Summieren wir diese Matrizen von $k = 1$ bis n und dividieren die Summe durch $n - 1$, so erhalten wir die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$. Es gibt noch eine weitere Darstellung der empirischen Varianz-Kovarianz-Matrix, auf die wir noch häufiger zurückkommen werden. Sei $\tilde{\mathbf{X}}$ die zentrierte Datenmatrix aus Gleichung (2.10) auf Seite 26. Dann gilt

$$\mathbf{S} = \frac{1}{n-1}\,\tilde{\mathbf{X}}'\tilde{\mathbf{X}}. \tag{2.17}$$

Das Element in der i-ten Zeile und j-ten Spalte von $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ erhält man dadurch, dass man das innere Produkt aus den Vektoren bildet, die in der i-ten und der j-ten Spalte von $\tilde{\mathbf{X}}$ stehen. Dieses ist

$$\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)\,(x_{kj} - \bar{x}_j).$$

Dividiert man diesen Ausdruck durch $n - 1$, so erhält man die empirische Kovarianz zwischen dem i-ten und j-ten Merkmal, wie man durch einen Vergleich mit Gleichung (2.15) erkennt.

Die empirische Kovarianz ist nicht skaleninvariant. Multipliziert man alle Werte des einen Merkmals mit einer Konstanten b und die Werte des anderen Merkmals mit einer Konstanten c, so wird die empirische Kovarianz bc-mal so groß.

Mit (2.5) gilt nämlich

$$\begin{aligned}\frac{1}{n-1}\sum_{k=1}^{n}(b\,x_{ki}-\overline{b\,x_i})(c\,x_{kj}-\overline{c\,x_j}) &= \frac{1}{n-1}\sum_{k=1}^{n}(b\,x_{ki}-b\,\overline{x}_i)(c\,x_{kj}-c\,\overline{x}_j)\\ &= b\,c\,\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j)\\ &= b\,c\,s_{ij}.\end{aligned}$$

Man kann die empirische Kovarianz normieren, indem man sie durch das Produkt der Standardabweichungen der beiden Merkmale dividiert. Man erhält dann den *empirischen Korrelationskoeffizienten*

$$r_{ij}=\frac{s_{ij}}{s_i\,s_j}. \tag{2.18}$$

Für den empirischen Korrelationskoeffizienten r_{ij} gilt:

1. $-1\le r_{ij}\le 1$,
2. $r_{ij}=1$ genau dann, wenn zwischen den beiden Merkmalen ein exakter linearer Zusammenhang mit positiver Steigung besteht,
3. $r_{ij}=-1$ genau dann, wenn zwischen den beiden Merkmalen ein exakter linearer Zusammenhang mit negativer Steigung besteht.

Wir wollen diese Eigenschaften hier nicht beweisen. Wir beweisen sie in Kapitel 3 für den Korrelationskoeffizienten. Hier wollen wir diese Eigenschaften aber interpretieren. Die erste Eigenschaft besagt, dass der empirische Korrelationskoeffizient Werte zwischen -1 und 1 annimmt, während die beiden anderen Eigenschaften erklären, wie wir die Werte des empirischen Korrelationskoeffizienten zu interpretieren haben. Liegt der Wert des empirischen Korrelationskoeffizienten in der Nähe von 1, so liegt ein positiver linearer Zusammenhang zwischen den beiden Merkmalen vor, während ein Wert in der Nähe von -1 auf einen negativen linearen Zusammenhang hindeutet. Ein Wert in der Nähe von 0 spricht dafür, dass kein linearer Zusammenhang zwischen den beiden Merkmalen vorliegt. Dies bedeutet aber nicht notwendigerweise, dass gar kein Zusammenhang zwischen den beiden Merkmalen besteht, wie das Beispiel in Tabelle 2.3 zeigt. Der Wert des Korrelationskoeffizienten zwischen den beiden Merkmalen beträgt 0. Schaut man sich die Werte in der Tabelle genauer an, so stellt man fest, dass $x_{k2}=x_{k1}^2$ gilt. Zwischen den beiden Merkmalen besteht also ein funktionaler Zusammenhang.

Table 2.3. Werte der Merkmale x_1 und x_2

k	x_{k1}	x_{k2}
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

Wir stellen die Korrelationen in der *empirischen Korrelationsmatrix* $\mathbf{R}$ zusammen:

$$\mathbf{R} = \begin{pmatrix} r_{11} & \dots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \dots & r_{pp} \end{pmatrix}. \tag{2.19}$$

hmcounterend. (fortgesetzt)

Example 15. Es gilt

$$\mathbf{R} = \begin{pmatrix} 1 & 0.916 & 0.953 \\ 0.916 & 1 & 0.932 \\ 0.953 & 0.932 & 1 \end{pmatrix}.$$

Es fällt auf, dass alle Elemente der empirischen Korrelationsmatrix positiv sind. □

Man kann die empirische Korrelationsmatrix auch mit Hilfe der Matrix der standardisierten Merkmale (2.14) bestimmen. Es gilt

$$\mathbf{R} = \frac{1}{n-1}\,\mathbf{X}^{*\prime}\mathbf{X}^{*}. \tag{2.20}$$

Das Element in der i-ten Zeile und j-ten Spalte von $\mathbf{X}^{*\prime}\mathbf{X}^{*}$ erhält man dadurch, dass man das innere Produkt aus den Vektoren bildet, die in der i-ten und der j-ten Spalte von $\mathbf{X}^{*}$ stehen. Dieses ist

$$\sum_{k=1}^{n} \frac{x_{ki} - \bar{x}_i}{s_i} \frac{x_{kj} - \bar{x}_j}{s_j} = \frac{1}{s_i\, s_j} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)\,(x_{kj} - \bar{x}_j).$$

Dividiert man diesen Ausdruck durch $n-1$, so erhält man den empirischen Korrelationskoeffizienten zwischen dem i-ten und j-ten Merkmal, wie man durch einen Vergleich mit Gleichung (2.18) erkennt. In der empirischen Korrelationsmatrix sind die Zusammenhänge zwischen allen Paaren von Merkmalen zusammengefasst. Eine hierzu analoge graphische Darstellung ist die *Streudiagrammmatrix*. Hier werden die Streudiagramme aller Paare von Merkmalen in einer Matrix zusammengefasst. hmcounterend. (fortgesetzt)

Example 15. Die Streudiagrammmatrix ist in Abbildung 2.8 zu finden. Wir sehen hier auf einen Blick, dass alle Merkmale miteinander positiv korreliert sind. □

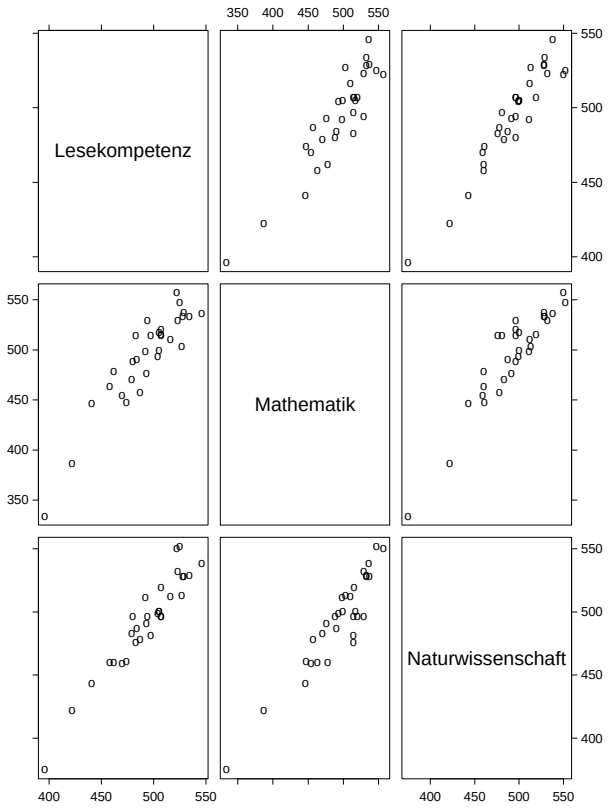


Fig. 2.8. Streudiagrammmatrix der drei Merkmale im Rahmen der PISA-Studie

Bisher haben wir mit Hilfe von Streudiagrammen versucht herauszufinden, welcher Zusammenhang zwischen zwei Merkmalen besteht. Die Objekte, an denen die Merkmale erhoben wurden, waren nicht von Interesse. Mit diesen wollen wir uns nun aber auch beschäftigen. hmcounterend. (fortgesetzt)

Example 15. Abbildung 2.9 zeigt das Streudiagramm der Merkmale `Lesekompetenz` und `Mathematische Grundbildung`, wobei wir aber an die Koordinaten jedes Landes den Namen des Landes schreiben. Wir sehen nun sehr schön, wo die einzelnen Länder liegen. Will man eine graphische Darstellung hinsichtlich aller drei Merkmale, so könnte man eine dreidimensionale Graphik erstellen. Bei mehr als drei Merkmalen ist eine direkte graphische

Darstellung der Objekte hinsichtlich aller Merkmale nicht mehr möglich. Wir werden aber Verfahren kennenlernen, die eine interpretierbare Darstellung von Objekten in einem zweidimensionalen Streudiagramm ermöglichen. □

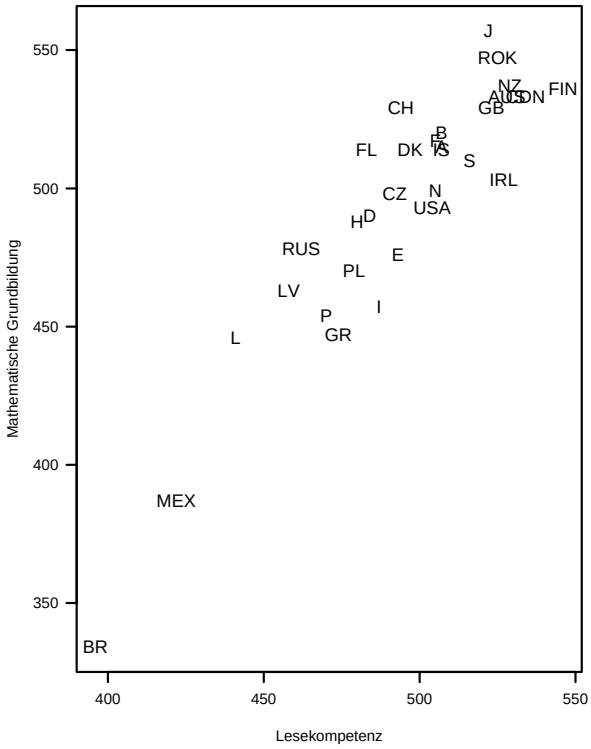


Fig. 2.9. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

2.2.2 Beschreibung und Darstellung von Datenmatrizen qualitativer Merkmale

Example 16. Im Beispiel 2 auf Seite 3 wurde eine Reihe qualitativer Merkmale erhoben. Die Datenmatrix ist in (2.2) auf Seite 14 zu finden. Wir wählen von dieser die Spalten 1, 2 und 4 mit den Merkmalen `Geschlecht`, `MatheLK` und `Abitur88` aus. □

Bei nur einem Merkmal haben wir eine Häufigkeitstabelle erstellt. Dies wird auch der erste Schritt bei mehreren qualitativen Merkmalen sein. Die klassis-

che Form der Darstellung einer (n, p)-Datenmatrix, die nur qualitative Merkmale enthält, ist die *Kontingenztabelle*. Eine Kontingenztabelle ist nichts anderes als eine Häufigkeitstabelle mehrerer qualitativer Merkmale. Schauen wir uns diese zunächst für zwei qualitative Merkmale A und B an. Wir bezeichnen die Merkmalsausprägungen von A mit $A_1, A_2, \ldots, A_I$ und die Merkmalsausprägungen von B mit $B_1, B_2, \ldots, B_J$. Wie im univariaten Fall bestimmen wir absolute Häufigkeiten, wobei wir aber die beiden Merkmale gemeinsam betrachten. Sei n_{ij} die Anzahl der Objekte, die beim Merkmal A die Ausprägung A_i und beim Merkmal B die Ausprägung B_j aufweisen. Tabelle 2.4 zeigt den allgemeinen Aufbau einer zweidimensionalen Kontingenztabelle. hmcounterend. (fortgesetzt)

Table 2.4. Allgemeiner Aufbau einer zweidimensionalen Kontingenztabelle

B / A	B_1	B_2	$\ldots$	B_J
A_1	n_{11}	n_{12}	$\cdots$	n_{1J}
A_2	n_{21}	n_{22}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}

Example 16. Sei A das Merkmal `Geschlecht` und B das Merkmal `MatheLK`. Fassen wir bei beiden Merkmalen die 0 als erste Merkmalsausprägung und die 1 als zweite Merkmalsausprägung auf, so gilt

$$n_{11} = 5, \quad n_{12} = 5, \quad n_{21} = 4, \quad n_{22} = 6.$$

Tabelle 2.5 zeigt die Kontingenztabelle. □

Table 2.5. Kontingenztabelle der Merkmale Geschlecht und MatheLK

`Geschlecht` / `MatheLK`	0	1
0	5	5
1	4	6

Die absoluten Häufigkeiten der Merkmalsausprägungen der univariaten Merkmale erhalten wir durch Summierung der Elemente der Zeilen bzw. Spalten. Wir bezeichnen die absolute Häufigkeit der Merkmalsausprägung A_i mit $n_{i\cdot}$ und die absolute Häufigkeit der Merkmalsausprägung B_j mit $n_{\cdot j}$. Es gilt

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$$

und

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij}.$$

hmcounterend. (fortgesetzt)

Example 16. Es gilt $n_{1\cdot} = 10$, $n_{2\cdot} = 10$, $n_{\cdot 1} = 9$ und $n_{\cdot 2} = 11$. □

Es ist von Interesse, ob zwischen den beiden Merkmalen ein Zusammenhang besteht. Hierzu schaut man sich zunächst die *bedingten relativen Häufigkeiten* an. Dies bedeutet, dass man unter der Bedingung, dass die einzelnen Kategorien des Merkmals A gegeben sind, die Verteilung des Merkmals B bestimmt. hmcounterend. (fortgesetzt)

Example 16. Wir betrachten zunächst nur die Männer. Von den 10 Männern haben 5 den Mathematik-Leistungskurs besucht, also 50 Prozent. Von den 10 Frauen haben 6 den Mathematik-Leistungskurs besucht, also 60 Prozent. Wir sehen, dass sich diese Häufigkeiten unterscheiden. Es stellt sich die Frage, ob dieser Unterschied signifikant ist. Wir werden diese Frage im Kapitel 10 über loglineare Modelle beantworten. □

Für die bedingte relative Häufigkeit der Merkmalsausprägung B_j unter der Bedingung, dass die Merkmalsausprägung A_i gegeben ist, schreiben wir $h_{j|i}$. Offensichtlich gilt

$$h_{j|i} = \frac{n_{ij}}{n_{i\cdot}}.$$

Den allgemeinen Aufbau einer Tabelle mit bedingten relativen Häufigkeiten zeigt Tabelle 2.6. Die Zeilen dieser Tabelle bezeichnet man auch als *Profile*.

Table 2.6. Allgemeiner Aufbau einer Kontingenztabelle mit bedingten relativen Häufigkeiten

B / A	B_1	B_2	$\dots$	B_J
A_1	$h_{1\|1}$	$h_{2\|1}$	$\cdots$	$h_{J\|1}$
A_2	$h_{1\|2}$	$h_{2\|2}$	$\cdots$	$h_{J\|2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_I	$h_{1\|I}$	$h_{2\|I}$	$\cdots$	$h_{J\|I}$

hmcounterend. (fortgesetzt)

Table 2.7. Kontingenztabelle der Merkmale Geschlecht und MatheLK mit bedingten relativen Häufigkeiten

MatheLK	0	1
Geschlecht		
0	0.5	0.5
1	0.4	0.6

Example 16. Für das Beispiel erhalten wir die bedingten relativen Häufigkeiten in Tabelle 2.7. □

Man kann natürlich auch die Verteilung von A unter der Bedingung bestimmen, dass die einzelnen Kategorien von B gegeben sind. Dies wollen wir aber nicht im Detail ausführen.

Die Kontingenztabelle von zwei qualitativen Merkmalen ist ein Rechteck. Nimmt man ein weiteres Merkmal hinzu, so erhält man einen Quader. Diesen stellt man nun nicht dreidimensional, sondern mit Hilfe von Schnitten zweidimensional dar. Gegeben seien also die qualitativen Merkmale A, B und C mit den Merkmalsausprägungen $A_1, \ldots, A_I$, $B_1, \ldots, B_J$ und $C_1, \ldots, C_K$. Dann ist n_{ijk} die absolute Häufigkeit des gemeinsamen Auftretens von A_i, B_j und C_k, $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$. Tabelle 2.8 beinhaltet den allgemeinen Aufbau einer dreidimensionalen Kontingenztabelle. hmcounterend. (fortgesetzt)

Table 2.8. Allgemeiner Aufbau einer dreidimensionalen Kontingenztabelle

		B		
C	A	B_1	$\ldots$	C_J
C_1	A_1	n_{111}	$\ldots$	n_{1J1}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_I	n_{I11}	$\ldots$	n_{IJ1}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
C_K	A_1	n_{11K}	$\ldots$	n_{IJK}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	A_I	n_{I1K}	$\ldots$	n_{IJK}

Example 16. Die dreidimensionale Kontingenztabelle der Merkmale `Geschlecht`, `MatheLK` und `Abitur88` ist in Abbildung 2.9 zu finden. □

Aus einer dreidimensionalen Tabelle kann man durch Summation über die Häufigkeiten eines Merkmals drei zweidimensionale Kontingenztabellen erhalten. hmcounterend. (fortgesetzt)

Table 2.9. Dreidimensionale Kontingenztabelle der Merkmale Geschlecht, MatheLK und Abitur88

		MatheLK	
Abitur88	Geschlecht	0	1
0	0	5	4
	1	1	3
1	0	0	1
	1	3	3

Example 16. Wir haben die Kontingenztabelle der Merkmale `Geschlecht` und `MatheLK` bereits erstellt. Sie ist in Tabelle 2.5 auf Seite 42 zu finden. Die beiden anderen Tabellen sind in den Abbildungen 2.10 und 2.11 zu finden.

Table 2.10. Kontingenztabelle der Merkmale Geschlecht und Abitur88

Abitur88	0	1
Geschlecht		
0	9	1
1	4	6

Table 2.11. Kontingenztabelle der Merkmale Abitur88 und MatheLK

MatheLK	0	1
Abitur88		
0	6	7
1	3	4

Schaut man sich die entsprechenden bedingten relativen Häufigkeiten an, so sieht es so aus, als ob zwischen den Merkmalen `Geschlecht` und `Abitur88` ein Zusammenhang besteht, während zwischen den Merkmalen `Abitur88` und `MatheLK` kein Zusammenhang zu bestehen scheint. □

Wir haben bisher nur die zweidimensionalen Kontingenztabellen betrachtet, die man aus einer dreidimensionalen Kontingenztabelle gewinnen kann und deskriptiv auf Zusammenhänge untersucht. In dreidimensionalen Kontingenztabellen können aber noch komplexere Zusammenhänge existieren. Mit diesen werden wir uns detailliert im Kapitel 10 im Zusammenhang mit loglinearen Modellen beschäftigen.

2.3 Datenbehandlung in S-PLUS

2.3.1 Univariate Datenanalyse

Quantitative Merkmale Wir wollen nun lernen, wie man in S-PLUS Daten elementar analysiert. S-PLUS bietet eine interaktive Umgebung, Befehlsmodus genannt, in der man die Daten direkt eingeben und analysieren kann. Durch das Bereitschaftszeichen > wird angezeigt, dass eine Eingabe erwartet wird. Der Befehlsmodus ist ein mächtiger Taschenrechner. Wir können hier die Grundrechenarten Addition, Subtraktion, Multiplikation und Division mit den Operatoren +, -, * und / durchführen:

```
> 3+4
[1] 7
> 3-4
[1] -1
> 3*4
[1] 12
> 3/4
[1] 0.75
```

Zum Potenzieren benutzen wir ^ :

```
> 3^4
[1] 81
```

Man kann aber auch komplizierte Analysen durchführen. Wir wollen die Punkte aller Länder im Bereich Mathematische Grundbildung aus dem Beispiel 14 auf Seite 17 analysieren, die wir hier noch einmal wiedergeben:

```
533 520 334 514 490 536 517 447 529 503 514 457 557
533 547 463 514 446 387 537 499 515 470 454 478 510
529 476 498 488 493 .
```

Die Standarddatenstruktur in S-PLUS ist der Vektor. Ein Vektor ist eine Zusammenfassung von Objekten zu einer endlichen Folge. Einen Vektor erstellt man mit der Funktion c. Diese macht aus einer Folge von Zahlen, die durch Kommata getrennt sind, einen Vektor, dessen Komponenten die einzelnen Zahlen sind. Die Zahlen sind die Argumente der Funktion c. Argumente einer Funktion stehen in runden Klammern hinter dem Funktionsnamen und sind durch Kommata voneinander getrennt. Der Aufruf

```
> c(533,520,334,514,490,536,517,447,529,503,514,457,557,
    533,547,463,514,446,387,537,499,515,470,454,478,510,
    529,476,498,488,493)
```

liefert am Bildschirm folgendes Ergebnis:

```
[1] 533 520 334 514 490 536 517 447 529 503 514 457 557
    533 547 463 514 446 387 537 499 515 470 454 478 510
    529 476 498 488 493
```

Die Elemente des Vektors werden ausgegeben. Am Anfang steht `[1]`. Dies zeigt, dass die erste Zahl gleich der ersten Komponente des Vektors ist. Um mit den Werten weiterhin arbeiten zu können, müssen wir sie in einer Variablen speichern. Dies geschieht mit dem Zuweisungsoperator `<-`, den man durch die Zeichen `<` und `-` erhält. Auf der linken Seite steht der Name der Variablen, der die Werte zugewiesen werden sollen, auf der rechten Seite steht der Aufruf der Funktion `c`. Die Namen von Variablen dürfen beliebig lang sein, dürfen aber nur aus Buchstaben, Ziffern und dem Punkt bestehen, wobei das erste Zeichen ein Buchstabe oder der Punkt sein muss. Beginnt ein Name mit einem Punkt, so dürfen nicht alle folgenden Zeichen Ziffern sein. Hierdurch erzeugt man nämlich eine Zahl. Wir nennen die Variable `Mathe`. `S-PLUS` unterscheidet Groß- und Kleinschreibung. Die Variablennamen `Mathe` und `mathe` beziehen sich also auf unterschiedliche Objekte. Wir geben ein

```
> Mathe<-c(533,520,334,514,490,536,517,447,529,503,514,
           457,557,533,547,463,514,446,387,537,499,515,
           470,454,478,510,529,476,498,488,493)
```

Den Inhalt einer Variablen kann man sich durch Eingabe des Namens anschauen. Der Aufruf

```
> Mathe
```

liefert das Ergebnis

```
 [1] 533 520 334 514 490 536 517 447 529 503 514 457 557
     533 547 463 514 446 387 537 499 515 470 454 478 510
     529 476 498 488 493
```

Man kann gleichlange Vektoren mit Operatoren verknüpfen. Dabei wird der Operator auf die entsprechenden Komponenten der Vektoren angewendet. Man kann aber auch einen Skalar mit einem Vektor über einen Operator verknüpfen. Dabei wird der Skalar mit jeder Komponente des Vektors über den Operator verknüpft. Will man also wissen, wie sich jede Komponente des Vektors `Mathe` von der Zahl 500 unterscheidet, so gibt man ein

```
> Mathe-500
 [1] 33 20 -166 14 -10 36 17 -53 29 3 14 -43 57 33 47 -37
     14 -54 -113 37 -1 15 -30 -46 -22 10 29 -24 -2 -12 -7
```

Auf Komponenten eines Vektors greift man durch Indizierung zu. Hierzu gibt man den Namen des Vektors gefolgt von eckigen Klammern ein, zwischen denen die Nummer der Komponente steht, auf die man zugreifen will. Will man also die Punkte des zweiten Landes wissen, so gibt man ein

```
> Mathe[2]
```

und erhält als Ergebnis

```
[1] 520
```

Will man auf die letzte Komponente zugreifen, so benötigt man die Länge des Vektors. Diese liefert die Funktion `length`:

```
> length(Mathe)
[1] 31
```

Die letzte Komponente des Vektors `Mathe` erhalten wir also durch

```
> Mathe[length(Mathe)]
[1] 493
```

Auf mehrere Komponenten eines Vektors greift man zu, indem man einen Vektor mit den Nummern der Komponenten bildet und mit diesem indiziert. So erhält man die Punkte der ersten drei Länder durch

```
> Mathe[c(1,2,3)]
[1] 533 520 334
```

Wir können auf Komponenten, die hintereinander stehen, einfacher zugreifen. Sind i und j natürliche Zahlen mit $i < j$, so liefert in `S-PLUS` der Ausdruck

```
i:j
```

die Zahlenfolge $i, i+1, \ldots, j-1, j$. Ist $i > j$, so erhalten wir die Zahlenfolge $i, i-1, \ldots, j+1, j$. Wollen wir also auf die ersten drei Komponenten von `Mathe` zugreifen, so geben wir ein

```
> Mathe[1:3]
[1] 533 520 334
```

Wollen wir den Vektor `Mathe` in umgekehrter Reihenfolge ausgeben, so geben wir ein

```
> Mathe[length(Mathe):1]
 [1] 493 488 498 476 529 510 478 454 470 515 499 537 387
     446 514 463 547 533 557 457 514 503 529 447 517 536
     490 514 334 520 533
```

Mit der Funktion `rev` hätten wir das gleiche Ergebnis erhalten.

Oft will man Komponenten eines Vektors selektieren, die bestimmte Eigenschaften besitzen. Hierzu benötigt man Vergleichsoperatoren, mit denen man auf Gleichheit mit `==`, Ungleichheit mit `!=`, kleiner mit `<`, kleiner gleich mit `<=`, größer mit `>` oder größer gleich mit `>=` überprüfen kann. Das Ergebnis des Vergleichs ist vom Typ `logical`, ist also entweder `T` oder `F`, wobei `T` für true und `F` für false steht:

```
> 3<4
[1] T
```

Man kann natürlich auch einen Vektor der Länge n und einen Skalar mit einem Vergleichsoperator verknüpfen:

```
> 1:5 <= 3
[1] T T T F F
```

Indiziert man einen Vektor der Länge n mit einem Vektor vom Typ `logical` der Länge n, so werden die Komponenten ausgewählt, bei denen im Vektor vom Typ `logical` ein `T` steht. Wollen wir die Punktezahlen der Länder wissen, die weniger als 480 Punkte erreicht haben, so geben wir ein

```
> Mathe[Mathe<480]
 [1] 334 447 457 463 446 387 470 454 478 476
```

Die Nummern der Länder erhalten wir durch

```
> (1:length(Mathe))[Mathe<480]
 [1]  3  8 12 16 18 19 23 24 25 28
```

Die Funktion `sum` bestimmt die Summe der Komponenten eines Vektors. Sind diese vom Typ `logical`, so wird `F` in 0 und `T` in 1 umgewandelt. Der Aufruf

```
> sum(Mathe<480)
[1] 10
```

liefert also die Anzahl der Länder mit weniger als 480 Punkten. Sollen mehrere Bedingungen erfüllt sein, so kann man die logischen Operatoren `&` und `|` verwenden. Der Operator `&` entspricht dem logischen `"und"` und der Operator `|` entspricht dem logischen `"oder"`. Die Indizes der Länder, die mindestens 490 und höchstens 510 Punkte erreicht haben, erhalten wir durch

```
> (1:length(Mathe))[Mathe>=490 & Mathe<=510]
[1]  5 10 21 26 29 31
```

In `S-PLUS` gibt es eine Vielzahl von Funktionen. Von diesen haben wir die Funktionen `c`, `sum`, `length` und `rev` kennengelernt. Mit den Funktionen `sum` und `length` können wir den Mittelwert folgendermaßen bestimmen:

```
> sum(Mathe)/length(Mathe)
[1] 493.1613
```

In `S-PLUS` gibt es zur Bestimmung des Mittelwerts die Funktion `mean`. Für die Variable `Mathe` erhalten wir

```
> mean(Mathe)
[1] 493.1613
```

Mit der Funktion `mean` kann man aber nicht nur den Mittelwert bestimmen.

Schauen wir uns die Funktion an:

```
> mean
 function(x, trim = 0, na.rm = F) {
     if(na.rm) {
         wnas <- which.na(x)
         if(length(wnas))
             x <- x[ - wnas]
     }
     if(mode(x) == "complex") {
         if(trim > 0)
             stop("trimming not allowed for complex data")
         return(sum(x)/length(x))
     }
     x <- as.double(x)
     if(trim > 0) {
         if(trim >= 0.5)
             return(median(x, na.rm = F))
         if(!na.rm && length(which.na(x)))
             return(NA)
         n <- length(x)
         i1 <- floor(trim * n) + 1
         i2 <- n - i1 + 1
         x <- sort(x, unique(c(i1, i2)))[i1:i2]
     }
     sum(x)/length(x)
}
```

Wie jede Funktion in S-PLUS besteht `mean` aus einem Kopf und einem Körper.

Der Funktionskopf besteht aus dem Namen der Funktion gefolgt von den Argumenten der Funktion, die in runden Klammern stehen und durch Kommata getrennt sind. Die Funktion `mean` hat die drei Argumente `x`, `trim` und `na.rm`. Die Argumente `trim` und `na.rm` sind in dem Sinne fakultativ, dass ihnen beim Aufruf der Funktion vom Benutzer keine Werte zugewiesen werden müssen. Sie sind vorbelegt durch `0` im Falle von `trim` und von `F` im Falle von `na.rm`. Das Argument `x` hingegen muss der Funktion `mean` übergeben werden. Hierzu haben wir zwei Möglichkeiten. Wir können die Funktion aufrufen mit

```
> mean(x=Mathe)
```

Hierdurch werden der lokalen Variablen `x` beim Aufruf von `mean` die Werte der Variablen `Mathe` zugewiesen. Wie wir weiter oben gesehen haben, können wir aber auch eingeben

```
> mean(Mathe)
```

Hierbei müssen die Argumente nur an der richtigen Position stehen. Da das erste Argument der Funktion `mean` der Datenvektor `x` ist, werden die Werte von `Mathe` für diesen eingesetzt.

Der Funktionskörper besteht aus den Anweisungen. Die Anweisungen stehen in geschweiften Klammern. Wir sehen, dass die Funktion `mean` aus einer Reihe von Anweisungen besteht. Wir wollen diese hier nicht alle diskutieren, sondern nur bemerken, dass durch das Argument `trim` ein getrimmter Mittelwert bestimmt werden kann und dass mit Hilfe des Arguments `na.rm` gesteuert werden kann, was mit fehlenden Beobachtungen bei der Berechnung des Mittelwerts geschehen soll. Für jede fehlende Beobachtung gibt man den Wert `NA` ein. Liegen keine fehlenden Beobachtungen vor und soll auch nicht getrimmt werden, so wird nur der letzte Befehl der Funktion `mean` ausgeführt:

```
sum(x)/length(x)
```

Diesen kennen wir bereits. Da `S-PLUS` das Ergebnis des letzten Ausdrucks einer Funktion als Ergebnis der Funktion zurückgibt, liefert die Funktion `mean` den Mittelwert als Ergebnis.

Wir haben schon erwähnt, dass man mit der Funktion `mean` auch getrimmte Mittelwerte bestimmen kann. Man muss in diesem Fall nur dem Argument `trim` den gewünschten Trimmanteil α zuweisen. Der Aufruf

```
> mean(Mathe,0.05)
```

liefert das Ergebnis

```
[1] 496.4483
```

Man kann mit der Funktion `mean` auch den Median bestimmen. Man muss nur das Argument `trim` auf 0.5 setzen:

```
> mean(Mathe,0.5)
[1] 503
```

Schaut man sich den Inhalt der Funktion `mean` an, so sieht man, dass in diesem Fall die Funktion `median` aufgerufen wird. Wir können den Median also direkt bestimmen durch

```
> median(Mathe)
[1] 503
```

Schauen wir uns die Stelle in der Funktion `mean` an, an der die Funktion `median` aufgerufen wird. Sie lautet

```
if(trim >= 0.5)
    return(median(x, na.rm = F))
```

Hierbei handelt es sich um eine bedingte Anweisung. Es wird die Bedingung `trim>=0.5` überprüft. Ist das Argument von `if` gleich `T`, so wird die Anweisungsfolge ausgeführt, die hinter dem Ausdruck `trim>=0.5` steht. Dabei besteht eine Anweisungsfolge in `S-PLUS` aus einer Folge von Anweisungen,

die von geschweiften Klammern umgeben sind. Liegt nur ein Befehl vor, so kann man auf die Klammern verzichten. Ist das Argument von `if` gleich `F`, so wird die Anweisungsfolge übersprungen, die hinter dem Ausdruck `trim>=0.5` steht, und der hinter dieser Befehlsfolge stehende Befehl wird ausgeführt. Ist `trim` also größer oder gleich 0.5, so wird der Befehl `return(median(x, na.rm = F))` ausgeführt. Es wird der Median berechnet und als Ergebnis der Funktion `mean` zurückgegeben. Der Ausdruck `return(x)` bewirkt, dass die Ausführung einer Funktion beendet wird, und `x` als Ergebnis der Funktion zurückgegeben wird.

Kehren wir zu den Funktionen zurück, mit denen man Daten analysieren kann. Die Varianz einer Variablen erhält man mit der Funktion `var`:

```
> var(Mathe)
[1] 2192.873
```

Für die Standardabweichung gibt es keine eigene Funktion in `S-PLUS`. Man kann sich aber eine eigene Funktion schreiben. Die Standardabweichung ist die Wurzel aus der Varianz. Die Funktion `sqrt` bestimmt die Wurzel. Wir erhalten die Standardabweichung also durch

```
> sqrt(var(Mathe))
[1] 46.82812
```

Wir wollen nun eine Funktion `std` schreiben, die die Standardabweichung der Elemente eines Objekts `x` bestimmt. Eine Funktion wird durch folgende Befehlsfolge deklariert:

```
fname<-function(Argumente}
{
   Koerper der Funktion
   return(Ergebnis)
}
```

Wir geben also ein

```
std<-function(x)
{
   return(sqrt(var(x)))
}
```

Wir können die Funktion sofort benutzen:

```
> std(Mathe)
[1] 46.82812
```

Der Zweck der Funktion `std` ist ersichtlich, aber auch hier verbessert die Verwendung von Kommentaren die Lesbarkeit.

Hier ist die kommentierte Version von `std`:

```
std<-function(x)
{
# Standardabweichung der Elemente von x
   return(sqrt(var(x)))
}
```

Bei der Beschreibung eines univariaten Merkmals haben wir auch die Fünf-Zahlen-Zusammenfassung betrachtet. Die Funktion `summary` bestimmt das Minimum $x_{(1)}$, das untere Quartil $x_{0.25}$, den Median $x_{0.5}$, das obere Quartil $x_{0.75}$ und das Maximum $x_{(n)}$. Der Aufruf

```
> summary(Mathe)
```

liefert das Ergebnis

```
Min. 1st Qu. Median  Mean 3rd Qu. Max.
 334     473    503 493.2   524.5  557
```

Wir sehen, dass neben den 5 Zahlen auch noch der Mittelwert bestimmt wird. Die Zahlen stimmen mit denen überein, die wir weiter oben bestimmt haben. `S-PLUS` liefert aber nicht für jeden Stichprobenumfang die Quartile so, wie es auf Seite 20 beschrieben wird. Der Aufruf

```
> summary(1:6)
```

liefert das Ergebnis

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
   1    2.25    3.5  3.5    4.75    6
```

Bei Tukey nimmt das untere Quartil den Wert 2 an. Hyndman & Fan (1996) geben an, wie `S-PLUS` die Quartile bestimmt. Wir wollen hierauf aber nicht eingehen. Im Anhang ist auf Seite 495 eine Funktion `quartile` zu finden, die die Quartile so bestimmt, wie es auf Seite 20 beschrieben wird.:

```
> quartile(1:6)
[1] 2 5
```

Mit Hilfe der 5 Zahlen kann man einen Boxplot erstellen. Der Aufruf

```
> boxplot(Mathe)
```

liefert die Abbildung 2.10. Der Boxplot sieht nicht so aus wie in Abbildung 2.3 auf Seite 22. Die Beschriftung der Ordinate unterscheidet sich in beiden Abbildungen. Wir sind eine Beschriftung wie in Abbildung 2.3 gewohnt. Diese erreichen wir, indem wir den Graphikparameter `las` auf den Wert 1 setzen. Damit der Boxplot wie in Abbildung 2.3 aussieht, müssen wir einige Argumente der Funktion `boxplot` mit speziellen Werten aufrufen. Der folgende Aufruf liefert den Boxplot in Abbildung 2.3:

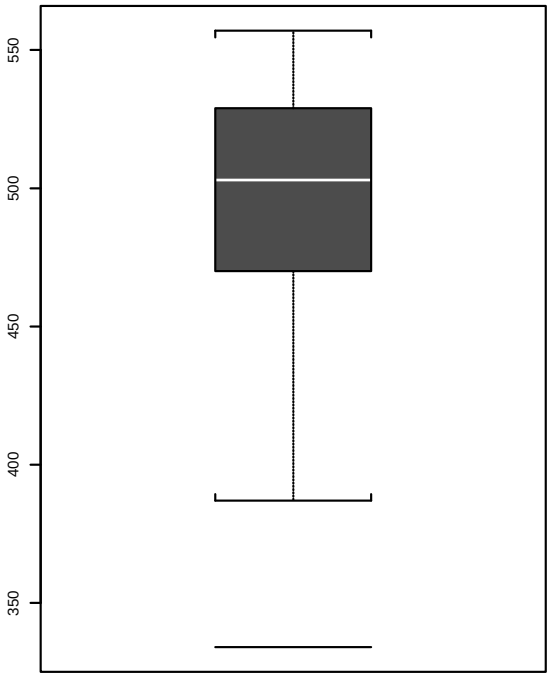


Fig. 2.10. Boxplot des Merkmals Mathematische Grundbildung

```
> par(las=1)
> boxplot(Mathe,names="Mathematische Grundbildung",boxcol=0,
   medline=T,medcol=1,outline=F,outpch="*",medlwd=0.5,col=1)
```

Das Argument `names` ist eine Zeichenkette. Eine Zeichenkette ist eine Folge von Zeichen, die in Hochkommata stehen. Wir werden uns gleich mit Zeichenketten beschäftigen. Schauen wir uns vorher die Befehlsfolge an, die das Histogramm in Abbildung 2.2 auf Seite 19 liefert:

```
>  hist(Mathe,prob=T,xlab="Mathematische Grundbildung")
```

Durch `prob=T` stellen wir sicher, dass die Fläche unter dem Histogramm gleich 1 ist. Setzen wir `prob` auf `F`, so haben die Rechtecke die Höhe der absoluten Häufigkeiten. `S-PLUS` wählt standardmäßig gleich große Klassen. Die Anzahl der Klassen ist proportional zu $\ln n$.

Qualitative Merkmale Wir wollen die Analyse des Merkmals `MatheLK` aus Beispiel 13 auf Seite 15 in `S-PLUS` nachvollziehen. Das Merkmal `MatheLK`

kann die Werte `j` und `n` annehmen. Hier sind noch einmal die Werte der 20 Studenten:

```
n n n n n n n n n j j j j j j j j j j
```

Wir wollen diese Werte der Variablen `MatheLK` zuweisen. Hierzu erzeugen wir uns einen Vektor der Länge 20, dessen Komponenten Zeichenketten sind. Die ersten 9 Komponenten sollen die Zeichenkette `"n"` und die letzten 11 Komponenten die Zeichenkette `"j"` enthalten. Um uns die Eingabe zu erleichtern, verwenden wir die Funktion `rep`. Der Aufruf

```
rep(x,times)
```

erzeugt einen Vektor, in dem das Argument `x` `times`-mal wiederholt wird:

```
> rep("n",9)
[1] "n" "n" "n" "n" "n" "n" "n" "n" "n"
```

Wir erzeugen den Vektor `MatheLK` also durch

```
> MatheLK<-c(rep("n",9),rep("j",11))
```

Schauen wir uns `MatheLK` an:

```
> MatheLK
 [1] "n" "j" "n" "n" "n" "n" "n" "n" "n" "n"
     "j" "j" "j" "j" "j" "j" "j" "j" "j" "j"
```

Das Merkmal `MatheLK` ist nominalskaliert. Ein nominalskaliertes qualitatives Merkmal ist in `S-PLUS` ein *Faktor*. Ein Faktor wird erzeugt mit der Funktion `factor`:

```
> MatheLK<-factor(MatheLK)
> MatheLK
 [1] n n n n n n n n n j j j j j j j j j j
```

Ein ordinalskaliertes qualitatives Merkmal ist in `S-PLUS` ein *geordneter Faktor*. Diesen erzeugt man mit der Funktion `ordered`. Die absoluten Häufigkeiten der Merkmalsausprägungen erhalten wir mit der Funktion `table`. Der Aufruf

```
> table(MatheLK)
```

liefert das Ergebnis

```
 j n
11 9
```

Die relativen Häufigkeiten erhalten wir, indem wir das Ergebnis der Funktion `table` durch die Anzahl der Beobachtungen teilen:

```
> table(MatheLK)/length(MatheLK)
    j    n
0.55 0.45
```

Mit der Funktion `barplot` erstellen wir das Stabdiagramm. Der Aufruf

```
> barplot(table(MatheLK)/length(MatheLK))
```

liefert die Abbildung 2.11. Wir sehen, dass bei diesem Stabdiagramm im

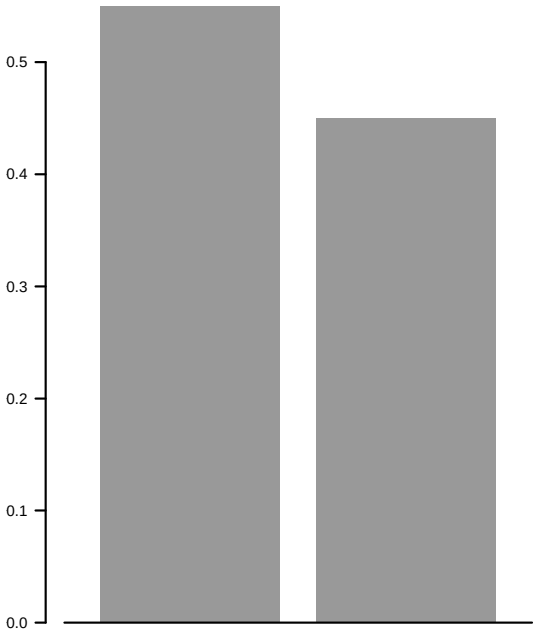


Fig. 2.11. Stabdiagramm des Merkmals MatheLK mit breiten Balken und ohne Achsenbeschriftung

Gegensatz zum Stabdiagramm in Abbildung 2.1 auf Seite 17 die Stäbe sehr breit sind. Außerdem fehlt die Achsenbeschriftung. Die Breite der Stäbe wird mit dem Parameter `space` festgelegt. Dieser gibt das Verhältnis aus dem Zwischenraum zwischen den Balken zur Breite der Balken an. Setzt man im Beispiel das Argument `space` auf den Wert 20, so erhält man die Balkenbreite in Abbildung 2.1. Die Achsenbeschriftung erhalten wir, indem wir dem Parameter `names` die Merkmalsausprägungen als Zeichenkettenvektor übergeben.

2.3.2 Multivariate Datenanalyse

Quantitative Merkmale Nun wollen wir die Merkmale `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` aus dem Beispiel 15 auf Seite 24 gemeinsam analysieren. Hierzu geben wir die Daten in Form einer Matrix ein. In `S-PLUS` erzeugt man eine Matrix mit der Funktion `matrix`. Der Aufruf von `matrix` ist

```
matrix(data,nrow=1,ncol=1,byrow=F)
```

Dabei ist `data` der Vektor mit den Elementen der Matrix. Das Argument `nrow` gibt die Anzahl der Zeilen und das Argument `ncol` die Anzahl der Spalten der Matrix an. Standardmäßig wird eine Matrix spaltenweise eingegeben. Sollen die Zeilen aufgefüllt werden, so muss das Argument `byrow` auf den Wert `T` gesetzt werden. Wir weisen die Punkte der 31 Länder der Matrix `PISA` zu, wobei wir hier die Daten verkürzt wiedergeben. Die drei Punkte stehen für die restlichen 87 Beobachtungen:

```
> PISA<-matrix(c(528,507,396,...,511,496,499),31,3)
```

Wir wollen nun noch den Zeilen und Spalten der Matrix `PISA` Namen geben. Dies geschieht mit der Funktion `dimnames`. Der Aufruf von `dimnames` für eine Matrix `mat` ist

```
> dimnames(mat)<-list(ZN,SN)
```

Dabei sind `ZN` und `SN` Vektoren mit den Namen der Zeilen beziehungsweise Spalten der Matrix `mat`. In der Regel werden dies Vektoren sein, die Zeichenketten enthalten. Die Funktion `list` verbindet ihre Argumente zu einer Liste. Eine Liste besteht aus Komponenten, die unterschiedliche `S-PLUS`-Objekte sein können. In einer Liste kann man zum Beispiel Vektoren und Matrizen zu einem Objekt zusammenfassen. Schauen wir uns dies für das Beispiel an. Wir erzeugen zunächst einen Vektor `laender` mit den Namen der Länder, wobei wir die Ländernamen durch die Autokennzeichen abkürzen:

```
> laender<-c("AUS","B","BR","DK","D","FIN","F","GR","GB",
             "IRL","IS","I","J","CDN","ROK","LV","FL","L",
             "MEX","NZ","N","A","PL","P","RUS","S","CH",
             "E","CZ","H","USA")
```

Dann erzeugen wir einen Vektor `bereiche` mit den drei Bereichen:

```
> bereiche<-c("Lesekompetenz","Mathematik",
              "Naturwissenschaft")
```

Wir weisen diese beiden Vektoren einer Liste mit Namen `namen.PISA` zu:

```
> namen.PISA<-list(laender,bereiche)
```

Schauen wir uns `namen.PISA` an:

```
> namen.PISA
[[1]]:
 [1] "AUS" "B"   "BR"  "DK"  "D"   "FIN" "F"   "GR"
     "GB"  "IRL" "IS"  "I"   "J"   "CDN" "ROK" "LV"
     "FL"  "L"   "MEX" "NZ"  "N"   "A"   "PL"
     "P"   "RUS" "S"   "CH"  "E"   "CZ"  "H"   "USA" [[2]]:
  [1] "Lesekompetenz" "Mathematik" "Naturwissenschaft"
```

Auf Komponenten einer Liste greift man mit doppelten eckigen Klammern zu:

```
> namen.PISA[[2]]
[1] "Lesekompetenz" "Mathematik" "Naturwissenschaft"
```

Nun geben wir den Zeilen und Spalten von `pisa` Namen:

```
> dimnames(PISA)<-namen.PISA
```

Einzelne Elemente der Matrix erhält man durch Indizierung, wobei man die Nummer der Zeile und die Nummer der Spalte in eckigen Klammern durch Komma getrennt eingeben muss. Die Punktezahl von Deutschland im Bereich Mathematik erhält man also durch

```
> PISA[5,2]
[1] 490
```

Die Punkte von Deutschland in allen Bereichen erhält man durch

```
> PISA[5,]
 Lesekompetenz Mathematik Naturwissenschaft
           484        490               487
```

Wendet man die Funktion `mean` auf eine Matrix an, so wird der Mittelwert aller Elemente dieser Matrix bestimmt:

```
> mean(PISA)
[1] 493.0753
```

Dieser interessiert aber in der Regel wenig, da man die einzelnen Variablen getrennt analysieren will. Will man die Mittelwerte aller Spalten einer Matrix bestimmen, so muss man die Funktion `apply` aufrufen. Der allgemeine Aufruf von `apply` ist

```
apply(X, MARGIN, FUN)
```

Dabei sind `X` die Matrix und `MARGIN` die Dimension der Matrix, bezüglich der die Funktion angewendet werden soll. Dabei steht 1 für die Zeilen und 2 für die Spalten. Das Argument `FUN` ist der Name der Funktion, die auf `MARGIN` von `X` angewendet werden soll. Der Aufruf `apply(PISA,1,mean)` bestimmt den Vektor der Mittelwerte der Zeilen der Datenmatrix `PISA` und der Aufruf `apply(PISA,2,mean)` bestimmt den Vektor der Mittelwerte der Spalten der Datenmatrix `PISA`. So sind die mittleren Punktezahlen in den Bereichen:

```
> apply(PISA,2,mean)
 Lesekompetenz Mathematik Naturwissenschaft
      493.4516   493.1613          492.6129
```

Die zentrierte Datenmatrix kann man auf drei Arten erhalten. Man kann die Funktion `scale` anwenden, die neben der Datenmatrix `m` noch die beiden Argumente `center` und `scale` besitzt. Diese sind standardmäßig auf `T` gesetzt. Ruft man die Funktion `scale` nur mit der Datenmatrix als Argument auf, so liefert diese die Matrix der standardisierten Variablen. Von jedem Wert jeder Variablen wird der Mittelwert subtrahiert und anschließend durch die Standardabweichung der Variablen dividiert. Setzt man das Argument `scale` auf `F`, so erhält man die Matrix der zentrierten Variablen. Der Aufruf

```
> scale(PISA,scale=F)
```

liefert also die zentrierte Datenmatrix. Man kann aber auch die Funktion `sweep` aufrufen. Der Aufruf von `sweep` für eine Matrix ist

```
sweep(M, MARGIN, STATS, FUN)
```

Dabei sind `M` die Matrix und `MARGIN` die Dimension der Matrix, bezüglich der die Funktion angewendet werden soll. Dabei steht 1 für die Zeilen und 2 für die Spalten. Das Argument `STATS` ist ein Vektor, dessen Länge der Größe der Dimension entspricht, die im Argument `MARGIN` gewählt wurde, und das

Argument `FUN` ist der Name der Funktion, die auf `MARGIN` von `M` angewendet werden soll. Standardmäßig wird die Subtraktion gewählt. Die Funktion `sweep` bewirkt, dass die Funktion `FUN` angewendet wird, um die Komponenten des Vektors aus der gewählten Dimension von `M` im wahrsten Sinne des Wortes herauszufegen. Stehen zum Beispiel in `STATS` die Mittelwerte der Spalten von `M`, und ist `FUN` gleich `"-"`, so liefert der Aufruf

```
> sweep(M,2,STATS,FUN="-")
```

die zentrierte Datenmatrix. Die Komponenten von `STATS` können wir mit Hilfe von `apply` bestimmen, sodass der folgende Aufruf für das Beispiel die Matrix der zentrierten Variablen liefert:

```
> sweep(PISA,2,apply(PISA,2,mean),FUN="-")
```

Man kann die zentrierte Datenmatrix aber auch mit der Gleichung (2.10) auf Seite 26 gewinnen. Die Matrix $\mathbf{M}$ liefert folgender Ausdruck:

```
> n<-dim(PISA)[1]
> M<-diag(n)-outer(rep(1,n),rep(1,n))/n
```

Die Funktion `outer` wird auf Seite 490 beschrieben. Die zentrierte Datenmatrix erhalten wir durch

```
> M%*%PISA
```

Um die Stichprobenvarianzen der drei Variablen zu bestimmen, benutzen wir wiederum die Funktion `apply`:

```
> apply(PISA,2,var)
 Lesekompetenz Mathematik Naturwissenschaft
      1109.389   2192.873          1418.978
```

Um ein Streudiagramm zu erstellen, verwendet man in `S-PLUS` die Funktion `plot`. Übergibt man dieser als Argumente zwei gleich lange Vektoren, so erstellt sie ein Streudiagramm, wobei die Komponenten des ersten Vektors der Abszisse und die des zweiten Vektors der Ordinate zugeordnet werden. Um das Streudiagramm der Merkmale `Lesekompetenz` und `Mathematische Grundbildung` zu erstellen, geben wir also ein

```
> plot(PISA[,1],PISA[,2])
```

Wir erhalten die Abbildung 2.12. Die Graphik kann man nun noch verbessern. Die Achsen können noch geeignet beschriftet werden durch die Argumente `xlab` und `ylab`. Die Beschriftung wird der Funktion `plot` als Argument in Form einer Zeichenkette übergeben. Die folgende Befehlsfolge erzeugt die Abbildung 2.4:

```
> plot(PISA[,1],PISA[,2],xlab="Lesekompetenz",
       ylab="Mathematische Grundbildung")
```

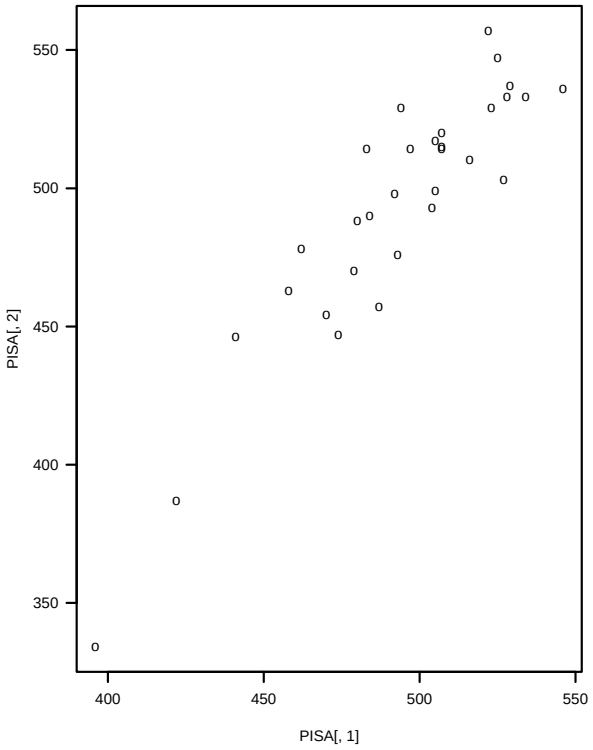


Fig. 2.12. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung im Rahmen der PISA-Studie

In Abbildung 2.9 haben wir die Punkte im Streudiagramm mit Kürzeln der Ländernamen versehen. Um dies zu erreichen, weisen wir beim Aufruf der Funktion `plot` dem Argument `type` den Wert `"n"` zu. In diesem Fall werden keine Punkte gezeichnet:

```
> plot(PISA[,1],PISA[,2],xlab="Lesekompetenz",
       ylab="Mathematische Grundbildung",type="n")
```

Nun müssen wir nur noch mit der Funktion `text` die Namen an den entsprechenden Stellen hinzufügen:

```
> text(PISA[,1],PISA[,2],laender)
```

Wendet man die Funktion `var` auf eine Datenmatrix an, so erhält man die empirische Varianz-Kovarianz-Matrix:

```
> var(PISA)
```

```
                 Lesekompetenz Mathematik Naturwissenschaft
      Lesekompetenz      1109.389   1428.325          1195.614
         Mathematik      1428.325   2192.873          1644.031
   Naturwissenschaft      1195.614   1644.031          1418.978
```

Um die Struktur besser erkennen zu können, runden wir mit der Funktion `round` auf eine Stelle nach dem Komma:

```
> round(var(PISA),1)
                 Lesekompetenz Mathematik Naturwissenschaft
    Lesekompetenz        1109.4     1428.3            1195.6
       Mathematik        1428.3     2192.9            1644.0
Naturwissenschaft        1195.6     1644.0            1419.0
```

In S-PLUS bestimmen wir die empirische Korrelationsmatrix mit der Funktion `cor`:

```
> cor(PISA)
                 Lesekompetenz Mathematik Naturwissenschaft
    Lesekompetenz     1.0000000  0.9157527         0.9529302
       Mathematik     0.9157527  1.0000000         0.9319989
Naturwissenschaft     0.9529302  0.9319989         1.0000000
```

Eine Streudiagrammmatrix liefert die Funktion `pairs`. Um die Abbildung 2.8 zu erhalten, geben wir ein

```
> pairs(PISA)
```

Nun fehlt uns aus dem Bereich der quantitativen Merkmale noch die konvexe Hülle. Die Indizes der Länder auf der konvexen Hülle aller Beobachtungen erhält man mit der Funktion `chull`. Der allgemeine Aufruf von `chull` ist:

```
chull(x, y, peel=F, maxpeel=<<see below>>, onbdy=peel,
      tol=.0001)
```

Die drei letzten Argumente sind für uns im Folgenden nicht wichtig. Schauen wir uns die anderen an. Das Argument `x` ist ein Vektor mit den ersten Koordinaten der Punkte und das Argument `y` ein Vektor mit den zweiten Koordinaten der Punkte. Das Argument `peel` ist eine logische Variable, über die gesteuert wird, ob eine Folge konvexer Hüllen erzeugt werden soll. Wenn das Argument `peel` gleich `F` ist, erhält man als Ergebnis einen Vektor mit den Indizes der Punkte auf der konvexen Hülle. Um diesen für das Beispiel zu erhalten, geben wir also ein

```
> chull(PISA[,1],PISA[,2])
```

und erhalten das Ergebnis

```
[1] 10  3 18 17 27 13  6
```

Der folgende Befehl liefert den 0.23-getrimmten Mittelwert:

```
> apply(PISA[-chull(PISA[,1],PISA[,2]),1:2],2,mean)
 Lesekompetenz Mathematik
      495.3333   494.5417
```

Um Abbildung 2.5 auf Seite 30 zu erhalten, benötigen wir die Funktion `polygon`. Der Aufruf

```
> polygon(x,y)
```

überlagert eine Graphik mit einem Polygon mit den Eckpunkten `(x,y)`. Wir geben also ein

```
> plot(PISA[,1],PISA[,2],xlab="Lesekompetenz",
       ylab="Mathematische Grundbildung")
> hull <- chull(PISA[,1],PISA[,2])
> polygon(PISA[hull,1],PISA[hull,2],density=0)
```

und erhalten Abbildung 2.5 auf Seite 30. Um den auf der konvexen Hülle beruhenden Median zu erhalten, setzen wir das Argument `peel` der Funktion `chull` auf `T`:

```
> p <- chull(PISA[,1],PISA[,2],peel=T)
```

und erhalten folgendes Ergebnis:

```
> p
 $depth:
 [1] 3 4 1 3 5 1 5 3 4 1 5 2 1 2 2 3
     1 1 2 3 5 5 5 4 2 3 1 4 6 4 4
 $hull:
 [1] 10 3 18 17 27 13 6 19 25 15 14 12 8 16 4 20
     1 26 28 24 31 2 9 30 23 5 7 22 11 21 29
 $count:
 [1] 7 5 6 6 6 1
```

Das Ergebnis ist eine Liste. Die erste Komponente gibt für jeden Punkt die Nummer der konvexen Hülle an, auf der er liegt. Dabei werden die Hüllen von außen nach innen nummeriert. Die zweite Komponente gibt die Indizes der Punkte auf den einzelnen Hüllen an. Die dritte Komponente gibt die Anzahl der Punkte auf jeder Hülle an. Um den Median zu bestimmen, benötigen wir nur die erste Komponente. Wir bestimmen die Punkte, die auf der Hülle mit der höchsten Nummer liegen:

```
> m<-PISA[p[[1]]==max(p[[1]]),1:2]
> m
 Lesekompetenz Mathematik
           492         498
```

Da es sich um einen Punkt handelt, haben wir den Median bereits gefunden. Bei mehr als einem Punkt bestimmen wir den Mittelwert dieser Punkte mit der Funktion `apply`.

Bisher haben wir Vektoren, Listen und Matrizen betrachtet. Von diesen bieten Listen die Möglichkeit, Variablen unterschiedlichen Typs in einem Objekt zu speichern. Die Elemente einer Matrix müssen vom gleichen Typ sein. In `S-PLUS` ist es aber auch möglich, Variablen unterschiedlichen Typs in einem Objekt zu speichern, auf das wie auf eine Matrix zugegriffen werden kann. Diese heißen *Dataframes*. Schauen wir uns exemplarisch die ersten 10 Beobachtungen der Daten in Tabelle 1.2 auf Seite 5 an. Wir erzeugen

zunächst die 5 Variablen. Da viele Werte mehrfach hintereinander vorkommen, verwenden wir die Funktion `rep`:

```
> Geschlecht<-c(rep("m",5),rep("w",4),"m")
> Geschlecht<-factor(Geschlecht)
> MatheLK<-c(rep("n",9),"j")
> MatheLK<-factor(MatheLK)
> MatheNote<-c(3,4,4,4,3,3,4,3,4,3)
> MatheNote<-ordered(MatheNote)
> Abitur88<-c(rep("n",6),rep("j",3),"n")
> Abitur88<-factor(Abitur88)
> Punkte<-c(8,7,4,2,7,6,3,7,14,19)
```

Mit der Funktion `data.frame` macht man aus diesen Variablen einen Dataframe:

```
> test<-data.frame(Geschlecht,MatheLK,MatheNote,
                   Abitur88,Punkte)
   Geschlecht MatheLK MatheNote Abitur88 Punkte
1           m       n         3        n      8
2           m       n         4        n      7
3           m       n         4        n      4
4           m       n         4        n      2
5           m       n         3        n      7
6           w       n         3        n      6
7           w       n         4        j      3
8           w       n         3        j      7
9           w       n         4        j     14
10          m       j         3        n     19
```

Die Werte der Variablen `Geschlecht` erhalten wir durch

```
> test[,1]
 [1] m m m m m w w w w m
```

oder durch

```
> test[[1]]
 [1] m m m m m w w w w m
```

Schauen wir uns noch die Beschreibung qualitativer Merkmale an. Wir betrachten wiederum die Merkmale `Geschlecht`, `MatheLK` und `Abitur88` des Beispiels 2 bei den ersten 10 Studenten. Wir bilden eine Matrix `qual` mit den Merkmalen

```
> qual<-test[,c(1,2,4)]
```

Schauen wir uns `qual` an:

```
> qual
   Geschlecht MatheLK Abitur88
 1          m       n        n
 2          m       n        n
 3          m       n        n
 4          m       n        n
 5          m       n        n
 6          w       n        n
 7          w       n        j
 8          w       n        j
 9          w       n        j
10          m       j        n
```

Mit Hilfe der Funktion `table` erstellen wir die zweidimensionale Kontingenztabelle der Merkmale `Geschlecht` und `MatheLK`:

```
> table(qual[,1],qual[,2])
   j n
 m 1 5
 w 0 4
```

In den Zeilen stehen die Ausprägungen des Merkmals `Geschlecht` und in den Spalten die Ausprägungen des Merkmals `MatheLK`. Die Matrix der bedingten relativen Häufigkeiten bestimmen wir mit der Funktion `sweep`. Hierzu weisen wir den obigen Aufruf einer Variablen zu:

```
> e<-table(qual[,1],qual[,2])
```

und rufen dann `sweep` auf:

```
> sweep(e,1,apply(e,1,sum),"/")
           j         n
 m 0.1666667 0.8333333
 w 0.0000000 1.0000000
```

Eine dreidimensionale Kontingenztabelle liefert der Aufruf von `table` mit drei Argumenten. Die dreidimensionale Tabelle der Merkmale `Geschlecht`, `MatheLK` und `Abitur88` erhalten wir durch

```
> e<-table(qual[,1],qual[,2],qual[,3])
> e
 , , j
   j n
 m 0 0
 w 0 3
 , , n
   j n
```

```
m 1 5
w 0 1
```

Die zweidimensionale Tabelle der Merkmale `Geschlecht` und `MatheLK` erhalten wir durch Anwenden von `apply` auf `e` mit der Funktion `sum`:

```
> apply(e,c(1,2),sum)
  j n
m 1 5
w 0 4
```

Entsprechend erhält man die beiden anderen zweidimensionalen Tabellen. Oft liegen die Daten in Form einer dreidimensionalen Kontingenztabelle vor. Dies ist im Beispiel 9 der Fall. Man kann diese in `S-PLUS` mit der Funktion `array` eingeben. Diese wird folgendermaßen aufgerufen:

```
array(data = NA, dim, dimnames = NULL)
```

Dabei ist `data` der Vektor mit den Daten, `dim` ein Vektor mit den Dimensionsangaben und `dimnames` eine Liste mit Namen der Dimensionen. In welcher Reihenfolge wird die dreidimensionale Tabelle nun aufgefüllt? Stellen wir uns die Tabelle als Schichten von Matrizen vor, so wird zuerst die Matrix der ersten Schicht spaltenweise aufgefüllt. Dann wird die Matrix jeder weiteren Schicht spaltenweise aufgefüllt. Wir geben also ein

```
> wahl<-array(c(4,2,12,2,46,4,24,6),c(2,2,2),
                  dimnames=list(c("BWL","VWL"),
                  c("CDU","SPD"),c("w","m")))
```

Schauen wir uns `wahl` an:

```
> wahl
 , , w
    CDU SPD
BWL   4  12
VWL   2   2
 , , m
    CDU SPD
BWL  46  24
VWL   4   6
```

2.4 Ergänzungen und weiterführende Literatur

In diesem Kapitel haben wir Verfahren kennengelernt, mit denen man die wesentlichen Charakteristika eines Datensatzes mit mehreren Merkmalen beschreiben kann. Hierbei haben wir einige Aspekte nicht berücksichtigt. Das Histogramm ist ein spezieller Dichteschätzer, der aber nicht glatt ist. Mit *Kerndichteschätzern* erhält man eine glatte Schätzung der Dichtefunktion. Univariate und multivariate Dichteschätzer werden bei Härdle (1990b) beschrieben. Dort sind auch Funktionen in `S` zur Dichteschätzung zu finden. Rousseeuw et al. (1999) entwickelten einen zweidimensionalen Boxplot, den sie *Bagplot* nennen. Von Rousseeuw (1984) wurden zwei robuste Schätzer der Varianz-Kovarianz-Matrix vorgeschlagen. Beim *MVE-Schätzer* wird das Ellipsoid mit kleinstem Volumen bestimmt, das h der n Beobachtungen enthält, während man beim *MCD-Schätzer* die h Beobachtungen sucht, deren empirische Varianz-Kovarianz-Matrix die kleinste Determinante besitzt. Bei beiden Schätzern wird die Varianz-Kovarianz-Matrix durch die empirische Varianz-Kovarianz-Matrix der h Beobachtungen geschätzt. Ein Algorithmus zur schnellen Bestimmung des MCD-Schätzers und einige Anwendungen sind bei Rousseeuw & van Driessen (1999) zu finden. Einen weiteren wichtigen Aspekt multivariater Datensätze haben wir nicht berücksichtigt. In der Regel enthalten multivariate Datensätze eine Vielzahl fehlender Beobachtungen. Es gibt eine Reihe von Verfahren zur Behandlung fehlender Beobachtungen. Man kann zum Beispiel alle Objekte aus dem Datensatz entfernen, bei denen mindestens eine Beobachtung fehlt. Eine solche Vorgehensweise führt meistens zu einer drastischen Verringerung des Datenbestandes und ist deshalb nicht sinnvoll. Man wird eher versuchen, die fehlenden Beobachtungen zu ersetzen. Wie man hierbei vorgehen sollte, kann man bei Bankhofer (1995) finden. Bei Schafer (1997) ist ein bayesianischer Zugang zur Behandlung fehlender Beobachtungen in multivariaten Datensätzen zu finden.

2.5 Übungen

Exercise 1. Im Rahmen der PISA-Studie wurde das Merkmal `Lesekompetenz` näher untersucht. Dabei wurden die mittleren Punktezahlen der Schüler in den Bereichen `Ermitteln von Informationen`, `Textbezogenes Interpretieren` und `Reflektieren und Bewerten` in jedem der 31 Ländern bestimmt. In Tabelle 2.12 sind die Ergebnisse zu finden.

Table 2.12. Mittelwert der Punkte in den Bereichen der Lesekompetenz im Rahmen der PISA-Studie, vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S.533

Land	Ermitteln von Informationen	Textbezogenes Interpretieren	Reflektieren und Bewerten
Australien	536	527	526
Belgien	515	512	497
Brasilien	365	400	417
Dänemark	498	494	500
Deutschland	483	488	478
Finnland	556	555	533
Frankreich	515	506	496
Griechenland	450	475	495
Grossbritannien	523	514	539
Irland	524	526	533
Island	500	514	501
Italien	488	489	483
Japan	526	518	530
Kanada	530	532	542
Korea	530	525	526
Lettland	451	459	458
Liechtenstein	492	484	468
Luxemburg	433	446	442
Mexiko	402	419	446
Neuseeland	535	526	529
Norwegen	505	505	506
Österreich	502	508	512
Polen	475	482	477
Portugal	455	473	480
Russland	451	468	455
Schweden	516	522	510
Schweiz	498	496	488
Spanien	483	491	506
Tschechien	481	500	485
Ungarn	478	480	481
USA	499	505	507

Benutzen Sie bei der Lösung der folgenden Aufgaben bitte `S-PLUS`.

1. Bestimmen Sie den Mittelwertvektor und die Matrix der zentrierten Merkmale.
2. Bestimmen Sie die Matrix der standardisierten Merkmale.
3. Bestimmen Sie die empirische Varianz-Kovarianz-Matrix und die empirische Korrelationsmatrix der Daten.
4. Erstellen und interpretieren Sie die Streudiagrammmatrix.

Exercise 2. Betrachten Sie die Daten in Tabelle 1.1 auf Seite 4.

1. Bestimmen Sie den Mittelwertvektor und die Matrix der zentrierten Merkmale mit `S-PLUS`.
2. Zeigen Sie, dass der Mittelwert der zentrierten Merkmale gleich 0 ist.
3. Prüfen Sie mit `S-PLUS` am Datensatz, dass der Mittelwert der zentrierten Merkmale gleich 0 ist.
4. Bestimmen Sie die Matrix der standardisierten Merkmale mit `S-PLUS`.
5. Zeigen Sie, dass die Stichprobenvarianz der standardisierten Merkmale gleich 1 ist.
6. Prüfen Sie mit `S-PLUS` am Datensatz, dass die Stichprobenvarianz der standardisierten Merkmale gleich 1 ist.
7. Bestimmen Sie die empirische Varianz-Kovarianz-Matrix und die empirische Korrelationsmatrix der Daten mit `S-PLUS`.
8. Bestimmen Sie mit `S-PLUS` die empirische Varianz-Kovarianz-Matrix der Daten mit Hilfe von Gleichung (2.17).
9. Bestimmen Sie mit `S-PLUS` die empirische Korrelationsmatrix der Daten mit Hilfe von Gleichung (2.20).

Exercise 3. Im Rahmen einer Weiterbildungsveranstaltung sollten die Teilnehmer einen Fragebogen ausfüllen. Neben dem Merkmal `Geschlecht` mit den Ausprägungsmöglichkeiten `w` und `m` wurde noch eine Reihe weiterer Merkmale erhoben. Die Teilnehmer wurden gefragt, ob sie den Film Titanic gesehen haben. Dieses Merkmal bezeichnen wir mit `Titanic`. Außerdem sollten Sie den folgenden Satz fortsetzen:

> Zu Risiken und Nebenwirkungen ...

Wir bezeichnen das Merkmal mit `Satz`. Es nimmt die Ausprägung `j`, wenn der Satz richtig fortgesetzt wurde. Ansonsten nimmt es den Wert `n` an. Die Ergebnisse sind in Tabelle 2.13 zu finden.

1. Erstellen Sie die dreidimensionale Kontingenztabelle.
2. Erstellen Sie die Kontingenztabelle der Merkmale `Geschlecht` und `Titanic` und bestimmen Sie die bedingten relativen Häufigkeiten.
3. Erstellen Sie die Kontingenztabelle der Merkmale `Geschlecht` und `Satz` und bestimmen Sie die bedingten relativen Häufigkeiten.
4. Erstellen Sie die Kontingenztabelle der Merkmale `Satz` und `Titanic` und bestimmen Sie die bedingten relativen Häufigkeiten.
5. Auf welche Zusammenhänge deuten die drei zweidimensionalen Kontingenztabellen hin?

Table 2.13. Ergebnisse einer Befragung in einer Weiterbildungsveranstaltung

Person	Geschlecht	Titanic	Satz	Person	Geschlecht	Titanic	Satz
1	m	n	n	14	w	j	j
2	w	j	n	15	w	j	n
3	w	j	j	16	m	j	n
4	m	n	n	17	m	n	n
5	m	n	n	18	m	j	n
6	m	j	j	19	w	n	n
7	w	j	n	20	w	j	n
8	m	n	n	21	w	j	j
9	w	j	j	22	w	j	j
10	m	n	n	23	w	j	n
11	w	j	j	24	w	j	j
12	m	j	n	25	m	n	j
13	m	j	j				

3 Mehrdimensionale Zufallsvariablen

3.1 Problemstellung

Im letzten Kapitel haben wir einfache Verfahren zur Darstellung hochdimensionaler Datensätze kennengelernt. Bei diesen Datensätzen handelt es sich in der Regel um Stichproben aus Populationen. Um Schlüsse über die zugrunde liegenden Populationen ziehen zu können, muss man Annahmen über die Merkmale machen. Hierzu benötigen wir das Konzept der *Zufallsvariablen*. Wir werden in diesem Kapitel zunächst *univariate* Zufallsvariablen betrachten. Anschließend werden wir die wesentlichen Eigenschaften von *mehrdimensionalen* Zufallsvariablen herleiten, die wir im weiteren Verlauf des Buches immer wieder benötigen werden.

3.2 Univariate Zufallsvariablen

Example 17. Im Beispiel 2 auf Seite 3 wurden die 20 Studenten unter anderem danach befragt, ob sie den Leistungskurs Mathematik besucht haben. Kodiert man `j` mit 1 und `n` mit 0, so erhält man folgende Daten:

```
0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 .
```

□

Im letzten Kapitel haben wir gesehen, wie wir diesen Datensatz beschreiben können. Hier wollen wir die Daten als Zufallsstichprobe aus einer Grundgesamtheit auffassen, über die man Aussagen treffen will. In der Regel will man Parameter schätzen oder Hypothesen testen. Um zu sinnvollen Schlussfolgerungen zu gelangen, muss man für das Merkmal eine Wahrscheinlichkeitsverteilung unterstellen. Das Merkmal wird dann zu einer Zufallsvariablen. Man unterscheidet *diskrete* und *stetige* Zufallsvariablen.

Definition 1. *Eine Zufallsvariable Y, die höchstens abzählbar viele Werte annehmen kann, heißt diskret. Dabei heißt $P(Y = y)$ die Wahrscheinlichkeitsfunktion von Y.*

Für die Wahrscheinlichkeitsfunktion einer diskreten Zufallsvariablen Y gilt

$$\sum_y P(Y = y) = 1$$

und

$$P(Y = y) \geq 0$$

für alle $y \in \mathbb{R}$. hmcounterend. (fortgesetzt)

Example 17. Es liegt nahe, die Zufallsvariable Y zu betrachten, die wir Leistungskurs nennen wollen. Sie kann die Werte 0 und 1 annehmen. Die Wahrscheinlichkeit des Wertes 1 sei p. Somit ist die Wahrscheinlichkeit des Wertes 0 gleich $1 - p$. Wir erhalten somit die Wahrscheinlichkeitsfunktion

$$P(Y = 0) = 1 - p\,,$$
$$P(Y = 1) = p\,.$$

□

Definition 2. *Die Zufallsvariable Y heißt bernoulliverteilt mit dem Parameter p, wenn gilt*

$$P(Y = 0) = 1 - p\,,$$
$$P(Y = 1) = p\,.$$

Man kann die Wahrscheinlichkeitsverteilung der *Bernoulli-Verteilung* auch kompakter schreiben. Es gilt

$$P(Y = y) = p^y\,(1 - p)^{1-y} \quad \text{für} \quad y = 0, 1\,.$$

Oft ist die Frage von Interesse, ob eine Zufallsvariable Y Werte annimmt, die kleiner oder gleich einem vorgegebenen Wert y sind.

Definition 3. *Sei Y eine Zufallsvariable. Dann heißt*

$$F_Y(y) = P(Y \leq y)$$

die Verteilungsfunktion von Y.

Schauen wir uns stetige Zufallsvariablen an.

Definition 4. *Eine Zufallsvariable Y heißt stetig, wenn eine Funktion $f_Y : \mathbb{R} \to \mathbb{R}$ existiert, sodass für die Verteilungsfunktion $F_Y(y)$ von Y gilt*

$$F_Y(y) = \int_{-\infty}^{y} f_Y(u)\,du\,.$$

Die Funktion $f_Y(y)$ heißt *Dichtefunktion* der Zufallsvariablen Y.

Die Dichtefunktion $f_Y(y)$ erfüllt folgende Bedingungen:

1.
$$f_Y(y) \geq 0 \qquad \text{für alle } y \in I\!R,$$

2.
$$\int_{-\infty}^{\infty} f_Y(y)\, dy = 1\,.$$

Man kann zeigen, dass jede Funktion, die diese Bedingungen erfüllt, als Dichtefunktion einer stetigen Zufallsvariablen aufgefasst werden kann.

Das wichtigste Verteilungsmodell für eine stetige Zufallsvariable ist die *Normalverteilung* .

Definition 5. *Die Zufallsvariable Y heißt normalverteilt mit den Parametern μ und σ^2, wenn ihre Dichtefunktion gegeben ist durch*

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\} \quad \textit{für} \quad y \in I\!R. \tag{3.1}$$

Wir schreiben $Y \sim N(\mu, \sigma^2)$.

Abbildung 3.1 zeigt die Dichtefunktion der Normalverteilung mit $\mu = 0$ und $\sigma = 1$, die *Standardnormalverteilung* heißt.
Charakteristika von Verteilungen werden durch *Maßzahlen* beschrieben. Die beiden wichtigsten sind der *Erwartungswert* und die *Varianz*. Der Erwartungswert ist die wichtigste Maßzahl für die Lage einer Verteilung.

Definition 6. *Sei Y eine Zufallsvariable mit Wahrscheinlichkeitsfunktion $P(Y = y)$ bzw. Dichtefunktion $f_Y(y)$. Der Erwartungswert $E(Y)$ von Y ist definiert durch*

$$E(Y) = \sum_y y\, P(Y = y),$$

falls Y diskret ist, und durch

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy,$$

falls Y stetig ist.

Schauen wir uns die Bernoulli-Verteilung und die Normalverteilung an. Für eine mit Parameter p bernoulliverteilte Zufallsvariable Y gilt

$$E(Y) = p. \tag{3.2}$$

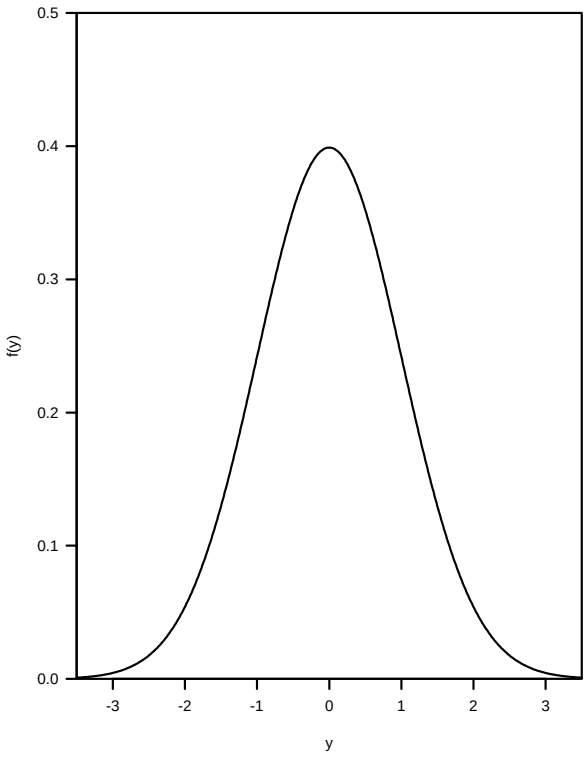


Fig. 3.1. Dichtefunktion der Standardnormalverteilung

Dies sieht man folgendermaßen:

$$E(Y) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Für eine mit den Parametern μ und σ^2 normalverteilte Zufallsvariable Y gilt

$$E(Y) = \mu. \tag{3.3}$$

Der Beweis ist zu finden bei Mood et al. (1974), S. 109.

Der Erwartungswert $E(g(Y))$ einer Funktion $g(Y)$ der Zufallsvariablen Y ist definiert durch

$$E(g(Y)) = \sum_y g(y)\, P(Y = y),$$

falls Y diskret ist und durch

$$E(g(Y)) = \int_{-\infty}^{\infty} g(y)\, f_Y(y)\, dy,$$

falls Y stetig ist.

Der Erwartungswert besitzt folgende wichtige Eigenschaft:

$$E(a\,Y + b) = a\,E(Y) + b. \tag{3.4}$$

Wir zeigen dies für eine diskrete Zufallsvariable Y mit Wahrscheinlichkeitsfunktion $P(Y = y)$:

$$\begin{aligned} E(a\,Y + b) &= \sum_y (a\,y + b)\,P(Y = y) = \sum_y (a\,y\,P(Y = y) + b\,P(Y = y)) \\ &= \sum_y a\,y\,P(Y = y) + \sum_y b\,P(Y = y) \\ &= a \sum_y y\,P(Y = y) + b \sum_y P(Y = y) = a\,E(Y) + b. \end{aligned}$$

Definition 7. *Sei Y eine Zufallsvariable. Dann ist die Varianz von Y definiert durch*

$$Var(Y) = E((Y - E(Y))^2). \tag{3.5}$$

Für eine mit Parameter p bernoulliverteilte Zufallsvariable Y gilt

$$Var(Y) = p \cdot (1 - p).$$

Dies sieht man folgendermaßen:

$$\begin{aligned} Var(Y) &= (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p \\ &= p \cdot (1 - p) \cdot (p + 1 - p) \\ &= p \cdot (1 - p). \end{aligned}$$

Für eine mit den Parametern μ und σ^2 normalverteilte Zufallsvariable Y gilt

$$Var(Y) = \sigma^2.$$

Der Beweis ist bei Mood et al. (1974) auf Seite 109 zu finden. Die Parameter μ und σ^2 sind also gerade der Erwartungswert und die Varianz.

Die Varianz besitzt folgende Eigenschaft:

$$Var(a\,Y + b) = a^2\,Var(Y). \tag{3.6}$$

Dies sieht man mit Gleichung (3.4) folgendermaßen:

$$\begin{aligned} Var(a\,Y + b) &= E\left[(a\,Y + b - E(a\,Y + b))^2\right] \\ &= E\left[(a\,Y + b - a\,E(Y) - b)^2\right] = E\left[(a\,(Y - E(Y)))^2\right] \\ &= a^2\,E\left[(Y - E(Y))^2\right] = a^2\,Var(Y)\,. \end{aligned}$$

3.3 Zufallsmatrizen und Zufallsvektoren

Wir wollen nun mehrer Zufallsvariablen gleichzeitig betrachten. Beginnen wir mit einem Theorem, dessen Aussage wir im Folgenden immer wieder benötigen.

Theorem 1. *Seien $Y_1, \ldots, Y_p$ univariate Zufallsvariablen. Dann gilt*

$$E\left(\sum_{i=1}^{p} Y_i\right) = \sum_{i=1}^{p} E(Y_i). \tag{3.7}$$

Der Beweis des Theorems ist bei Rice (1988), S. 112-113 zu finden.

Man kann mehrere Zufallsvariablen zu einer *Zufallsmatrix* oder einem *Zufallsvektor* zusammenfassen.

Definition 8. *Seien $W_{11}, \ldots, W_{1n}, \ldots, W_{m1}, \ldots, W_{mn}$ univariate Zufallsvariablen. Dann heißt*

$$\mathbf{W} = \begin{pmatrix} W_{11} & \ldots & W_{1n} \\ W_{21} & \ldots & W_{2n} \\ \vdots & \ddots & \vdots \\ W_{m1} & \ldots & W_{mn} \end{pmatrix}$$

Zufallsmatrix.

Wie auch bei univariaten Zufallsvariablen ist bei Zufallsmatrizen der Erwartungswert von Interesse:

Definition 9. *Sei $\mathbf{W}$ eine Zufallsmatrix. Dann heißt*

$$E(\mathbf{W}) = \begin{pmatrix} E(W_{11}) & \ldots & E(W_{1n}) \\ E(W_{21}) & \ldots & E(W_{2n}) \\ \vdots & \ddots & \vdots \\ E(W_{m1}) & \ldots & E(W_{mn}) \end{pmatrix}$$

der Erwartungswert von $\mathbf{W}$.

Der Erwartungswert einer Zufallsmatrix besitzt eine wichtige Eigenschaft:

Theorem 2. *Seien $\mathbf{W}$ eine (m,n)-Zufallsmatrix, $\mathbf{A}$ eine (l,m)-Matrix und $\mathbf{B}$ eine (n,p)-Matrix, dann gilt*

$$E(\mathbf{AWB}) = \mathbf{A}E(\mathbf{W})\mathbf{B}. \tag{3.8}$$

Beweis:
Das Element in der i-ten Zeile und j-ten Spalte von $\mathbf{AWB}$ erhält man, indem

man das innere Produkt der i-ten Zeile von $\mathbf{AW}$ *und der j-ten Spalte der Matrix* $\mathbf{B}$ *bildet. Die i-te Zeile der Matrix* $\mathbf{AW}$ *ist*

$$\left(\sum_{r=1}^{m} a_{ir} W_{r1}, \ldots, \sum_{r=1}^{m} a_{ir} W_{rn} \right)$$

und die j-te Spalte von $\mathbf{B}$ *ist*

$$\begin{pmatrix} b_{1j} \\ \vdots \\ b_{nj} \end{pmatrix}.$$

Bildet man das innere Produkt dieser beiden Vektoren, so erhält man

$$b_{1j} \sum_{r=1}^{m} a_{ir} W_{r1} + \ldots + b_{nj} \sum_{r=1}^{m} a_{ir}\, W_{rn} = \sum_{r=1}^{m} a_{ir} W_{r1} b_{1j} + \ldots + \sum_{r=1}^{m} a_{ir} W_{rn} b_{nj}$$

$$= \sum_{s=1}^{n} \sum_{r=1}^{m} a_{ir} W_{rs} b_{sj}.$$

Nun gilt

$$E\left(\sum_{s=1}^{n} \sum_{r=1}^{m} a_{ir}\, W_{rs}\, b_{sj} \right) = \sum_{s=1}^{n} \sum_{r=1}^{m} a_{ir}\, E(W_{rs})\, b_{sj}.$$

Dies ist aber gerade das Element in der i-ten Zeile und j-ten Spalte der Matrix $\mathbf{A}E(\mathbf{W})\mathbf{B}$.

Diese Eigenschaft von Zufallsmatrizen werden wir gleich benutzen. Wir werden uns im Folgenden aber nicht mit Zufallsmatrizen, sondern mit Zufallsvektoren beschäftigen, die wir auch als *mehrdimensionale Zufallsvariablen* bezeichnen.

Definition 10. *Seien* $Y_1, \ldots, Y_p$ *univariate Zufallsvariablen. Dann heißt*

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$$

p-dimensionale Zufallsvariable.

In Analogie zum Erwartungswert einer Zufallsmatrix definieren wir den Erwartungswert einer p-dimensionalen Zufallsvariablen.

Definition 11. *Sei* $\mathbf{Y}$ *eine p-dimensionale Zufallsvariable. Dann heißt*

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ \vdots \\ E(Y_p) \end{pmatrix}$$

Erwartungswert von $\mathbf{Y}$.

Wir haben bei univariaten Zufallsvariablen Y gesehen, dass der Erwartungswert einer Lineartransformation von Y gleich der Lineartransformation des Erwartungswertes ist. Eine entsprechende Eigenschaft gilt für mehrdimensionale Zufallsvariablen.

Theorem 3. *Sei* $\mathbf{Y}$ *eine* p*-dimensionale Zufallsvariable,* $\mathbf{A}$ *eine* (m, p)*-Matrix,* $\mathbf{b}$ *ein* m*-dimensionaler Vektor. Dann gilt*

$$E(\mathbf{AY} + \mathbf{b}) = \mathbf{A}E(\mathbf{Y}) + \mathbf{b}. \tag{3.9}$$

Beweis:
Die i-te Komponente des Vektors $\mathbf{AY} + \mathbf{b}$ *ist*

$$\sum_{k=1}^{p} a_{ik}Y_k + b_i.$$

Nun gilt

$$E\left(\sum_{k=1}^{p} a_{ik}Y_k + b_i\right) = \sum_{k=1}^{p} a_{ik}E(Y_k) + b_i.$$

Dies ist aber gerade die i-te Komponente von $\mathbf{A}E(\mathbf{Y}) + \mathbf{b}$.

Wie das folgende Theorem zeigt, gilt die Aussage von Theorem 1 auch für p-dimensionale Zufallsvariablen.

Theorem 4. *Seien* $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ p*-dimensionale Zufallsvariablen mit*

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{ip} \end{pmatrix}.$$

Dann gilt

$$E\left(\sum_{i=1}^{n} \mathbf{Y}_i\right) = \sum_{i=1}^{n} E(\mathbf{Y}_i). \tag{3.10}$$

Beweis:

$$E\left(\sum_{i=1}^{n} \mathbf{Y}_i\right) = E\left[\begin{pmatrix} \sum_{i=1}^{n} Y_{i1} \\ \vdots \\ \sum_{i=1}^{n} Y_{ip} \end{pmatrix}\right] = \begin{pmatrix} E(\sum_{i=1}^{n} Y_{i1}) \\ \vdots \\ E(\sum_{i=1}^{n} Y_{ip}) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} E(Y_{i1}) \\ \vdots \\ \sum_{i=1}^{n} E(Y_{ip}) \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} E(Y_{i1}) \\ \vdots \\ E(Y_{ip}) \end{pmatrix} = \sum_{i=1}^{n} E(\mathbf{Y}_i).$$

Im Kapitel 2 haben wir die empirische Kovarianz als Maß für den linearen Zusammenhang zwischen zwei Merkmalen kennengelernt. Sie ist definiert durch

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j).$$

Diesen Ausdruck können wir problemlos auf zwei Zufallsvariablen übertragen.

Definition 12. *Seien Y und Z univariate Zufallsvariablen. Dann ist die Kovarianz $Cov(Y, Z)$ zwischen Y und Z definiert durch*

$$Cov(Y, Z) = E\left[(Y - E(Y))\,(Z - E(Z))\right]. \tag{3.11}$$

Offensichtlich gilt

$$Cov(Y, Z) = Cov(Z, Y). \tag{3.12}$$

Setzen wir in (3.11) Z gleich Y, so ergibt sich die Varianz von Y:

$$Var(Y) = Cov(Y, Y). \tag{3.13}$$

Das folgende Theorem gibt wichtige Eigenschaften der Kovarianz an.

Theorem 5. *Seien U, V, Y und Z univariate Zufallsvariablen und a, b, c und d reelle Zahlen. Dann gilt*

$$\begin{aligned} Cov(U + V, Y + Z) = {} & Cov(U, Y) + Cov(U, Z) \\ & + Cov(V, Y) + Cov(V, Z) \end{aligned} \tag{3.14}$$

und

$$Cov(a\,V + b, c\,Y + d) = a\,c\,Cov(V, Y). \tag{3.15}$$

Beweis:
Wir beweisen zunächst Gleichung (3.14):

$$\begin{aligned} & Cov(U + V, Y + Z) = \\ & E[(U + V - E(U + V))\,(Y + Z - E(Y + Z))] = \\ & E[(U + V - E(U) - E(V))(Y + Z - E(Y) - E(Z))] = \\ & E[(U - E(U) + V - E(V))\,(Y - E(Y) + Z - E(Z))] = \\ & E[(U - E(U))(Y - E(Y)) + (U - E(U))(Z - E(Z)) + \\ & (V - E(V))(Y - E(Y)) + (V - E(V))(Z - E(Z))] = \\ & E[(U - E(U))(Y - E(Y))] + E[(U - E(U))(Z - E(Z))] + \\ & E[(V - E(V))(Y - E(Y))] + E[(V - E(V))(Z - E(Z))] = \\ & Cov(U, Y) + Cov(U, Z) + Cov(V, Y) + Cov(V, Z). \end{aligned}$$

Gleichung (3.15) gilt wegen

$$\begin{aligned} Cov(a\,V + b, c\,Y + d) &= E[(a\,V + b - E(a\,V + b))\,(c\,Y + d - E(c\,Y + d))] \\ &= E[(a\,V + b - a\,E(V) - b)\,(c\,Y + d - c\,E(Y) - d)] \\ &= a\,c\,E[(V - E(V))\,(Y - E(Y))] \\ &= a\,c\,Cov(V, Y). \end{aligned}$$

Setzen wir in Gleichung (3.14) U gleich Y und V gleich Z, so gilt

$$\begin{aligned} Cov(Y + Z, Y + Z) &= Cov(Y, Y) + Cov(Y, Z) \\ &\quad + Cov(Z, Y) + Cov(Z, Z) \\ &= Var(Y) + 2\,Cov(Y, Z) + Var(Z). \end{aligned}$$

Es gilt also

$$Var(Y + Z) = Var(Y) + Var(Z) + 2\,Cov(Y, Z). \tag{3.16}$$

Entsprechend kann man zeigen

$$Var(Y - Z) = Var(Y) + Var(Z) - 2\,Cov(Y, Z). \tag{3.17}$$

Gleichung (3.15) zeigt, dass die Kovarianz nicht skaleninvariant ist. Misst man die Körpergröße in Zentimetern und das Körpergewicht in Gramm und bestimmt die Kovarianz zwischen diesen beiden Zufallsvariablen, so ist die Kovarianz 100000-mal so groß, als wenn man die Körpergröße in Metern und das Körpergewicht in Kilogramm bestimmt. Eine skaleninvariante Maßzahl für den Zusammenhang zwischen zwei Zufallsvariablen erhält man, indem man die beiden Zufallsvariablen standardisiert. Wir bilden

$$Y^* = \frac{Y - E(Y)}{\sqrt{Var(Y)}} \tag{3.18}$$

und

$$Z^* = \frac{Z - E(Z)}{\sqrt{Var(Z)}}. \tag{3.19}$$

Wie man mit (3.4) und (3.6) leicht zeigen kann, gilt

$$E(Y^*) = E(Z^*) = 0 \tag{3.20}$$

und

$$Var(Y^*) = Var(Z^*) = 1\,. \tag{3.21}$$

Für die Kovarianz zwischen Y^* und Z^* gilt

$$Cov(Y^*, Z^*) = \frac{Cov(Y, Z)}{\sqrt{Var(Y)}\,\sqrt{Var(Z)}}. \tag{3.22}$$

Dies sieht man mit Gleichung (3.15) folgendermaßen:

$$\begin{aligned} Cov(Y^*, Z^*) &= Cov\left(\frac{Y - E(Y)}{\sqrt{Var(Y)}}, \frac{Z - E(Z)}{\sqrt{Var(Z)}}\right) \\ &= \frac{1}{\sqrt{Var(Y)}\,\sqrt{Var(Z)}}\, Cov(Y - E(Y), Z - E(Z)) \\ &= \frac{Cov(Y, Z)}{\sqrt{Var(Y)}\,\sqrt{Var(Z)}}. \end{aligned}$$

Definition 13. *Seien Y und Z Zufallsvariablen. Der Korrelationskoeffizient $\rho_{Y,Z}$ zwischen Y und Z ist definiert durch*

$$\rho_{Y,Z} = \frac{Cov(Y, Z)}{\sqrt{Var(Y)}\,\sqrt{Var(Z)}}. \tag{3.23}$$

Der Korrelationskoeffizient ist skaleninvariant. Sind a und c positive reelle Zahlen, so gilt

$$\rho_{aY,cZ} = \rho_{Y,Z}\,.$$

Mit (3.6) und (3.15) sieht man dies folgendermaßen:

$$\begin{aligned} \rho_{aY,cZ} &= \frac{Cov(a\,Y, c\,Z)}{\sqrt{Var(a\,Y)}\,\sqrt{Var(c\,Z)}} = \frac{a\,c\,Cov(Y, Z)}{\sqrt{a^2\,Var(Y)}\,\sqrt{c^2\,Var(Z)}} \\ &= \frac{a\,c\,Cov(Y, Z)}{a\,c\,\sqrt{Var(Y)}\,\sqrt{Var(Z)}} = \frac{Cov(Y, Z)}{\sqrt{Var(Y)}\,\sqrt{Var(Z)}} = \rho_{Y,Z}\,. \end{aligned}$$

Das folgende Theorem gibt eine wichtige Eigenschaft des Korrelationskoeffizienten an.

Theorem 6. *Für den Korrelationskoeffizienten $\rho_{Y,Z}$ zwischen den Zufallsvariablen Y und Z gilt*

$$-1 \le \rho_{Y,Z} \le 1. \tag{3.24}$$

Dabei ist $|\rho_{Y,Z}| = 1$ genau dann, wenn Konstanten a und $b \neq 0$ existieren, sodass gilt

$$P(Z = a \pm b\,Y) = 1.$$

Beweis:
Seien Y^* und Z^* die standardisierten Variablen. Dann gilt wegen (3.16)

$$Var(Y^* + Z^*) = Var(Y^*) + Var(Z^*) + 2\,Cov(Y^*, Z^*) = 2 + 2\,\rho_{Y,Z}.$$

Da die Varianz nichtnegativ ist, gilt

$$2 + 2\,\rho_{Y,Z} \geq 0$$

und somit

$$\rho_{Y,Z} \geq -1.$$

Außerdem gilt wegen (3.17)

$$Var(Y^* - Z^*) = Var(Y^*) + Var(Z^*) - 2\,Cov(Y^*, Z^*) = 2 - 2\,\rho_{Y,Z}.$$

Hieraus folgt

$$\rho_{Y,Z} \leq 1.$$

Also gilt

$$-1 \leq \rho_{Y,Z} \leq 1.$$

Ist

$$\rho_{Y,Z} = 1,$$

so gilt

$$Var(Y^* - Z^*) = 0.$$

Somit gilt

$$P(Y^* - Z^* = 0) = 1,$$

siehe dazu Rice (1988), S. 119.
Also gilt

$$P(Y = a + b\,Z) = 1$$

mit

$$a = E(Y) - \frac{\sqrt{Var(Y)}}{\sqrt{Var(Z)}}\,E(Z)$$

und

$$b = \frac{\sqrt{Var(Y)}}{\sqrt{Var(Z)}}.$$

Eine analoge Beziehung erhält man für $\rho_{Y,Z} = -1$.

Das Konzept der Kovarianz kann auch auf mehrdimensionale Zufallsvariablen übertragen werden.

Definition 14. *Sei* $\mathbf{Y}$ *eine p-dimensionale Zufallsvariable und* $\mathbf{Z}$ *eine q-dimensionale Zufallsvariable, dann heißt*

$$Cov(\mathbf{Y},\mathbf{Z}) = \begin{pmatrix} Cov(Y_1,Z_1) & \dots & Cov(Y_1,Z_q) \\ \vdots & \ddots & \vdots \\ Cov(Y_p,Z_1) & \dots & Cov(Y_p,Z_q) \end{pmatrix} \tag{3.25}$$

die Kovarianzmatrix von $\mathbf{Y}$ *und* $\mathbf{Z}$.

Man kann die Kovarianzmatrix von $\mathbf{Y}$ und $\mathbf{Z}$ auch mit Hilfe des äußeren Produkts darstellen:

$$Cov(\mathbf{Y},\mathbf{Z}) = E\left[(\mathbf{Y}-E(\mathbf{Y}))\,(\mathbf{Z}-E(\mathbf{Z}))'\right]. \tag{3.26}$$

Dies sieht man folgendermaßen:

$$Cov(\mathbf{Y},\mathbf{Z}) =$$

$$\begin{pmatrix} Cov(Y_1,Z_1) & \dots & Cov(Y_1,Z_q) \\ \vdots & \ddots & \vdots \\ Cov(Y_p,Z_1) & \dots & Cov(Y_p,Z_q) \end{pmatrix} =$$

$$\begin{pmatrix} E[(Y_1-E(Y_1))\,(Z_1-E(Z_1))] & \dots & E[(Y_1-E(Y_1))\,(Z_q-E(Z_q))] \\ \vdots & \ddots & \vdots \\ E[(Y_p-E(Y_p))\,(Z_1-E(Z_1))] & \dots & E[(Y_p-E(Y_p))\,(Z_q-E(Z_q))] \end{pmatrix} =$$

$$E\begin{pmatrix} (Y_1-E(Y_1))\,(Z_1-E(Z_1)) & \dots & (Y_1-E(Y_1))\,(Z_q-E(Z_q)) \\ \vdots & \ddots & \vdots \\ (Y_p-E(Y_p))\,(Z_1-E(Z_1)) & \dots & (Y_p-E(Y_p))\,(Z_q-E(Z_q)) \end{pmatrix} =$$

$$E\left[\begin{pmatrix} Y_1-E(Y_1) \\ \vdots \\ Y_p-E(Y_p) \end{pmatrix} \begin{pmatrix} Z_1-E(Z_1) & \dots & Z_q-E(Z_q) \end{pmatrix}\right] =$$

$$E\left[(\mathbf{Y}-E(\mathbf{Y}))\,(\mathbf{Z}-E(\mathbf{Z}))'\right].$$

In Theorem 5 wurden Eigenschaften von $Cov(Y,Z)$ bewiesen. $Cov(\mathbf{Y},\mathbf{Z})$ besitzt analoge Eigenschaften.

Theorem 7. *Seien* $\mathbf{U}$ *und* $\mathbf{V}$ *p-dimensionale Zufallsvariablen,* $\mathbf{Y}$ *und* $\mathbf{Z}$ *q-dimensionale Zufallsvariablen,* $\mathbf{A}$ *eine* (m,p)*-Matrix,* $\mathbf{C}$ *eine* (n,q)*-Matrix,* $\mathbf{b}$ *ein* m*-dimensionaler Vektor und* $\mathbf{d}$ *ein* n*-dimensionaler Vektor. Dann gilt*

$$\begin{aligned} Cov(\mathbf{U}+\mathbf{V},\mathbf{Y}+\mathbf{Z}) &= Cov(\mathbf{U},\mathbf{Y}) + Cov(\mathbf{V},\mathbf{Y}) \\ &\quad + Cov(\mathbf{U},\mathbf{Z}) + Cov(\mathbf{V},\mathbf{Z}) \end{aligned} \tag{3.27}$$

und

$$Cov(\mathbf{A}\,\mathbf{V}+\mathbf{b},\mathbf{C}\,\mathbf{Y}+\mathbf{d}) = \mathbf{A}\,Cov(\mathbf{V},\mathbf{Y})\,\mathbf{C}'. \tag{3.28}$$

Beweis*:*
Wir zeigen zunächst Gleichung (3.27):

$$\begin{aligned} &Cov(\mathbf{U}+\mathbf{V},\mathbf{Y}+\mathbf{Z}) = \\ &E[(\mathbf{U}+\mathbf{V}-E(\mathbf{U}+\mathbf{V}))\,(\mathbf{Y}+\mathbf{Z}-E(\mathbf{Y}+\mathbf{Z}))'] = \\ &E[(\mathbf{U}+\mathbf{V}-E(\mathbf{U})-E(\mathbf{V}))(\mathbf{Y}+\mathbf{Z}-E(\mathbf{Y})-E(\mathbf{Z}))'] = \\ &E[(\mathbf{U}-E(\mathbf{U})+\mathbf{V}-E(\mathbf{V}))\,(\mathbf{Y}-E(\mathbf{Y})+\mathbf{Z}-E(\mathbf{Z}))'] = \\ &E[(\mathbf{U}-E(\mathbf{U}))(\mathbf{Y}-E(\mathbf{Y}))' + (\mathbf{U}-E(\mathbf{U}))(\mathbf{Z}-E(\mathbf{Z}))' + \\ &(\mathbf{V}-E(\mathbf{V}))(\mathbf{Y}-E(\mathbf{Y}))' + (\mathbf{V}-E(\mathbf{V}))(\mathbf{Z}-E(\mathbf{Z}))'] = \\ &E[(\mathbf{U}-E(\mathbf{U}))(\mathbf{Y}-E(\mathbf{Y}))'] + E[(\mathbf{U}-E(\mathbf{U}))(\mathbf{Z}-E(\mathbf{Z}))'] + \\ &E[(\mathbf{V}-E(\mathbf{V}))(\mathbf{Y}-E(\mathbf{Y}))'] + E[(\mathbf{V}-E(\mathbf{V}))(\mathbf{Z}-E(\mathbf{Z}))'] = \\ &Cov(\mathbf{U},\mathbf{Y}) + Cov(\mathbf{U},\mathbf{Z}) + Cov(\mathbf{V},\mathbf{Y}) + Cov(\mathbf{V},\mathbf{Z}). \end{aligned}$$

Gleichung (3.28) ist erfüllt wegen

$$\begin{aligned} &Cov(\mathbf{AV}+\mathbf{b},\mathbf{CY}+\mathbf{d}) = \\ &E[(\mathbf{AV}+\mathbf{b}-E(\mathbf{AV}+\mathbf{b}))(\mathbf{CY}+\mathbf{d}-E(\mathbf{CY}+\mathbf{d}))'] = \\ &E[(\mathbf{AV}-\mathbf{A}E(\mathbf{V}))(\mathbf{CY}-\mathbf{C}E(\mathbf{Y}))'] = \\ &E[(\mathbf{A}(\mathbf{V}-E(\mathbf{V}))(\mathbf{C}(\mathbf{Y}-E(\mathbf{Y})))'] = \\ &E[\mathbf{A}(\mathbf{V}-E(\mathbf{V}))(\mathbf{Y}-E(\mathbf{Y}))'\mathbf{C}'] = \qquad (3.29) \\ &\mathbf{A}E[(\mathbf{V}-E(\mathbf{V}))(\mathbf{Y}-E(\mathbf{Y}))']\mathbf{C}' = \qquad (3.30) \\ &\mathbf{A}\,Cov(\mathbf{V},\mathbf{Y})\,\mathbf{C}'. \end{aligned}$$

Der Übergang von (3.29) zu (3.30) gilt aufgrund von (3.8).

Definition 15. *Sei* $\mathbf{Y}$ *eine* p*-dimensionale Zufallsvariable. Dann nennt man* $Cov(\mathbf{Y},\mathbf{Y})$ *die Varianz-Kovarianz-Matrix* $Var(\mathbf{Y})$ *von* $\mathbf{Y}$*.*

Für die Varianz-Kovarianz-Matrix schreiben wir auch $\boldsymbol{\Sigma}$. Sie sieht folgendermaßen aus:

$$\boldsymbol{\Sigma} = \begin{pmatrix} Var(Y_1) & Cov(Y_1,Y_2) & \dots & Cov(Y_1,Y_p) \\ Cov(Y_2,Y_1) & Var(Y_2) & \dots & Cov(Y_2,Y_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_p,Y_1) & Cov(Y_p,Y_2) & \dots & Var(Y_p) \end{pmatrix}.$$

Wegen $Cov(Y_i,Y_j) = Cov(Y_j,Y_i)$ ist die Varianz-Kovarianz-Matrix symmetrisch. Diese Eigenschaft wird im Folgenden von zentraler Bedeutung sein.

Theorem 8. *Sei* $\mathbf{Y}$ *eine p-dimensionale Zufallsvariable,* $\mathbf{A}$ *eine* (m,p)*-Matrix und* $\mathbf{b}$ *ein m-dimensionaler Vektor. Dann gilt*

$$Var(\mathbf{AY}+\mathbf{b}) = \mathbf{A}Var(\mathbf{Y})\mathbf{A}'. \tag{3.31}$$

Beweis:
Wegen (3.28) gilt:

$$\begin{aligned} Var(\mathbf{AY}+\mathbf{b}) &= Cov(\mathbf{AY}+\mathbf{b}, \mathbf{AY}+\mathbf{b}) \\ &= \mathbf{A}\,Cov(\mathbf{Y},\mathbf{Y})\mathbf{A}' = \mathbf{A}\,Var(\mathbf{Y})\mathbf{A}'. \end{aligned}$$

Ist in (3.31) $\mathbf{A}$ eine $(1,p)$-Matrix, also ein p-dimensionaler Zeilenvektor $\mathbf{a}'$, und $\mathbf{b} = \mathbf{0}$, so gilt

$$Var(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'Var(\mathbf{Y})\,\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}. \tag{3.32}$$

Da die Varianz nichtnegativ ist, gilt für jeden p-dimensionalen Vektor $\mathbf{a}$

$$\mathbf{a}'Var(\mathbf{Y})\,\mathbf{a} \geq 0.$$

Also ist eine Varianz-Kovarianz-Matrix immer nichtnegativ definit.

Setzen wir in (3.32) für $\mathbf{a}$ den Einservektor $\mathbf{1}$ ein, erhalten wir

$$\begin{aligned} Var\left(\sum_{i=1}^{p} Y_i\right) &= Var(\mathbf{1}'\mathbf{Y}) = \mathbf{1}'Var(\mathbf{Y})\,\mathbf{1} \\ &= \sum_{i=1}^{p} Var(Y_i) + \sum_{i\neq j} Cov(Y_i,Y_j). \end{aligned}$$

Sind die Zufallsvariablen $Y_1,\dots,Y_p$ also unkorreliert, so gilt

$$Var\left(\sum_{i=1}^{p} Y_i\right) = \sum_{i=1}^{p} Var(Y_i).$$

Die Varianz-Kovarianz-Matrix der standardisierten Zufallsvariablen nennt man auch *Korrelationsmatrix* $\mathbf{P}$:

$$\mathbf{P} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix}. \tag{3.33}$$

Dabei gilt

$$\rho_{ij} = \frac{Cov(Y_i, Y_j)}{\sqrt{Var(Y_i)}\sqrt{Var(Y_j)}}.$$

Wir bezeichnen $E(Y_i)$ mit μ_i und $\sqrt{Var(Y_i)}$ mit σ_i. Sei $\tilde{\mathbf{Y}}$ die zentrierte p-dimensionale Zufallsvariable $\mathbf{Y}$. Es gilt also

$$\tilde{\mathbf{Y}} = \begin{pmatrix} Y_1 - \mu_1 \\ \vdots \\ Y_p - \mu_p \end{pmatrix}. \tag{3.34}$$

Die standardisierte p-dimensionale Zufallsvariable $\mathbf{Y}$ bezeichnen wir mit $\mathbf{Y}^*$. Es gilt

$$\mathbf{Y}^* = \begin{pmatrix} \frac{Y_1 - \mu_1}{\sigma_1} \\ \vdots \\ \frac{Y_p - \mu_p}{\sigma_p} \end{pmatrix}. \tag{3.35}$$

Bilden wir die Diagonalmatrix

$$\mathbf{D} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p \end{pmatrix},$$

so gilt

$$\mathbf{Y}^* = \mathbf{D}^{-1}\,\tilde{\mathbf{Y}}.$$

Da Korrelationen gerade die Kovarianzen zwischen den standardisierten Zufallsvariablen sind, folgt

$$\mathbf{P} = Var(\mathbf{D}^{-1}\,\tilde{\mathbf{Y}}).$$

Diese Beziehung werden wir im Kapitel 9 benötigen.

3.4 Die multivariate Normalverteilung

Die wichtigste stetige Verteilung ist die Normalverteilung. Die Dichtefunktion einer mit den Parametern μ und σ^2 normalverteilten Zufallsvariablen ist gegeben durch

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\} \quad \text{für} \quad y \in I\!R.$$

Wir können diesen Ausdruck direkt übertragen auf eine p-dimensionale Zufallsvariable $\mathbf{Y}$.

Definition 16. *Die p-dimensionale Zufallsvariable* $\mathbf{Y}$ *heißt p-variat normalverteilt mit den Parametern* $\boldsymbol{\mu}$ *und* $\boldsymbol{\Sigma}$*, falls die Dichtefunktion von* $\mathbf{Y}$ *gegeben ist durch*

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\,\pi)^{-p/2}\,|\boldsymbol{\Sigma}|^{-0.5} \exp\{-0.5\,(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\}. \qquad (3.36)$$

Das folgende Theorem gibt den Erwartungswert und die Varianz-Kovarianz-Matrix einer p-dimensionalen Normalverteilung an.

Theorem 9. *Sei* $\mathbf{Y}$ *eine mit den Parametern* $\boldsymbol{\mu}$ *und* $\boldsymbol{\Sigma}$ *p-variat normalverteilte Zufallsvariable. Dann gilt*

$$E(\mathbf{Y}) = \boldsymbol{\mu}$$

und

$$Var(\mathbf{Y}) = \boldsymbol{\Sigma}.$$

Ein Beweis dieses Theorems ist bei Seber (1977), S. 23-24 zu finden. Das folgenden Theorem benötigen wir im Kapitel 9.

Theorem 10. *Sei* $\mathbf{Y}$ *eine mit den Parametern* $\boldsymbol{\mu}$ *und* $\boldsymbol{\Sigma}$ *p-variat normalverteilte Zufallsvariable,* $\mathbf{A}$ *eine (m,p)-Matrix vom Rang m und* $\mathbf{b}$ *ein m-dimensionaler Vektor. Dann ist* $\mathbf{AY} + \mathbf{b}$ *m-variat normalverteilt mit den Parametern* $\mathbf{A}\,\boldsymbol{\mu} + \mathbf{b}$ *und* $\mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}'$

Ein Beweis dieses Theorems ist bei Seber (1977), S. 28 zu finden.

4 Ähnlichkeits- und Distanzmaße

4.1 Problemstellung

Wir gehen von n Objekten aus, an denen p Merkmale erhoben wurden und wollen bestimmen, wie ähnlich sich die Objekte sind.

Example 18. Wir betrachten das Beispiel 3 aus Kapitel 1. Im Wintersemester `1996/97` wurden 265 Studenten in der Statistik I Vorlesung befragt. Die Daten von 5 Studenten sind in Tabelle 1.3 auf Seite 6 zu finden. □

Wir suchen eine Zahl s_{ij}, die die Ähnlichkeit zwischen dem i-ten und j-ten Objekt misst. Wir nennen diese *Ähnlichkeitskoeffizient*. Diesen werden wir so wählen, dass er umso größer ist, je ähnlicher sich die beiden Objekte sind. Oft sind Ähnlichkeitskoeffizienten normiert. Sie erfüllen die Bedingung

$$0 \leq s_{ij} \leq 1.$$

Statt eines Ähnlichkeitskoeffizienten kann man auch ein *Distanzmaß* d_{ij} betrachten, das die Unähnlichkeit zwischen dem i-ten und j-ten Objekt misst. Dieses Distanzmaß sollte umso größer sein, je mehr sich zwei Objekte unterscheiden. Gilt für einen Ähnlichkeitskoeffizienten s_{ij}

$$0 \leq s_{ij} \leq 1,$$

so erhält man mit $d_{ij} = 1 - s_{ij}$ ein Distanzmaß, für das gilt

$$0 \leq d_{ij} \leq 1.$$

Wir werden hierauf noch öfter zurückkommen.

Die Distanzen zwischen allen n Objekten stellen wir in einer sogenannten *Distanzmatrix* $\mathbf{D}$ dar:

$$\mathbf{D} = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nn} \end{pmatrix}.$$

Wir betrachten nun einige Distanzmaße und Ähnlichkeitskoeffizienten. Hierbei ist es sinnvoll, die unterschiedlichen Messniveaus getrennt zu behandeln.

4.2 Bestimmung der Distanzen und Ähnlichkeiten aus der Datenmatrix

4.2.1 Quantitative Merkmale

Wir wollen die Distanz zwischen zwei Objekten, bei denen nur quantitative Merkmale erhoben wurden, durch eine Maßzahl beschreiben. hmcounterend. (fortgesetzt)

Example 18. Die Merkmale `Alter`, `Größe` und `Gewicht` in Tabelle 1.3 sind quantitativ. In Tabelle 4.1 sind die Werte dieser Merkmale bei den 5 Studenten zu finden.

Table 4.1. Alter, Größe und Gewicht von 5 Studenten

Student	Alter	Größe	Gewicht
1	23	171	60
2	21	187	75
3	20	180	65
4	20	165	55
5	23	193	81

□

Beginnen wir mit zwei Merkmalen, da wir diese graphisch leicht in einem Streudiagramm darstellen können. hmcounterend. (fortgesetzt)

Example 18. In Abbildung 4.1 stellen wir die 5 Studenten hinsichtlich der Merkmale `Alter` und `Gewicht` in einem kartesischen Koordinatensystem dar, wobei wir die Punkte durch die Nummern der Studenten markieren.

□

Es liegt nahe, als Distanz zwischen zwei Objekten den kürzesten Abstand der zugehörigen Punkte zu wählen. Seien

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix}$$

und

$$\mathbf{x}_j = \begin{pmatrix} x_{j1} \\ x_{j2} \end{pmatrix}$$

zwei Punkte aus dem $\mathbb{R}^2$. Dann ist auf Grund des Satzes von Pythagoras der kürzeste Abstand zwischen $\mathbf{x}_i$ und $\mathbf{x}_j$ gegeben durch

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}.$$

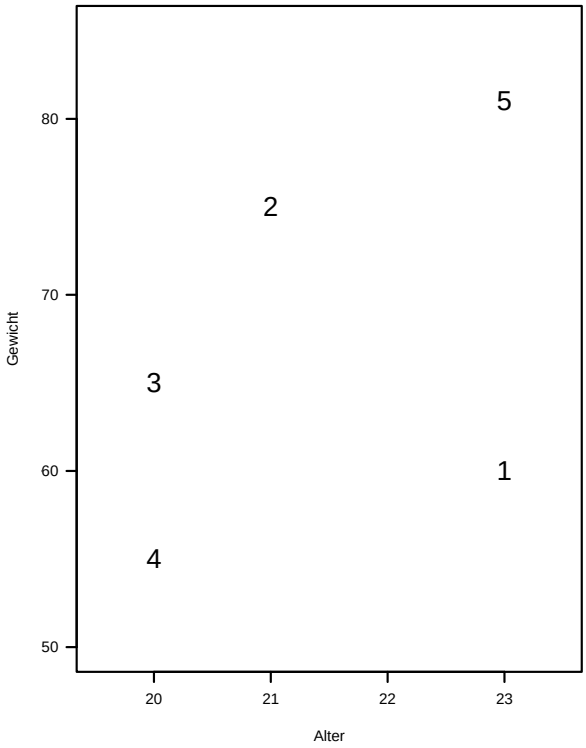


Fig. 4.1. Alter und Gewicht von 5 Studenten

Abbildung 4.2 veranschaulicht diesen Sachverhalt.
Das Konzept kann problemlos auf Punkte in höherdimensionalen Räumen übertragen werden. Die euklidische Distanz d_{ij} zwischen dem i-ten und j-ten Objekt mit den Merkmalsvektoren

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$$

und

$$\mathbf{x}_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jp} \end{pmatrix}$$

ist definiert durch

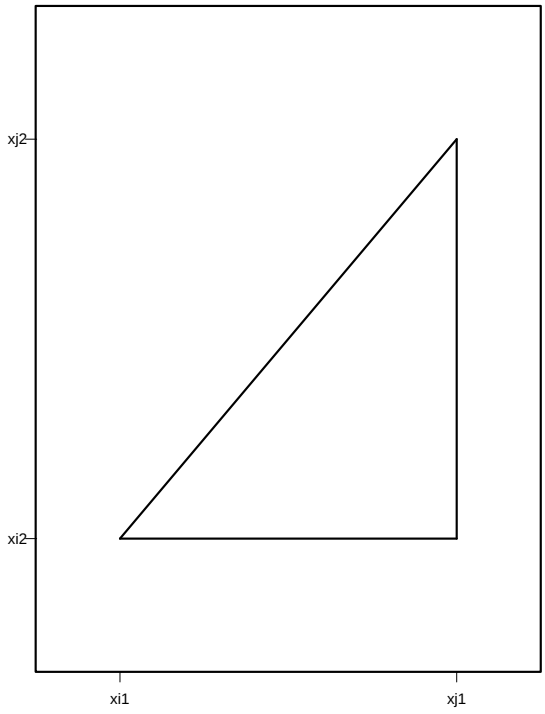


Fig. 4.2. Euklidische Distanz zwischen zwei Punkten

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}. \tag{4.1}$$

hmcounterend. (fortgesetzt)

Example 18. Wir betrachten alle drei Merkmale. Es gilt

$$d_{23} = \sqrt{(21-20)^2 + (187-180)^2 + (75-65)^2} = \sqrt{150} = 12.25.$$

Für die 5 Studenten erhalten wir folgende Distanzmatrix mit den euklidischen Distanzen der Merkmale `Alter`, `Größe` und `Gewicht`:

$$\mathbf{D} = \begin{pmatrix} 0 & 22.02 & 10.72 & 8.37 & 30.41 \\ 22.02 & 0 & 12.25 & 29.75 & 8.72 \\ 10.72 & 12.25 & 0 & 18.03 & 20.83 \\ 8.37 & 29.75 & 18.03 & 0 & 38.33 \\ 30.41 & 8.72 & 20.83 & 38.33 & 0 \end{pmatrix}.$$

□

Oft unterscheiden sich die Merkmale hinsichtlich ihrer Streuung. Dies führt dazu, dass die Distanz zwischen zwei Objekten durch die Merkmale dominiert wird, die eine große Streuung besitzen. hmcounterend. (fortgesetzt)

Example 18. Die Varianz des Merkmals `Alter` beträgt 2.3, die Varianz des Merkmals `Größe` 130.2 und die Varianz des Merkmals `Gewicht` 114.2. Wenn man zwei Studenten hinsichtlich dieser Merkmale vergleicht, so wird der Einfluss des Merkmals `Alter` gering sein. Ein Altersunterschied von einem Jahr wird bei der Berechnung des Abstandes genauso stark gewichtet wie ein Unterschied von einem Zentimeter Körpergröße. □

Dieses Problem kann man dadurch lösen, dass man die Merkmale skaliert, bevor man den Abstand bestimmt. Man dividiert den Wert x_{ij} durch die Standardabweichung s_j des j-ten Merkmals und bestimmt die Distanzen auf Basis der skalierten Merkmale. Die Distanz zwischen dem i-ten und j-ten Objekt ist somit

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{s_k^2}}.$$

hmcounterend. (fortgesetzt)

Example 18. Die Standardabweichung des Merkmals `Alter` beträgt 1.52, die Standardabweichung des Merkmals `Größe` 11.41 und die Standardabweichung des Merkmals `Gewicht` 10.69. Die Distanzmatrix der skalierten Merkmale lautet:

$$\mathbf{D} = \begin{pmatrix} 0 & 2.38 & 2.18 & 2.10 & 2.75 \\ 2.38 & 0 & 1.30 & 2.77 & 1.53 \\ 2.18 & 1.30 & 0 & 1.61 & 2.73 \\ 2.10 & 2.77 & 1.61 & 0 & 3.98 \\ 2.75 & 1.53 & 2.73 & 3.98 & 0 \end{pmatrix}.$$

Vergleichen wir diese Distanzen mit den Distanzen der unskalierten Merkmale, so sehen wir, dass sich bei den unskalierten Merkmalen die Studenten 1 und 4 am ähnlichsten sind, während bei den skalierten Merkmalen sich die Studenten 2 und 3 am meisten ähneln. □

Bei der euklidischen Distanz wird der kürzeste Abstand zwischen den beiden Punkten gewählt. Dies ist im rechtwinkligen Dreieck die Länge der Hypothenuse. Ein anderes Distanzmaß erhält man, wenn man die Summe der Längen der beiden Katheten bestimmt. Das ist dann die kürzeste Verbindung zwischen zwei Punkten, wenn man eine Stadt mit einem rechtwinkligen Straßennetz betrachtet. Da dies in Manhattan der Fall ist, spricht man von

der *Manhattan-Metrik* oder *City-Block-Metrik*. Dies führt zum City-Block-(Manhattan)-Abstand zwischen dem i-ten und j-ten Objekt mit den Merkmalsvektoren

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$$

und

$$\mathbf{x}_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jp} \end{pmatrix}.$$

Er ist definiert durch

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|. \tag{4.2}$$

hmcounterend. (fortgesetzt)

Example 18. Für die Distanz zwischen den Studenten 2 und 3 gilt

$$d_{23} = |21 - 20| + |187 - 180| + |75 - 65| = 18.$$

Die Distanzmatrix sieht folgendermaßen aus:

$$\mathbf{D} = \begin{pmatrix} 0 & 33 & 17 & 14 & 43 \\ 33 & 0 & 18 & 43 & 14 \\ 17 & 18 & 0 & 25 & 32 \\ 14 & 43 & 25 & 0 & 57 \\ 43 & 14 & 32 & 57 & 0 \end{pmatrix}.$$

□

Auch bei der Manhattan-Metrik liegt es nahe, die Merkmale zu skalieren. Man dividiert den Wert x_{ij} des Merkmals j beim i-ten Objekt durch die Spannweite R_j des j-ten Merkmals und bestimmt die Distanzen auf Basis der skalierten Merkmale. Die Distanz zwischen dem i-ten und j-ten Objekt ist somit

$$d_{ij} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{R_k}.$$

hmcounterend. (fortgesetzt)

Example 18. Die Spannweite des Merkmals `Alter` beträgt 3, die Spannweite des Merkmals `Größe` 28 und die Spannweite des Merkmals `Gewicht` 26. Wir bestimmen die Matrix der Distanzen:

$$\mathbf{D} = \begin{pmatrix} 0 & 1.82 & 1.51 & 1.41 & 1.59 \\ 1.82 & 0 & 0.97 & 1.89 & 1.11 \\ 1.51 & 0.97 & 0 & 0.92 & 2.08 \\ 1.41 & 1.89 & 0.92 & 0 & 3.00 \\ 1.59 & 1.11 & 2.08 & 3.00 & 0 \end{pmatrix}.$$

Vergleichen wir diese Distanzen mit den Distanzen der unskalierten Merkmale, so sehen wir, dass sich bei den unskalierten Merkmalen die Studenten 1 und 4 am ähnlichsten sind, während bei den skalierten Merkmalen sich die Studenten 3 und 4 am meisten ähneln. □

4.2.2 Binäre Merkmale

Binäre Merkmale haben nur zwei Ausprägungen, die wir mit 1 und 0 kodieren wollen. Dabei bedeutet 1, dass ein Objekt die durch das Merkmal beschriebene Eigenschaft besitzt. Auf den ersten Blick scheint es einfach zu sein, die Distanz beziehungsweise Ähnlichkeit zweier Objekte hinsichtlich eines binären Merkmals zu bestimmen. Betrachten wir zum Beispiel das Merkmal Geschlecht. Sind beide Personen weiblich, so ist die Distanz zwischen beiden 0, sind beide männlich, so ist die Distanz ebenfalls 0. Ist hingegen eine Person weiblich und die andere männlich, so ist die Distanz 1. Problematisch wird diese Vorgehensweise aber, wenn man am Vorhandensein eines seltenen Merkmals interessiert ist. Hier sind sich zwei Objekte nicht notwendigerweise ähnlich, wenn beide das Merkmal nicht besitzen. Man spricht in diesem Fall von *asymmetrischen* binären Merkmalen. Bei asymmetrischen binären Merkmalen sind sich zwei Objekte weder ähnlich noch unähnlich, wenn beide Objekte die interessierende Eigenschaft nicht besitzen. Bei *symmetrischen* binären Merkmalen sind sie sich auch ähnlich, wenn beide eine Eigenschaft nicht besitzen. Wir wollen uns im Folgenden mit symmetrischen und asymmetrischen binären Merkmalen beschäftigen und den Fall betrachten, dass mehrere binäre Merkmale erhoben wurden. hmcounterend. (fortgesetzt)

Example 18. Wir berücksichtigen nur die binären Merkmale aus Tabelle 1.3 auf Seite 6. Diese sind für die ersten beiden Studenten in Tabelle 4.2 zu finden.

Wir sehen, dass die beiden Studenten männlich sind und nicht rauchen. Beide haben ein eigenes Auto, aber nur der erste Student hat den Leistungskurs Mathematik besucht. □

Wir gehen zunächst davon aus, dass alle binären Merkmale symmetrisch sind. Dann können wir die Ähnlichkeit zwischen den beiden Objekten durch Zählen

Table 4.2. Merkmale Geschlecht, Raucher, Auto und MatheLK bei zwei Studenten

Student	Geschlecht	Raucher	Auto	MatheLK
1	0	0	1	1
2	0	0	1	0

bestimmen. Wir bestimmen den Anteil der Merkmale, bei denen sie übereinstimmen. Man spricht in diesem Fall vom *Simple-Matching-Koeffizienten*. Entsprechend erhalten wir ein Distanzmaß, indem wir den Anteil der Merkmale bestimmen, bei denen sie sich unterscheiden. hmcounterend. (fortgesetzt)

Example 18. Die beiden Studenten stimmen bei 3 von 4 Merkmalen überein. Also beträgt der Wert des Simple-Matching-Koeffizienten in diesem Fall 0.75. Sie unterscheiden sich bei einem von 4 Merkmalen. Also ist die Distanz gleich 0.25. □

Gehen wir davon aus, dass alle Merkmale asymmetrisch sind, so schließen wir zunächst alle Merkmale aus der weiteren Betrachtung aus, bei denen beide Objekte den Wert 0 aufweisen. Unter den restlichen Merkmalen bestimmen wir dann den Anteil, bei denen beide Objekte den gleichen Wert aufweisen. Man spricht in diesem Fall vom *Jaccard-Koeffizienten*, da dieser von Jaccard (1908) zum ersten Mal vorgeschlagen wurde. hmcounterend. (fortgesetzt)

Example 18. Die Merkmale `Geschlecht` und `Raucher` schließen wir aus, da beide Studenten hier den Wert 0 aufweisen. Bei einem der beiden anderen Merkmale stimmen sie überein, sodass der Jaccard-Koeffizient den Wert 0.5 annimmt. □

Es gibt noch eine Reihe von Koeffizienten, die auf den gleichen Ideen beruhen. Um diese konstruieren zu können, bestimmen wir für das i-te und j-te Objekt vier Größen. Die Anzahl der Merkmale, bei denen beide Objekte den Wert 1 annehmen, bezeichnen wir mit a, die Anzahl der Merkmale, bei denen Objekt i den Wert 1 und Objekt j den Wert 0 annimmt, mit b, die Anzahl der Merkmale, bei denen Objekt i den Wert 0 und Objekt j den Wert 1 annimmt, mit c und die Anzahl der Merkmale, bei denen beide den Wert 0 annehmen, mit d. Tabelle 4.3 veranschaulicht dies.

hmcounterend. (fortgesetzt)

Example 18. Es gilt:

$$a = 1, \quad b = 1, \quad c = 0, \quad d = 2.$$

Wir erhalten Tabelle 4.4.

□

Table 4.3. Hilfstabelle zur Berechnung von Distanzmaßen zwischen binären Merkmalen

Objekt i \ Objekt j	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	p

Table 4.4. Hilfstabelle zur Berechnung von Distanzmaßen zwischen binären Merkmalen mit den Werten von zwei Studenten

1. Student \ 2. Student	1	0	
1	1	1	2
0	0	2	2
	1	3	4

Schauen wir uns zunächst symmetrische binäre Merkmale an. Die Werte eines Ähnlichkeitsmaßes sollten sich nicht ändern, wenn die Kodierung der binären Merkmale vertauscht wird. Das Ähnlichkeitsmaß sollte also von $a + d$ und $b + c$ abhängen.

Gower & Legendre (1986) betrachten eine Klasse von Ähnlichkeitskoeffizienten s_{ij}^{GL1}, die für symmetrische Merkmale geeignet sind. Diese sind definiert durch

$$s_{ij}^{GL1} = \frac{a+d}{a+d+\theta(b+c)}, \tag{4.3}$$

wobei $\theta > 0$ gilt. Die Distanzmaße d_{ij}^{GL1} erhält man durch

$$d_{ij}^{GL1} = 1 - s_{ij}^{GL1}\,.$$

Durch θ kann man steuern, ob die Anzahl $a+d$ der Übereinstimmungen oder die Anzahl $b + c$ der Nichtübereinstimmungen ein stärkeres Gewicht erhält. Den Simple-Matching-Koeffizienten erhält man für $\theta = 1$. Es gilt

$$s_{ij}^{SM} = \frac{a+d}{a+b+c+d} \tag{4.4}$$

und

$$d_{ij}^{SM} = \frac{b+c}{a+b+c+d}\,. \tag{4.5}$$

hmcounterend. (fortgesetzt)

Example 18. Es gilt

$$s_{12}^{SM} = \frac{3}{4}, \qquad d_{12}^{SM} = \frac{1}{4}.$$

□

Für $\theta = 2$ ergibt sich der von Rogers & Tanimoto (1960) vorgeschlagene Koeffizient s_{ij}^{RT}:

$$s_{ij}^{RT} = \frac{a+d}{a+d+2(b+c)} \tag{4.6}$$

und

$$d_{ij}^{RT} = \frac{2(b+c)}{a+d+2(b+c)}. \tag{4.7}$$

hmcounterend. (fortgesetzt)

Example 18. Es gilt

$$s_{12}^{RT} = \frac{3}{5}, \qquad d_{12}^{RT} = \frac{2}{5}.$$

□

Gower & Legendre (1986) betrachten noch weitere Spezialfälle. Wir wollen auf diese aber nicht eingehen, sondern wenden uns den Koeffizienten für asymmetrische Merkmale zu. In Analogie zu symmetrischen Merkmalen betrachten Gower & Legendre (1986) folgende Klasse von Ähnlichkeitskoeffizienten:

$$s_{ij}^{GL2} = \frac{a}{a+\theta(b+c)}, \tag{4.8}$$

wobei $\theta > 0$ gilt. Die Distanzmaße d_{ij}^{GL2} erhält man durch

$$d_{ij}^{GL2} = 1 - s_{ij}^{GL2}.$$

Für $\theta = 1$ erhält man den Jaccard-Koeffizienten. Es gilt

$$s_{ij}^{JA} = \frac{a}{a+b+c} \tag{4.9}$$

und

$$d_{ij}^{JA} = \frac{b+c}{a+b+c}. \tag{4.10}$$

hmcounterend. (fortgesetzt)

Example 18. Es gilt

$$s_{12}^{JA} = \frac{1}{2}, \qquad d_{12}^{JA} = \frac{1}{2}.$$

□

Für $\theta = 2$ erhält man den von Sneath & Sokal (1973) vorgeschlagenen Koeffizienten:

$$s_{ij}^{SO} = \frac{a}{a + 2(b + c)} \tag{4.11}$$

und

$$d_{ij}^{SO} = \frac{2(b + c)}{a + 2(b + c)}. \tag{4.12}$$

hmcounterend. (fortgesetzt)

Example 18. Es gilt

$$s_{12}^{SO} = \frac{1}{3}, \qquad d_{12}^{SO} = \frac{2}{3}.$$

□

4.2.3 Qualitative Merkmale mit mehr als zwei Merkmalsausprägungen

Wir betrachten nun qualitative Merkmale mit mehr als zwei Merkmalsausprägungen. Sind alle p Merkmale nominal, so wird von Sneath vorgeschlagen:

$$s_{ij} = \frac{u}{p}$$

und

$$d_{ij} = \frac{p - u}{p},$$

wobei u die Anzahl der Merkmale ist, bei denen beide Objekte dieselbe Merkmalsausprägung besitzen.

4.2.4 Qualitative Merkmale, deren Merkmalsausprägungen geordnet sind

Die Ausprägungen eines ordinalen Merkmals seien der Größe nach geordnet, z.B. `sehr gut`, `gut`, `mittel`, `schlecht` und `sehr schlecht`. Wir ordnen den Ausprägungen die Ränge 1, 2, 3, 4 und 5 zu. Die Distanz zwischen zwei Objekten bei einem ordinalen Merkmal erhalten wir dadurch, dass wir den Absolutbetrag der Differenz durch die Spannweite der Ausprägungen des Merkmals dividieren. hmcounterend. (fortgesetzt)

Example 18. Das Merkmal `Cola` ist ordinal. Die Spannweite beträgt $3-1 = 2$. Die Distanz zwischen dem ersten und zweiten Studenten mit den Merkmalsausprägungen 2 beziehungsweise 1 beträgt also 0.5. □

4.2.5 Unterschiedliche Messniveaus

Reale Datensätze bestehen immer aus Merkmalen mit unterschiedlichem Messniveau. Von Gower (1971) wurde folgender Koeffizient für gemischte Merkmale vorgeschlagen:

$$d_{ij} = \frac{\sum_{k=1}^{p} \delta_{ij}^{(k)} \, d_{ij}^{(k)}}{\sum_{f=1}^{p} \delta_{ij}^{(k)}}.$$

Durch $\delta_{ij}^{(k)}$ werden zum einen fehlende Beobachtungen und zum anderen die Symmetrie binärer Merkmale berücksichtigt. Fehlende Beobachtungen werden dadurch berücksichtigt, dass $\delta_{ij}^{(k)}$ gleich 1 ist, wenn das k-te Merkmal bei beiden Objekten beobachtet wurde. Fehlt bei mindestens einem Objekt der Wert des k-ten Merkmals, so ist $\delta_{ij}^{(k)}$ gleich 0. Asymmetrische binäre Merkmale werden dadurch berücksichtigt, dass $\delta_{ij}^{(k)}$ gleich 0 gesetzt wird, wenn bei einem asymmetrischen binären Merkmal beide Objekte den Wert 0 annehmen.

In Abhängigkeit vom Messniveau des Merkmals k wird die Distanz $d_{ij}^{(k)}$ zwischen dem i-ten und j-ten Objekt mit den Merkmalsausprägungen x_{ik} beziehungsweise x_{jk} folgendermaßen bestimmt:

– Bei binären und nominalskalierten Merkmalen gilt

$$d_{ij}^{(k)} = \begin{cases} 1 \text{ wenn} & x_{ik} \neq x_{jk} \\ 0 \text{ wenn} & x_{ik} = x_{jk}. \end{cases}$$

– Bei quantitativen Merkmalen und ordinalen Merkmalen, deren Ausprägungsmöglichkeiten gleich den Rängen $1, \ldots, r$ sind, gilt

$$d_{ij}^{(k)} = \frac{|x_{ik} - x_{jk}|}{R_k}$$

mit $R_k = \max\limits_i x_{ik} - \min\limits_i x_{ik}$ für $i = 1, \ldots, n$.

Sind alle Merkmale quantitativ, und fehlen keine Beobachtungen, dann ist der *Gower-Koeffizient* gleich der Manhattan-Metrik angewendet auf die durch die Spannweite skalierten Merkmale. Dies folgt sofort aus der Definition des Gower-Koeffizienten. Sind alle Merkmale ordinal, dann ist der Gower-Koeffizient gleich der Manhattan-Metrik angewendet auf die durch die Spannweite skalierten Ränge. Sind alle Merkmale symmetrisch binär, so ist der Gower-Koeffizient gleich dem Simple-Matching-Koeffizienten. Die Distanz zwischen zwei Objekten ist nämlich bei einem Merkmal gleich 0, wenn die beiden Objekte bei dem Merkmal unterschiedliche Werte annehmen, ansonsten

ist sie gleich 1. Wir zählen also, bei wie vielen Merkmalen sich die Objekte unterscheiden und dividieren diese Anzahl durch die Anzahl der Merkmale. Sind alle Merkmale asymmetrisch binär, so ist der Gower-Koeffizient gleich dem Jaccard-Koeffizienten. Merkmale, bei denen beide Objekte den Wert 0 annehmen, werden bei der Zählung nicht berücksichtigt. Ansonsten liefern zwei Objekte mit unterschiedlichen Merkmalsausprägungen den Wert 0 und Objekte mit identischer Merkmalsausprägung den Wert 1.

hmcounterend. (fortgesetzt)

Example 18. Wir bestimmen den Gower-Koeffizienten zwischen dem 2-ten und 5-ten Studenten in Tabelle 1.3 auf Seite 6. Aus Gründen der Übersichtlichkeit geben wir die Werte hier noch einmal an. Sie sind in Tabelle 4.5 zu finden.

Table 4.5. Ergebnis der Befragung von 5 Erstsemesterstudenten

Student	Geschlecht	Alter	Größe	Gewicht	Raucher	Auto	Cola	MatheLK
1	0	23	171	60	0	1	2	1
2	0	21	187	75	0	1	1	0
3	1	20	180	65	0	0	3	1
4	1	20	165	55	1	0	2	1
5	0	23	193	81	0	0	3	0

Die Merkmale `Geschlecht`, `Raucher`, `Auto` und `MatheLK` sind binär, wobei wir annehmen, dass alle symmetrisch sind. Die Merkmale `Alter`, `Groesse`, `Gewicht` und `Cola` sind metrisch beziehungsweise ordinal. Da keine fehlenden Beobachtungen vorliegen und alle binären Merkmale symmetrisch sind, sind alle $\delta_{ij}^{(k)}$ gleich 1. Die Spannweite des Merkmals `Alter` ist gleich 3, die Spannweite des Merkmals `Groesse` ist gleich 28, die Spannweite des Merkmals `Gewicht` ist gleich 26 und die Spannweite des Merkmals `Cola` ist gleich 2. Für den Gower-Koeffizienten zwischen dem 2-ten und 5-ten Studenten gilt also

$$d_{25} = \frac{1}{8}\left(0 + \frac{2}{3} + \frac{6}{28} + \frac{6}{26} + 0 + 1 + \frac{2}{2} + 0\right) = 0.389.$$

Die folgende Matrix enthält die Distanzen zwischen den Studenten, die wir nach Anwendung des Gower-Koeffizienten erhalten:

$$\mathbf{D} = \begin{pmatrix} 0 & 0.414 & 0.502 & 0.551 & 0.512 \\ 0.414 & 0 & 0.621 & 0.799 & 0.389 \\ 0.502 & 0.621 & 0 & 0.303 & 0.510 \\ 0.551 & 0.799 & 0.303 & 0 & 0.813 \\ 0.512 & 0.389 & 0.510 & 0.813 & 0 \end{pmatrix}. \tag{4.13}$$

Wir sehen, dass sich die Studenten 3 und 4 am ähnlichsten und die Studenten 4 und 5 am unähnlichsten sind.

Wir wollen nun noch den Gower-Koeffizienten zwischen dem 2-ten und 5-ten Studenten für den Fall bestimmen, dass das Merkmal `MatheLK` asymmetrisch binär ist:

$$d_{25} = \frac{1}{7}\left(0 + \frac{2}{3} + \frac{6}{28} + \frac{6}{26} + 0 + 1 + \frac{2}{2}\right) = 0.445.$$

□

4.3 Distanzmaße in S-PLUS

Wir wollen in `S-PLUS` Distanzmaße für die Werte in Tabelle 1.3 auf Seite 6 bestimmen. Die Daten mögen in der Matrix `student5` stehen:

```
> student5
  Geschlecht Alter Groesse Gewicht Raucher Auto Cola MatheLK
1          0    23     171      60       0    1    2       1
2          0    21     187      75       0    1    1       0
3          1    20     180      65       0    0    3       1
4          1    20     165      55       1    0    2       1
5          0    23     193      81       0    0    3       0
```

In `S-PLUS` gibt es die Funktion `dist`, mit der man eine Reihe von Distanzmaßen bestimmen kann. Der Aufruf von `dist` ist

```
dist(x, metric = "euclidean")
```

Dabei ist `x` die Datenmatrix, die aus n Zeilen und p Spalten besteht. Die Metrik übergibt man mit dem Argument `metric`. Die euklidische Distanz erhält man, wenn man das Argument `metric` auf `"euclidean"` setzt. Weist man `metric` beim Aufruf den Wert `"manhattan"` zu, so wird die Manhattan-Metrik bestimmt. Sind alle Merkmale asymmetrisch binär, so wird der Jaccard-Koeffizient bestimmt. Das Ergebnis von `dist` ist ein Vektor der Länge $0.5 \cdot n \cdot (n-1)$, der die Elemente der unteren Hälfte der Distanzmatrix enthält. Neben den Distanzen liefert die Funktion `dist` noch das Attribut `size`, das die Anzahl der Beobachtungen enthält.

Wir wollen die Funktion `dist` anwenden. Beginnen wir mit den metrischen Merkmalen. Wir wählen die metrischen Merkmale aus der Matrix `student5` aus und weisen sie der Variablen `quant` zu:

```
> quant<-student5[,2:4]
> quant
  Alter Groesse Gewicht
1    23     171      60
2    21     187      75
3    20     180      65
4    20     165      55
5    23     193      81
```

Die euklidischen Distanzen erhalten wir durch

```
> d<-dist(quant,metric="euclidean")
> d
 [1] 22.022715 10.723805  8.366600 30.413813 12.247449
     29.748949  8.717798 18.027756 20.832666 38.327538
 attr(, "Size"):
 [1] 5
```

Die meisten Funktionen in S-PLUS, deren Argument eine Distanzmatrix ist, können auch mit dem Ergebnis der Funktion `dist` aufgerufen werden. An einigen Stellen benötigen wir jedoch die volle Distanzmatrix. Diese gewinnen wir dadurch aus `d`, dass wir zunächst die Anzahl der Beobachtungen bestimmen:

```
> n<-attr(d,"Size")
> n
[1] 5
```

und dann eine (n, n)-Matrix `dm` erzeugen, die aus Nullen besteht:

```
> dm<-matrix(0,n,n)
```

Anschließend weisen wir den Elementen unterhalb der Hauptdiagonalen von `dm` die Elemente des Vektors `d` zu. Hierzu verwenden wir die Funktion `lower.tri`:

```
> dm[lower.tri(dm)]<-d
> dm
          [,1]      [,2]     [,3]     [,4] [,5]
 [1,]  0.00000  0.000000  0.00000  0.00000    0
 [2,] 22.02271  0.000000  0.00000  0.00000    0
 [3,] 10.72381 12.247449  0.00000  0.00000    0
 [4,]  8.36660 29.748949 18.02776  0.00000    0
 [5,] 30.41381  8.717798 20.83267 38.32754    0
```

Nun addieren wir zu `dm` die Transponierte von `dm` mit der Funktion `t` und sind fertig:

```
> dm<-dm+t(dm)
> dm
          [,1]      [,2]     [,3]     [,4]      [,5]
 [1,]  0.00000 22.022715 10.72381  8.36660 30.413813
 [2,] 22.02271  0.000000 12.24745 29.74895  8.717798
 [3,] 10.72381 12.247449  0.00000 18.02776 20.832666
 [4,]  8.36660 29.748949 18.02776  0.00000 38.327538
 [5,] 30.41381  8.717798 20.83267 38.32754  0.000000
```

Im Anhang ist auf Seite 495 eine Funktion `distfull` zu finden, die aus einem Vektor mit Distanzen die Distanzmatrix erstellt. Wir werden diese Funktion häufiger verwenden.

Wir wenden die Funktion `distfull` an und runden die Werte mit der Funktion `round` auf zwei Stellen nach dem Dezimalpunkt:

```
> round(distfull(dist(quant,metric="euclidean")),2)
```

Dieser Aufruf liefert das Ergebnis

```
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,]  0.00 22.02 10.72  8.37 30.41
[2,] 22.02  0.00 12.25 29.75  8.72
[3,] 10.72 12.25  0.00 18.03 20.83
[4,]  8.37 29.75 18.03  0.00 38.33
[5,] 30.41  8.72 20.83 38.33  0.00
```

Entsprechend erhalten wir die Matrix mit den Distanzen der Manhattan-Metrik:

```
> distfull(dist(quant,metric="manhattan"))
     [,1] [,2] [,3] [,4] [,5]
[1,]    0   33   17   14   43
[2,]   33    0   18   43   14
[3,]   17   18    0   25   32
[4,]   14   43   25    0   57
[5,]   43   14   32   57    0
```

Um die skalierte euklidische und Manhattan-Metrik zu erhalten, müssen wir die Daten erst entsprechend skalieren. Fangen wir mit der skalierten euklidischen Metrik an. Hier dividieren wir die Werte jedes Merkmals durch ihre Standardabweichung. Einen Vektor mit den Standardabweichungen erhalten wir mit Hilfe der Funktion `apply`:

```
> sqrt(apply(quant,2,var))
    Alter  Groesse  Gewicht
 1.516575 11.41052 10.68644
```

Nun müssen wir noch jede Spalte durch die entsprechende Standardabweichung dividieren. Hierzu verwenden wir die Funktion `sweep`. Der Aufruf

```
> sweep(quant,2,sqrt(apply(quant,2,var)),"/")
      Alter  Groesse  Gewicht
 1 15.16575 14.98617 5.614592
 2 13.84699 16.38838 7.018240
 3 13.18761 15.77491 6.082475
 4 13.18761 14.46034 5.146709
 5 15.16575 16.91421 7.579699
```

liefert die Matrix der skalierten Werte. Auf diese können wir die Funktion `dist` anwenden:

```
> dist(sweep(quant,2,sqrt(apply(quant,2,var)),"/"))
 [1] 2.382344 2.180384 2.099632 2.752998 1.298762
     2.766725 1.526716 1.613619 2.729968 3.981706
 attr(, "Size"):
 [1] 5
```

Um die skalierte Manhattan-Metrik zu erhalten, definieren wir eine Funktion `Spannweite` durch

```
> Spannweite<-function(x) {max(x)-min(x)}
```

Die Funktionen `min` und `max` bestimmen das Minimum und Maximum der Komponenten eines Vektors. Wir wenden wieder die Funktionen `sweep` und `apply` an:

```
> sweep(quant,2,apply(quant,2,Spannweite),"/")
      Alter  Groesse  Gewicht
 1 7.666667 6.107143 2.307692
 2 7.000000 6.678571 2.884615
 3 6.666667 6.428571 2.500000
 4 6.666667 5.892857 2.115385
 5 7.666667 6.892857 3.115385
```

Die Distanzen zwischen den skalierten Merkmalen erhalten wir also durch

```
> dist(sweep(quant,2,apply(quant,2,Spannweite),"/"),
       metric="manhattan")
  [1] 1.8150179 1.5137360 1.4065936 1.5934064 0.9679489
      1.8882785 1.1117215 0.9203296 2.0796704 3.0000000
 attr(, "Size"): [1] 5
```

Schauen wir uns die binären Merkmale an. Wir weisen diese der Matrix `binaer` zu:

```
> binaer<-student5[,c(1,5,6,8)]
```

Diese sieht folgendermaßen aus:

```
> binaer
   Geschlecht Raucher Auto MatheLK
 1          0       0    1       1
 2          0       0    1       0
 3          1       0    0       1
 4          1       1    0       1
 5          0       0    0       0
```

Unterstellen wir, dass alle Merkmale asymmetrisch sind, dann erhalten wir den Jaccard-Koeffizienten mit

```
> dist(binaer,metric="binary")
  [1] 0.5000000 0.6666667 0.7500000 1.0000000 1.0000000
      1.0000000 1.0000000 0.3333333 1.0000000 1.0000000
 attr(, "Size"):
 [1] 5
```

Auch den Simple-Matching-Koeffizienten können wir mit der Funktion `dist` bestimmen. Wendet man die Manhattan-Distanzen auf eine Datenmatrix an, die aus Nullen und Einsen besteht, so erhält man für jedes Objektpaar die Anzahl der Fälle, in denen sie nicht übereinstimmen. Diese Anzahl müssen wir nur noch durch die Anzahl der Beobachtungen teilen, um den Simple-Matching-Koeffizienten zu erhalten:

```
> dist(binaer,metric="manhattan")/dim(binaer)[2]
  [1] 0.25 0.50 0.75 0.50 0.75 1.00 0.25 0.25 0.50 0.75
  attr(, "Size"):
  [1] 5
```

Die Funktion `dim` gibt die Dimension einer Matrix, also die Anzahl der Zeilen und Spalten, an. Liegen keine qualitativen Merkmale mit mehr als zwei Kategorien vor, fehlen keine Beobachtungen, und sind alle binären Merkmale symmetrisch, so erhält man den Gower-Koeffizienten dadurch, dass man alle Merkmale durch ihre Spannweite dividiert, auf das Ergebnis die Funktion `dist` mit `metric="manhattan"` anwendet und das Ergebnis durch die Anzahl der Merkmale dividiert:

```
> dist(sweep(student5,2,apply(student5,2,Spannweite),"/"),
       metric="manhattan")/8
  [1] 0.4143772 0.5017170 0.5508242 0.5116758 0.6209936
      0.7985348 0.3889652 0.3025412 0.5099588 0.8125000
 attr(, "Size"):
 [1] 5
```

Leider können hierbei keine fehlenden Beobachtungen und auch keine asymmetrischen binären Variablen berücksichtigt werden. Dies ist mit der Funktion `daisy` möglich, die bei Kaufman & Rousseeuw (1990) beschrieben wird. Diese kann man nur verwenden, wenn man die Programmbibliothek `cluster` geladen hat. Dies geschieht durch

```
> library(cluster)
```

Schauen wir uns an, wie man `daisy` verwendet, wenn man für die Daten in `student5` den Gower-Koeffizienten berechnen will. Dabei gehen wir davon aus, dass die Merkmale in den Spalten 1, 5, 6 und 8 symmetrisch binär sind. Die Variablen in den Spalten 2, 3 und 4 sind metrisch und die Variable in Spalte 7 ist ordinal.

Wir müssen zuerst die binären und ordinalen Merkmale geeignet transformieren. Wir machen aus jedem der binären Merkmale einen Faktor. Dies geschieht mit der Funktion `factor`:

```
> f1<-factor(student5[,1])
> f2<-factor(student5[,5])
> f3<-factor(student5[,6])
> f4<-factor(student5[,8])
```

Aus dem ordinalen Merkmal in der 7-ten Spalte machen wir mit der Funktion `ordered` einen geordneten Faktor:

```
> o1<-ordered(student5[,7])
```

Anschließend machen wir aus allen Variablen einen Dataframe, wie dies auf Seite 64 beschrieben wird:

```
> m<-data.frame(f1,student5[,2:4],f2,f3,o1,f4)
```

Der Aufruf

```
> daisy(m)
```

liefert die Werte des Gower-Koeffizienten:

```
Dissimilarities :
 [1] 0.4143773 0.5017170 0.5508242 0.5116758 0.6209936
     0.7985348 0.3889652 0.3025412 0.5099588 0.8125000
 Metric :  mixed
 Number of objects :  5
```

Soll das Merkmal `MatheLK` in der achten Spalte binär asymmetrisch sein, so rufen wir die Funktion `daisy` folgendermaßen auf:

```
> daisy(m,type=list(asymm=8))
Dissimilarities :
 [1] 0.4143773 0.5017170 0.5508242 0.5116758 0.6209936
     0.7985348 0.4445317 0.3025412 0.5099588 0.8125000

 Metric :  mixed
 Number of objects :  5
```

Es hat sich nur die Distanz zwischen dem zweiten und fünften Studenten geändert, da diese keinen Mathematik-Leistungskurs besucht haben.

4.4 Direkte Bestimmung der Distanzen

Bisher haben wir die Distanzen aus der Datenmatrix bestimmt. Man kann die Distanzen auch direkt bestimmen. In Kapitel 1 haben wir dafür auf Seite 6 ein Beispiel kennengelernt. Dort sind die Entfernungen zwischen Städten in Deutschland angegeben.

In den Sozialwissenschaften will man untersuchen, wie n Subjekte oder Objekte wahrgenommen werden. Hierzu bestimmt man die Distanzen zwischen allen möglichen Paaren der Subjekte beziehungsweise Objekte. Hierzu gibt es eine Reihe von Verfahren, die wir im Folgenden näher betrachten wollen.

Example 19. Es soll untersucht werden, wie ein Student 5 Politiker beurteilt. Hierzu werden alle Paare dieser 5 Politiker betrachtet. Diese sind in Tabelle 4.6 zu finden.

Table 4.6. Alle Paare aus einer Menge von 5 Politikern

1. Politiker	2. Politiker
Fischer	Merkel
Fischer	Schröder
Fischer	Stoiber
Fischer	Westerwelle
Merkel	Schröder
Merkel	Stoiber
Merkel	Westerwelle
Schröder	Stoiber
Schröder	Westerwelle
Stoiber	Westerwelle

□

Bei der Beurteilung der Ähnlichkeit von Politikern innerhalb eines Paares gibt es mehrere Möglichkeiten, von denen wir zwei näher betrachten wollen. Die erste Möglichkeit besteht darin, die Ähnlichkeit jedes Paares von Personen auf einer Skala von 1 bis n zu bewerten, wobei das Paar den Wert 1 erhält, wenn sich die beiden Personen sehr ähnlich sind, und den Wert n, wenn sich die Personen sehr unähnlich sind. Man spricht in diesem Fall vom *Ratingverfahren*.

hmcounterend. (fortgesetzt)

Example 19. Tabelle 4.7 zeigt die Ergebnisse eines Studenten, wobei $n = 7$ gilt.

Table 4.7. Vergleich aller Paare aus einer Menge von 5 Politikern mit dem Ratingverfahren

1. Politiker	2. Politiker	Rating
Fischer	Merkel	5
Fischer	Schröder	2
Fischer	Stoiber	7
Fischer	Westerwelle	5
Merkel	Schröder	3
Merkel	Stoiber	1
Merkel	Westerwelle	2
Schröder	Stoiber	5
Schröder	Westerwelle	3
Stoiber	Westerwelle	2

□

Das andere Verfahren besteht darin, die Paarvergleiche der Größe nach zu ordnen, wobei das ähnlichste Paar eine 1, das zweitähnlichste eine 2, u.s.w. erhält. Man spricht in diesem Fall von *Rangreihung*. hmcounterend. (fortgesetzt)

Example 19. Die Ergebnisse eines Studenten sind in Tabelle 4.8 zu finden.

Table 4.8. Vergleich aller Paare aus einer Menge von 5 Politikern mit dem Verfahren der Rangreihung

1. Politiker	2. Politiker	Rang
Fischer	Merkel	9
Fischer	Schröder	4
Fischer	Stoiber	10
Fischer	Westerwelle	7
Merkel	Schröder	3
Merkel	Stoiber	1
Merkel	Westerwelle	2
Schröder	Stoiber	8
Schröder	Westerwelle	6
Stoiber	Westerwelle	5

□

4.5 Übungen

Exercise 4. Betrachten Sie das Beispiel 12 auf Seite 11.

1. Bestimmen Sie die Distanzen zwischen den Regionen mit der euklidischen und der standardisierten euklidischen Distanz. Welche der beiden Vorgehensweisen halten Sie für angemessener?
2. Bestimmen Sie die Distanzen zwischen den Regionen mit der Manhattan-Metrik und der standardisierten Manhattan-Metrik. Welche der beiden Vorgehensweisen halten Sie für angemessener?

Exercise 5. Im Wintersemester 1999/2000 wurden 191 Studenten in der Veranstaltung Einführung in die Ökonometrie befragt. Neben dem Merkmal `Geschlecht` mit den Merkmalsausprägungen `w` und `m` wurden noch die Merkmale `Alter`, `Größe`, `Gewicht` und `Miete` der Wohnung erhoben. Außerdem wurden die Studierenden gefragt, ob sie bei den Eltern wohnen, ob sie einen eigenen PC besitzen und ob sie nach dem Abitur eine Berufsausbildung gemacht haben. Wir bezeichnen diese Merkmale mit `Eltern`, `PC` und `Ausbildung`. Ihre Ausprägungsmöglichkeiten sind `j` und `n`. Als letztes sollten die Studierenden ihre Durchschnittsnote im Abitur angeben. Dieses Merkmal bezeichnen wir mit `Note`. Die Ergebnisse der Befragung von 3 Studenten sind in Tabelle 4.9 zu finden.

Table 4.9. Ergebnis der Befragung von 3 Studenten

Geschlecht	Alter	Größe	Gewicht	Eltern	Miete	PC	Ausbildung	Note
w	22	179	65	n	550	n	n	4.0
w	19	205	80	n	1200	j	n	1.0
m	29	175	75	n	200	j	j	1.7

1. Wählen Sie die quantitativen Merkmale aus und bestimmen Sie die Distanzen zwischen den drei Studenten hinsichtlich dieser Merkmale bezüglich der euklidischen Distanz.
2. Wählen Sie die quantitativen Merkmale aus und bestimmen Sie die Distanzen zwischen den drei Studenten hinsichtlich dieser Merkmale bezüglich der Manhattan-Metrik und der skalierten Manhattan-Metrik.
3. Erklären Sie kurz, warum es im Beispiel sinnvoll ist, die skalierte Manhattan-Metrik zu betrachten.
4. Was ist der Unterschied zwischen symmetrischen und asymmetrischen binären Merkmalen?
5. Wählen Sie die binären Merkmale aus und gehen Sie davon aus, dass alle binären Merkmale symmetrisch sind. Bestimmen Sie die Ähnlichkeiten und Distanzen zwischen den drei Studenten hinsichtlich des Simple-Matching-Koeffizienten.

6. Welche Idee steckt hinter dem Simple-Matching-Koeffizienten?
7. Bestimmen Sie die Werte des Gower-Koeffizienten zwischen den drei Studenten.

Part II

Darstellung hochdimensionaler Daten in niedrigdimensionalen Räumen

5 Hauptkomponentenanalyse

5.1 Problemstellung

Ausgangspunkt vieler Anwendungen ist eine Datenmatrix, die mehr als zwei quantitative Merkmale enthält. In einer solchen Situation ist man sehr oft daran interessiert die Objekte der Größe nach zu ordnen, wobei alle Merkmale in Betracht gezogen werden sollen. Außerdem will man die Objekte in einem Streudiagramm darstellen. Auch hier sollen alle Merkmale bei der Darstellung berücksichtigt werden. Es liegt nahe, das zweite Ziel durch Zeichnen der Streudiagrammmatrix erreichen zu wollen. Bei dieser werden aber immer nur Paare von Merkmalen betrachtet, sodass der Zusammenhang zwischen allen p Merkmalen nicht berücksichtigt wird. Wir werden im Folgenden ein Verfahren kennenlernen, mit dessen Hilfe man ein Streudiagramm zeichnen kann, bei dem bei beiden Achsen alle Merkmale berücksichtigt werden. Außerdem kann man auch die Objekte ordnen, indem man ihre Werte bezüglich der ersten Achse des Streudiagramms betrachtet. Um das erste Ziel zu erreichen, bestimmt man oft den Mittelwert aller Merkmale bei jedem Objekt. Sei x_{ij} der Wert des j-ten Merkmals beim i-ten Objekt, dann ist der Mittelwert $\bar{x}_i$ der Werte des i-ten Objekts folgendermaßen definiert:

$$\bar{x}_i = \frac{1}{p} \sum_{j=1}^{p} x_{ij}.$$

Example 20. Im Beispiel 4 auf Seite 5 betrachten wir die Noten der Studenten in den Bereichen `Mathematik`, `BWL`, `VWL` und `Methoden`. Die Mittelwerte sind in Tabelle 5.1 zu finden. Wählt man die Durchschnittsnote als Kriterium, ist der 17. Student am besten und der 16. am schlechtesten.

□

Man kann den Mittelwert auch schreiben als

$$\bar{x}_i = \sum_{j=1}^{p} a_j \, x_{ij}$$

mit $a_j = \frac{1}{p}$.

Table 5.1. Durchschnittsnoten der 17 Studenten

Student	Durchschnittsnote
1	1.48
2	1.92
3	3.00
4	1.44
5	2.86
6	2.26
7	2.71
8	2.77
9	1.89
10	2.77
11	2.03
12	1.81
13	1.54
14	2.42
15	1.92
16	3.12
17	1.22

Der Mittelwert ist eine Linearkombination der Merkmalswerte, bei der alle Gewichte gleich sind. Das ist nicht notwendigerweise die beste Wahl. Die Objekte werden sich bezüglich der p Merkmale unterscheiden. Diese Unterschiede sollten beim Übergang zu einem Wert mit Hilfe einer Linearkombination im Wesentlichen erhalten bleiben. Wie ist das möglich? Und woran bemerkt man, wie gut dieses Ziel erreicht wurde?

Example 21. Wir schauen uns das Beispiel 1 auf Seite 3 an und wählen 5 Länder und die Merkmale `Lesekompetenz` und `Mathematische Grundbildung` aus. Die Werte sind in Tabelle 5.2 zu finden.

Table 5.2. Die Merkmale Lesekompetenz und Mathematische Grundbildung von 5 Ländern im Rahmen der PISA-Studie

Land	Lesekompetenz	Mathematische Grundbildung
DK	497	514
GR	474	447
I	487	457
P	470	454
S	516	510

Abbildung 5.1 zeigt das Streudiagramm der Daten.

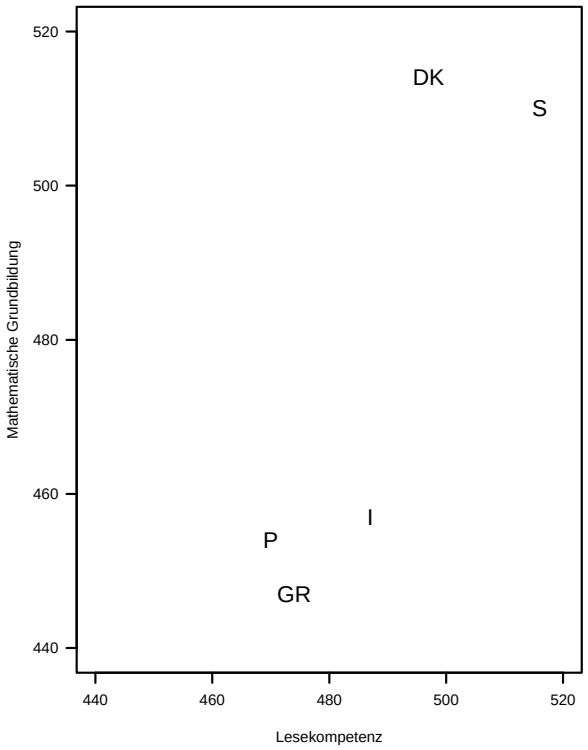


Fig. 5.1. Streudiagramm der Merkmale Lesekompetenz und Mathematische Grundbildung von 5 Ländern im Rahmen der PISA-Studie

Wir erkennen, dass das Merkmal `Mathematische Grundbildung` viel stärker streut als das Merkmal `Lesekompetenz`. Außerdem sehen wir, dass es zwei Gruppen gibt. Die eine Gruppe besteht aus Portugal, Griechenland und Italien, die andere aus Dänemark und Schweden. Wir wollen nun die beiden Merkmale durch eine Linearkombination ersetzen, die möglichst viel von der zweidimensionalen Struktur enthält. Die Interpretation der Linearkombination vereinfacht sich beträchtlich, wenn man die Merkmale zentriert. Wir subtrahieren also von den Werten eines Merkmals den Mittelwert des Merkmals. Der Mittelwert des Merkmals `Lesekompetenz` beträgt 488.8 und der Mittelwert des Merkmals `Mathematische Grundbildung` 476.4. Subtrahieren wir die Mittelwerte von den Werten der jeweiligen Merkmale, so erhalten wir folgende Datenmatrix der zentrierten Merkmale:

$$\tilde{\mathbf{X}} = \begin{pmatrix} 8.2 & 37.6 \\ -14.8 & -29.4 \\ -1.8 & -19.4 \\ -18.8 & -22.4 \\ 27.2 & 33.6 \end{pmatrix}.$$

□

Wir bezeichnen die zentrierten Merkmale mit $\tilde{x}_1$ und $\tilde{x}_2$ und bilden eine Linearkombination

$$a_1\,\tilde{x}_1 + a_2\,\tilde{x}_2$$

dieser Merkmale. Wie sollen wir a_1 und a_2 wählen? Am einfachsten ist es, nur eines der beiden Merkmale zu betrachten. Das bedeutet, dass man bei der Linearkombination a_1 gleich 1 und a_2 gleich 0 setzt, wenn nur das erste Merkmal und a_1 gleich 0 und a_2 gleich 1 setzt, wenn nur das zweite Merkmal berücksichtigt werden soll.

hmcounterend. (fortgesetzt)

Example 21. Die folgende Graphik zeigt die beiden eindimensionalen Darstellungen:

1

Die erste Zeile zeigt die Verteilung der Länder hinsichtlich des Merkmals `Lesekompetenz`, die zweite Zeile hinsichtlich des Merkmals `Mathematische Grundbildung`. Vergleichen wir diese Abbildung mit der zweidimensionalen Konfiguration in Abbildung 5.1, so sehen wir, dass die eindimensionale Darstellung bezüglich des Merkmals `Mathematische Grundbildung` die Beziehungen zwischen den Ländern viel besser wiedergibt als die Darstellung bezüglich des Merkmals `Lesekompetenz`. Die Länder Griechenland, Portugal und Italien liegen bezüglich des Merkmals `Mathematische Grundbildung` relativ nahe beieinander und sind von den Ländern Dänemark und Schweden relativ weit entfernt. Das ist auch bei der zweidimensionalen Darstellung der Fall. Woran liegt das? Schauen wir uns die Stichprobenvarianzen der beiden Merkmale an. Die Stichprobenvarianz des Merkmals `Mathematische Grundbildung` beträgt 1071.3, die des Merkmals `Lesekompetenz` beträgt 345.7. Die Streuung des Merkmals `Mathematische Grundbildung` ist viel größer als die Streuung des Merkmals `Lesekompetenz`. Eine große Streuung diskriminiert viel stärker zwischen den Ländern. Somit sollte man eine Linearkombination der beiden Merkmale wählen, die eine sehr große Streuung besitzt. □

Bisher haben wir nur Linearkombinationen betrachtet, bei denen genau eines der Merkmale betrachtet wird. Die Gewichte sind hierbei $a_1 = 1$ und $a_2 = 0$ oder $a_1 = 0$ und $a_2 = 1$. Betrachten wir nun eine andere Wahl der Gewichte, so sollte sie mit den bisher betrachteten Gewichten vergleichbar sein. Wir können die Varianz einer Linearkombination beliebig groß machen, indem wir die Gewichte a_1 und a_2 entsprechend vergrößern. Wir müssen die Gewichte also normieren. Naheliegend ist die Wahl

$$a_1 + a_2 = 1.$$

Aus technischen Gründen ist es sinnvoller,

$$a_1^2 + a_2^2 = 1$$

zu wählen. Diese Bedingung ist für die obigen Gewichte erfüllt. hmcounterend. (fortgesetzt)

Example 21. Wir wählen $a_1 = 0.6$ und $a_2 = 0.8$. Bildet man mit diesen Werten die Linearkombination der zentrierten Merkmale, so erhält man die in Tabelle 5.3 angegebenen Werte.

Table 5.3. Werte von $0.6\,\tilde{X}_1 + 0.8\,\tilde{X}_2$ von 5 Ländern im Rahmen der PISA-Studie

Land	Wert der Linearkombination
DK	35.0
GR	-32.4
I	-16.6
P	-29.2
S	43.2

Ein Land erhält bei dieser Linearkombination einen hohen Wert, wenn die Werte beider Merkmale größer als der jeweilige Mittelwert sind. Es erhält einen niedrigen Wert, wenn die Punktezahl des Landes in beiden Bereichen unter dem Durchschnitt liegt. Die Stichprobenvarianz dieser Linearkombination ist 1317.3. Wir sehen, dass diese Linearkombination eine größere Varianz besitzt als jedes der einzelnen Merkmale. Es lohnt sich also, beide Merkmale zu betrachten. □

Schauen wir uns noch eine andere Wahl der Gewichte an. hmcounterend. (fortgesetzt)

Example 21. Sei $a_1 = -0.6$ und $a_2 = 0.8$. Bildet man mit diesen Werten die Linearkombination der zentrierten Merkmale, so erhält man die Werte in Tabelle 5.4. Ein Land erhält bei dieser Linearkombination einen hohen Wert, wenn der Wert des Merkmals `Lesekompetenz` unter dem Durchschnitt und der Wert des Merkmals `Mathematische Grundbildung` über dem Durchschnitt liegt. Es erhält einen niedrigen Wert, wenn der Wert des Merkmals `Lesekompetenz` über dem Durchschnitt und der Wert des Merkmals `Mathematische Grundbildung` unter dem Durchschnitt liegt. Länder, die bei dieser Linearkombination extreme Werte annehmen, sind in einem der Bereiche gut und in dem anderen schlecht. Die Stichprobenvarianz dieser Linearkombination beträgt 302.868.

Table 5.4. Werte von $-0.6\,\tilde{X}_1 + 0.8\,\tilde{X}_2$ von 5 Ländern im Rahmen der PISA-Studie

Land	Wert der Linear-kombination
DK	25.16
GR	-14.64
I	-14.44
P	-6.64
S	10.56

□

5.2 Hauptkomponentenanalyse bei bekannter Varianz-Kovarianz-Matrix

Nun wollen wir zeigen, wie man die optimalen Gewichte bestimmen kann. Hierbei gehen wir aus von der p-dimensionalen Zufallsvariablen $\mathbf{X}$. Es gelte $Var(\mathbf{X}) = \boldsymbol{\Sigma}$. Gesucht ist die Linearkombination $\mathbf{a}_1'\mathbf{X}$ mit größter Varianz unter der Nebenbedingung

$$\mathbf{a}_1'\mathbf{a}_1 = 1.$$

Wie wir in (3.32) auf Seite 88 gezeigt haben, gilt

$$Var(\mathbf{a}_1'\mathbf{X}) = \mathbf{a}_1'\boldsymbol{\Sigma}\mathbf{a}_1.$$

Wir betrachten also folgendes Optimierungsproblem:

$$\max_{\mathbf{a}} \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \tag{5.1}$$

unter der Nebenbedingung

$$\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = 1. \tag{5.2}$$

Wir stellen die Lagrange-Funktion auf (siehe dazu Anhang A.2.3 auf Seite 487):

$$L(\mathbf{a}, \lambda) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} - \lambda\left(\mathbf{a}'\mathbf{a} - 1\right).$$

Die partiellen Ableitungen lauten

$$\frac{\partial}{\partial\,\mathbf{a}} L(\mathbf{a}, \lambda) = 2\boldsymbol{\Sigma}\mathbf{a} - 2\lambda\mathbf{a},$$

$$\frac{\partial}{\partial\,\lambda} L(\mathbf{a}, \lambda) = 1 - \mathbf{a}'\mathbf{a}.$$

Der Vektor $\mathbf{a}_1$ erfüllt die notwendigen Bedingungen für einen Extremwert, wenn gilt

$$2\boldsymbol{\Sigma}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = \mathbf{0} \tag{5.3}$$

und

$$\mathbf{a}_1'\mathbf{a}_1 = 1. \tag{5.4}$$

Aus (5.3) folgt

$$\boldsymbol{\Sigma}\mathbf{a}_1 = \lambda\mathbf{a}_1. \tag{5.5}$$

Bei Gleichung (5.5) handelt es sich um ein Eigenwertproblem. Die notwendigen Bedingungen eines Extremwerts von (5.1) unter der Nebenbedingung (5.2) werden also von den normierten Eigenvektoren der Matrix $\boldsymbol{\Sigma}$ erfüllt. Aufgrund von Gleichung (A.51) auf Seite 479 existieren zu jeder symmetrischen (p,p)-Matrix $\boldsymbol{\Sigma}$ p Eigenwerte $\lambda_1,\ldots,\lambda_p$ mit zugehörigen orthogonalen Eigenvektoren $\mathbf{a}_1,\ldots,\mathbf{a}_p$. Welcher der Eigenvektoren liefert nun die Linearkombination mit der größten Varianz? Um diese Frage zu beantworten, schauen wir uns die Varianz der Linearkombination $\mathbf{a}_1'\mathbf{X}$ an, wenn der Vektor $\mathbf{a}_1$ die Gleichungen (5.4) und (5.5) erfüllt:

$$Var(\mathbf{a}_1'\mathbf{X}) = \mathbf{a}_1'\boldsymbol{\Sigma}\mathbf{a}_1 = \mathbf{a}_1'\lambda\mathbf{a}_1 = \lambda\,\mathbf{a}_1'\mathbf{a}_1 = \lambda.$$

Also ist der Eigenwert λ zum Eigenvektor $\mathbf{a}_1$ die Varianz der Linearkombination $\mathbf{a}_1'\mathbf{X}$. Da wir die Linearkombination mit der größten Varianz suchen, wählen wir den Eigenvektor, der zum größten Eigenwert gehört. Wir nennen diesen die erste *Hauptkomponente*. Wir nummerieren die Eigenwerte und zugehörigen Eigenvektoren nach der Größe der Eigenwerte, wobei der größte Eigenwert den Index 1 erhält.

Wie kann man die anderen Eigenvektoren interpretieren? Wir schauen uns dies exemplarisch für $\mathbf{a}_2$ an. Dieser liefert die Koeffizienten der Linearkombination $\mathbf{a}_2'\mathbf{X}$ mit der größten Varianz unter den Nebenbedingungen

$$\mathbf{a}_2'\mathbf{a}_2 = 1 \tag{5.6}$$

und

$$\mathbf{a}_2'\mathbf{a}_1 = 0\,, \tag{5.7}$$

wobei $\mathbf{a}_1$ die erste Hauptkomponente ist. Die zweite Nebenbedingung besagt, dass $\mathbf{a}_1$ und $\mathbf{a}_2$ orthogonal sind. Wir zeigen nun diesen Sachverhalt.

Wir betrachten also das Optimierungsproblem

$$\max_{\mathbf{a}} \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$$

unter den Nebenbedingungen

$$\mathbf{a}'\mathbf{a} = 1$$

und

$$\mathbf{a}'\mathbf{a}_1 = 0.$$

Die Lagrange-Funktion lautet

$$L(\mathbf{a}, \lambda, \nu) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} - \lambda\left(\mathbf{a}'\mathbf{a} - 1\right) - 2\nu\mathbf{a}'\mathbf{a}_1.$$

Die partiellen Ableitungen lauten

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{a}} L(\mathbf{a}, \lambda, \nu) &= 2\boldsymbol{\Sigma}\mathbf{a} - 2\lambda\mathbf{a} - 2\nu\mathbf{a}_1, \\
\frac{\partial}{\partial \lambda} L(\mathbf{a}, \lambda, \nu) &= 1 - \mathbf{a}'\mathbf{a}, \\
\frac{\partial}{\partial \nu} L(\mathbf{a}, \lambda, \nu) &= -2\mathbf{a}'\mathbf{a}_1.
\end{aligned}$$

Der Vektor $\mathbf{a}_2$ erfüllt also die notwendigen Bedingungen für einen Extremwert, wenn gilt

$$2\boldsymbol{\Sigma}\mathbf{a}_2 - 2\lambda\mathbf{a}_2 - 2\nu\mathbf{a}_1 = 0, \tag{5.8}$$

$$\mathbf{a}_2'\mathbf{a}_2 = 1 \tag{5.9}$$

und

$$\mathbf{a}_2'\mathbf{a}_1 = 0. \tag{5.10}$$

Gleichung (5.8) können wir umformen zu

$$\boldsymbol{\Sigma}\,\mathbf{a}_2 = \lambda\mathbf{a}_2 + \nu\mathbf{a}_1. \tag{5.11}$$

Multiplizieren wir Gleichung (5.11) von links mit dem Eigenvektor $\mathbf{a}_1'$, so gilt

$$\mathbf{a}_1'\boldsymbol{\Sigma}\mathbf{a}_2 = \mathbf{a}_1'\lambda\mathbf{a}_2 + \mathbf{a}_1'\nu\mathbf{a}_1. \tag{5.12}$$

Mit (5.5) und (5.10) gilt

$$\mathbf{a}_1'\boldsymbol{\Sigma}\mathbf{a}_2 = \left(\mathbf{a}_1'\boldsymbol{\Sigma}\mathbf{a}_2\right)' = \mathbf{a}_2'\boldsymbol{\Sigma}\mathbf{a}_1 = \mathbf{a}_2'\lambda_1\mathbf{a}_1 = \lambda_1\mathbf{a}_2'\mathbf{a}_1 = 0. \tag{5.13}$$

Mit Gleichung (5.13) und Gleichung (5.10) vereinfacht sich Gleichung (5.12) zu

$$\nu\mathbf{a}_1'\mathbf{a}_1 = 0.$$

Und hieraus folgt mit (5.4) sofort $\nu = 0$. Somit vereinfacht sich (5.11) zu

$$\boldsymbol{\Sigma}\mathbf{a}_2 = \lambda\mathbf{a}_2. \tag{5.14}$$

$\mathbf{a}_2$ ist also auch Eigenvektor von $\boldsymbol{\Sigma}$. Außerdem ist er orthogonal zu $\mathbf{a}_1$. Wir bezeichnen ihn als zweite Hauptkomponente. Die anderen Eigenvektoren kann man analog interpretieren.

Fassen wir zusammen: Um die Linearkombination $\mathbf{a}'\mathbf{X}$ mit der größten Varianz zu erhalten, bestimmen wir die Eigenwerte und Eigenvektoren der Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$. Der normierte Eigenvektor $\mathbf{a}_1$, der zum größten Eigenwert λ_1 gehört, bildet die Gewichte der gesuchten Linearkombination. Der Eigenvektor $\mathbf{a}_2$, der zum zweitgrößten Eigenwert λ_2 gehört, bildet die Gewichte der Linearkombination mit der größten Varianz unter allen Vektoren, die zum ersten Eigenvektor orthogonal sind. Entsprechend kann man die anderen Eigenvektoren interpretieren.

Bisher haben wir die Hauptkomponentenanalyse nur unter dem Aspekt einer bekannten Varianz-Kovarianz-Matrix kennengelernt. Im nächsten Abschnitt werden wir sehen, wie man datengestützt vorgeht.

5.3 Hauptkomponentenanalyse bei unbekannter Varianz-Kovarianz-Matrix

In der Praxis ist die Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$ in der Regel unbekannt und muss durch die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ geschätzt werden. Die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ ist dann der Ausgangspunkt der Hauptkomponentenanalyse. hmcounterend. (fortgesetzt)

Example 21. Es gilt

$$\mathbf{S} = \begin{pmatrix} 345.70 & 528.35 \\ 528.35 & 1071.30 \end{pmatrix}.$$

Es gilt

$$\det(\mathbf{S} - \lambda\,\mathbf{I}_2) = (345.7 - \lambda)(1071.3 - \lambda) - 279153.7.$$

Wir müssen also folgende quadratische Gleichung lösen:

$$(345.7 - \lambda)(1071.3 - \lambda) - 279153.7 = 0.$$

Wir können diese auch schreiben als

$$\lambda^2 - 1417\,\lambda + 91194.69 = 0.$$

Die Lösungen dieser Gleichung sind $\lambda_1 = 1349.42$ und $\lambda_2 = 67.58$. Der Eigenvektor $\mathbf{a}_1$ zum Eigenwert $\lambda_1 = 1349.42$ muss folgendes Gleichungssystem erfüllen:

$$\begin{aligned} -1003.72\,a_{11} + 528.35\,a_{12} &= 0, \\ 528.35\,a_{11} - 278.12\,a_{12} &= 0. \end{aligned}$$

Hieraus folgt $a_{12} = 1.8997\,a_{11}$.

Wegen $a_{11}^2 + a_{12}^2 = 1$ ist der Eigenvektor $\mathbf{a}_1$ zum Eigenwert λ_1 also

$$\mathbf{a}_1 = \begin{pmatrix} 0.466 \\ 0.885 \end{pmatrix}.$$

Der Eigenvektor $\mathbf{a}_2$ zum Eigenwert $\lambda_2 = 2.91$ muss folgendes Gleichungssystem erfüllen:

$$\begin{aligned} 278.12\, a_{21} + 528.35\, a_{22} &= 0, \\ 528.35\, a_{21} + 1003.72\, a_{22} &= 0. \end{aligned}$$

Hieraus folgt $a_{22} = -0.5264\, a_{21}$. Der Eigenvektor $\mathbf{a}_2$ zum Eigenwert λ_2 ist also

$$\mathbf{a}_2 = \begin{pmatrix} 0.885 \\ -0.466 \end{pmatrix}.$$

Die erste Hauptkomponente stellt eine Art Mittelwert der beiden Merkmale dar, wobei das Merkmal `Mathematische Grundbildung` stärker gewichtet wird. Die zweite Hauptkomponente ist ein Kontrast aus den beiden Merkmalen. Abbildung 5.2 erleichtert die Interpretation der Eigenvektoren.

□

In der Einleitung zu diesem Kapitel hatten wir zwei Ziele formuliert. Zum einen wollten wir die Objekte hinsichtlich aller Merkmale ordnen. Dies ist problemlos möglich, wenn wir die Werte jedes Objekts hinsichtlich einer Linearkombination bestimmen. Es bietet sich die Linearkombination mit der größten Varianz an, da bei dieser die Unterschiede zwischen den Objekten am besten zu Tage treten. hmcounterend. (fortgesetzt)

Example 21. In Tabelle 5.5 findet man die Werte der Linearkombination $0.466\, \tilde{X}_1 + 0.885\, \tilde{X}_2$. Wir sehen, dass Schweden am besten und Griechenland am schlechtesten abschneidet.

Table 5.5. Werte von $0.466\, \tilde{X}_1 + 0.885\, \tilde{X}_2$ von 5 Ländern im Rahmen der PISA-Studie

Land	Wert der Linearkombination
DK	37.1
GR	-32.9
I	-18.0
P	-28.6
S	42.4

□

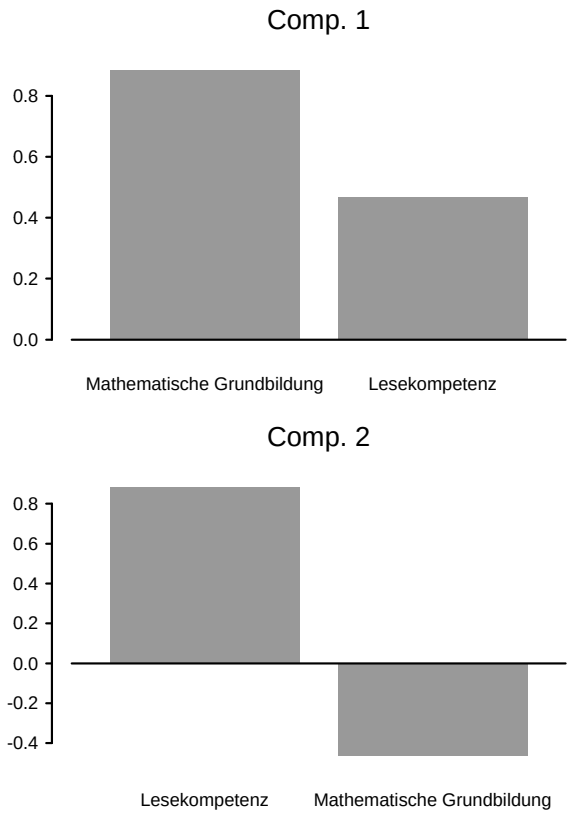


Fig. 5.2. Hauptkomponenten der Merkmale Lesekompetenz und Mathematische Grundbildung von 5 Ländern im Rahmen der PISA-Studie

Das zweite Ziel, das wir am Anfang dieses Kapitels formulierten, war, die Objekte in einem Streudiagramm darzustellen, wobei bei jeder Achse alle Merkmale berücksichtigt werden sollten. Auch dieses Ziel haben wir erreicht. Als erste Achse wählen wir die erste Hauptkomponente und als zweite Achse die zweite Hauptkomponente. Dann zeichnen wir die Werte der Objekte bezüglich dieser Hauptkomponenten ein, also $\mathbf{z}_1 = \tilde{\mathbf{X}}\mathbf{a}_1$ und $\mathbf{z}_2 = \tilde{\mathbf{X}}\mathbf{a}_2$. Dabei ist $\tilde{\mathbf{X}}$ die Matrix der zentrierten Merkmale. Man nennt $\mathbf{z}_1$ und $\mathbf{z}_2$ *Scores*. hmcounterend. (fortgesetzt)

Example 21. Da nur zwei Merkmale betrachtet wurden, erhalten wir eine Wiedergabe des ursprünglichen Streudiagramms in einem gedrehten Koordinatensystem. Abbildung 5.3 zeigt dies.

□

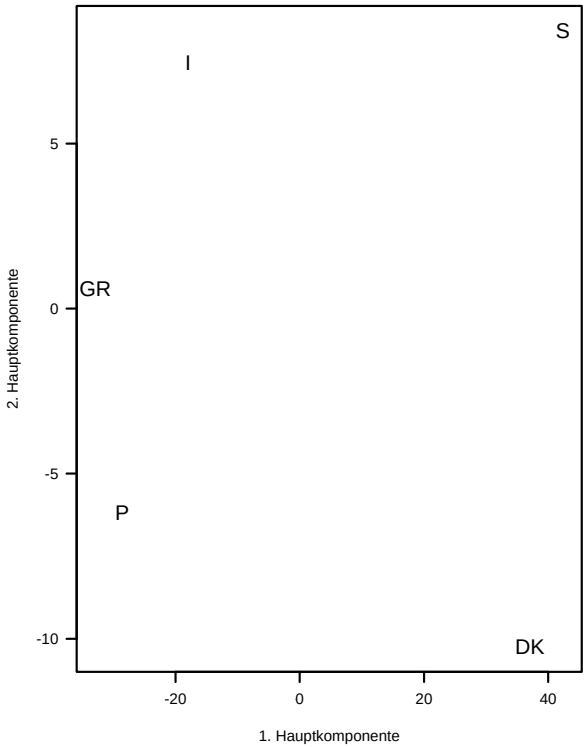


Fig. 5.3. Darstellung der Länder bezüglich der ersten beiden Hauptkomponenten

5.4 Praktische Aspekte

Wir haben im letzten Abschnitt gesehen, wie man auf Basis einer Datenmatrix eine Hauptkomponentenanalyse durchführt. In diesem Abschnitt werden wir einige praxisrelevante Aspekte betrachten. Diese werden wir im Wesentlichen anhand eines Beispiels illustrieren.

Example 22. Im Beispiel 4 auf Seite 5 betrachten wir die Noten der Studenten in den Bereichen `Mathematik`, `BWL`, `VWL` und `Methoden`. Wir führen für diese Daten eine Hauptkomponentenanalyse durch. Die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ ist gegeben durch

$$\mathbf{S} = \begin{pmatrix} 0.6382 & 0.3069 & 0.4695 & 0.3761 \\ 0.3069 & 0.4825 & 0.2889 & 0.1388 \\ 0.4695 & 0.2889 & 0.4944 & 0.3198 \\ 0.3761 & 0.1388 & 0.3198 & 0.3966 \end{pmatrix}.$$

Die Eigenwerte sind

$$\lambda_1 = 1.497, \quad \lambda_2 = 0.323, \quad \lambda_3 = 0.103, \quad \lambda_4 = 0.088. \tag{5.15}$$

Die Eigenvektoren lauten

$$\mathbf{a}_1 = \begin{pmatrix} 0.617 \\ 0.397 \\ 0.536 \\ 0.417 \end{pmatrix}, \mathbf{a}_2 = \begin{pmatrix} 0.177 \\ -0.855 \\ 0.051 \\ 0.485 \end{pmatrix}, \mathbf{a}_3 = \begin{pmatrix} 0.602 \\ -0.289 \\ 0.095 \\ -0.738 \end{pmatrix}, \mathbf{a}_4 = \begin{pmatrix} 0.474 \\ 0.169 \\ -0.837 \\ 0.214 \end{pmatrix}.$$

Abbildung 5.4 erleichtert die Interpretation der Hauptkomponenten.

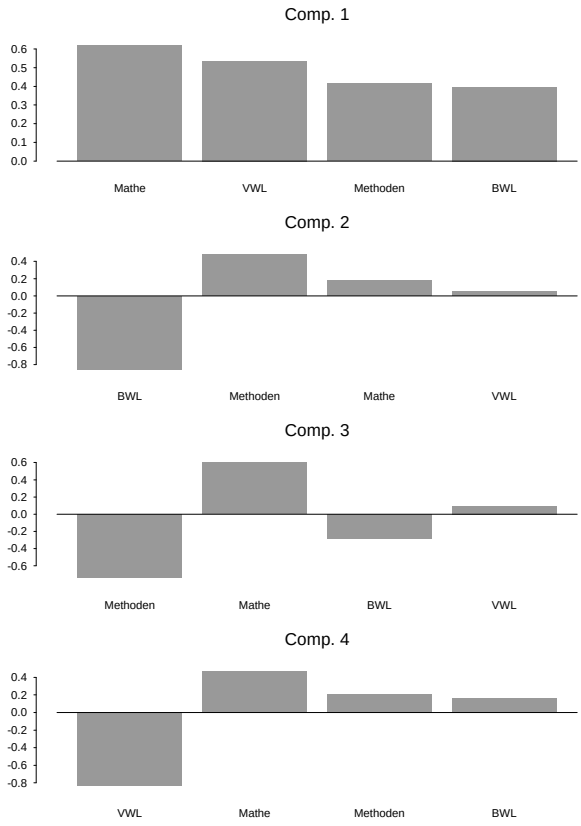


Fig. 5.4. Hauptkomponenten bei 17 Studenten auf der Basis der Noten in 4 Fachgebieten

Die erste Hauptkomponente stellt eine Art Mittelwert der vier Merkmale dar, wobei das Merkmal `Mathematik` am stärksten gewichtet wird. Die zweite

Hauptkomponente ist ein Kontrast aus der Note in `Methoden` und `BWL`. Ein Student hat hier einen großen Wert, wenn er einen großen Wert in `Methoden` und einen kleinen Wert in `BWL` hat. Hier wird also unterschieden zwischen Studenten, die gut in `BWL` und schlecht in `Methoden` und solchen, die gut in `Methoden` und schlecht in `BWL` sind. Wir sehen, dass die dritte Komponente ein Kontrast aus `Methoden` und `Mathematik` ist, während die vierte Hauptkomponente ein Kontrast aus `VWL` und `Mathematik` ist. Abbildung 5.5 zeigt die Werte der 17 Studenten bezüglich der ersten beiden Hauptkomponenten, wobei wir auch hier wieder von den zentrierten Merkmalen ausgehen.

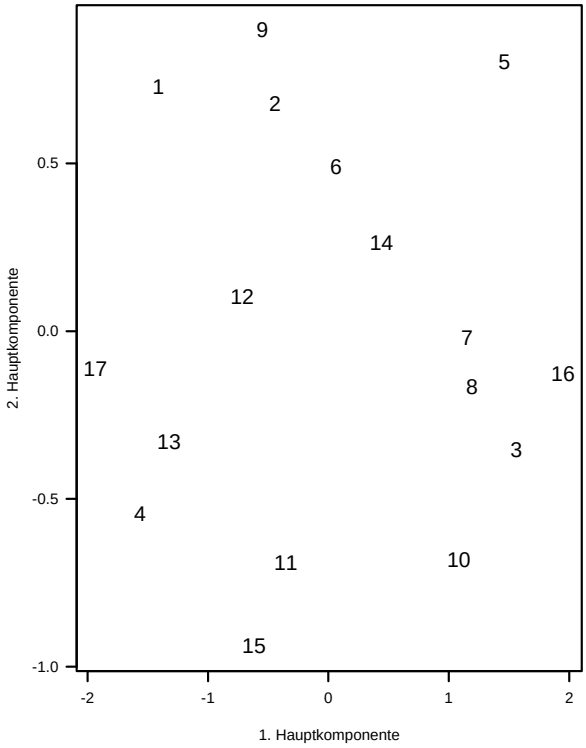


Fig. 5.5. Werte von 17 Studenten bezüglich der ersten beiden Hauptkomponenten

□

Es sind nun noch einige Fragen offen:

1. Wie viele Hauptkomponenten benötigt man?
2. Wie kann man beurteilen, ob die Darstellung im $\mathbb{R}^2$ gut ist?

3. Soll die Analyse auf Basis der Varianz-Kovarianz-Matrix oder auf Basis der Korrelationsmatrix durchgeführt werden?

Diese Fragen werden wir in den nächsten Abschnitten beantworten.

5.4.1 Anzahl der Hauptkomponenten

In diesem Abschnitt wollen wir uns mit der Frage beschäftigen, wie viele Hauptkomponenten ausreichen, um die wesentliche Struktur des Datensatzes zu reproduzieren. Um dies entscheiden zu können, schauen wir uns die Eigenwerte der empirischen Varianz-Kovarianz-Matrix $\mathbf{S}$ an. Wir gehen dabei davon aus, dass die Eigenwerte $\lambda_1, \ldots, \lambda_p$ der Größe nach geordnet sind, wobei λ_1 der größte Eigenwert ist. hmcounterend. (fortgesetzt)

Example 22. Die Eigenwerte sind

$$\lambda_1 = 1.497, \quad \lambda_2 = 0.323, \quad \lambda_3 = 0.103, \quad \lambda_4 = 0.088.$$

□

Die Spur von $\mathbf{S}$ ist gleich der Summe der Stichprobenvarianzen:

$$tr(\mathbf{S}) = \sum_{i=1}^{p} s_i^2.$$

Auf Seite 480 wird gezeigt, dass die Spur einer symmetrischen Matrix gleich der Summe ihrer Eigenwerte ist. Da die Eigenwerte der empirischen Varianz-Kovarianz-Matrix $\mathbf{S}$ aber gerade die Stichprobenvarianzen der Hauptkomponenten sind, können wir bestimmen, welcher Anteil der Gesamtstreuung durch die einzelnen Hauptkomponenten erklärt wird. Um sich für eine Anzahl von Hauptkomponenten zu entscheiden, gibt man einen Anteil α vor und wählt als Anzahl den kleinsten Wert von r, für den gilt

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq \alpha.$$

Typische Werte für α sind 0.75, 0.8 und 0.85. hmcounterend. (fortgesetzt)

Example 22. Die erste Hauptkomponente erklärt 74.4 Prozent der Gesamtstreuung. Die ersten beiden Hauptkomponenten erklären zusammen 90.5 Prozent der Gesamtstreuung. Es sollten also zwei Komponenten zur Beschreibung des Datensatzes herangezogen werden. □

Cattell (1966) hat vorgeschlagen, die λ_i gegen den Index i zu zeichnen. In der Regel ist in dieser Graphik ein Knick zu beobachten. Es werden nur die Hauptkomponenten betrachtet, die zu Eigenwerten gehören, die vor dem Knick liegen. Man spricht in diesem Fall vom *Screeplot* (scree heißt Geröllhalde). hmcounterend. (fortgesetzt)

Example 22. Abbildung 5.6 zeigt den Screeplot.

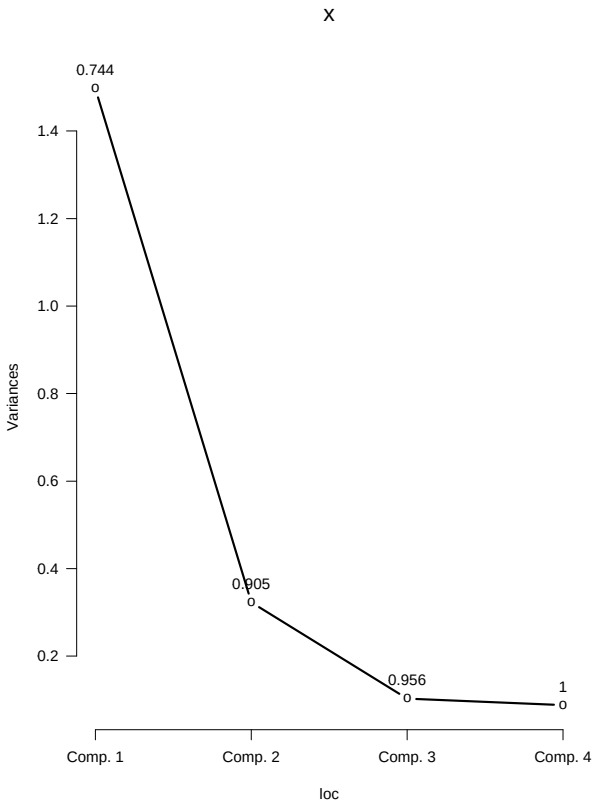


Fig. 5.6. Screeplot der Eigenwerte der Noten der 17 Studenten

Hier ist es nicht einfach, eine Entscheidung zu fällen. Sowohl eine als auch zwei Hauptkomponenten erscheinen akzeptabel. □

Ein anderer Zugang wurde von Kaiser (1960) gewählt. Dieser hat vorgeschlagen, nur die Hauptkomponenten zu berücksichtigen, deren zugehörige Eigenwerte größer sind als der Mittelwert aller Eigenwerte, d.h. alle λ_i mit $\lambda_i > \bar{\lambda}$, wobei gilt

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

hmcounterend. (fortgesetzt)

Example 22. Der Mittelwert der Eigenwerte ist 0.503. Nur der erste Eigenwert ist größer als dieser Mittelwert. Nach dem *Kriterium von Kaiser* benötigen wir also nur eine Hauptkomponente. □

Kaisers Kriterium beruht auf der Hauptkomponentenanalyse der standardisierten Merkmale. In diesem Fall ist die Stichprobenvarianz jedes Merkmals gleich 1. Der Mittelwert aller Stichprobenvarianzen ist somit auch gleich 1. Jede Hauptkomponente, deren zugehöriger Eigenwert größer als 1 ist, trägt zur Gesamtstreuung mehr bei als jedes einzelne Merkmal. In diesem Sinne ist sie wichtig und sollte berücksichtigt werden. Dieses Konzept kann problemlos auf eine Hauptkomponente der nichtstandardisierten Merkmale übertragen werden. Dies führt dann zu Kaisers Kriterium. Jolliffe (1972) hat festgestellt, dass die Forderung Kaisers zu hoch ist. Er hat vorgeschlagen, nur die Hauptkomponenten zu berücksichtigen, deren zugehörige Eigenwerte größer sind als das 0.7-fache des Mittelwerts aller Eigenwerte, d.h. alle λ_i mit $\lambda_i > 0.7\,\bar{\lambda}$, wobei gilt

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

hmcounterend. (fortgesetzt)

Example 22. Es gilt $0.7\,\bar{\lambda} = 0.352$. Auch nach dem *Kriterium von Jolliffe* würde man nur eine Hauptkomponente verwenden. □

5.4.2 Überprüfung der Güte der Anpassung

Im Folgenden wollen wir überprüfen, wie gut die durch eine Hauptkomponentenanalyse gewonnene zweidimensionale Darstellung ist. Hierzu benötigen wir einige Begriffe aus der Graphentheorie. Ausgangspunkt sei eine *Menge von Punkten*.

Example 23. Wir betrachten die folgenden Punkte:

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad \mathbf{x}_4 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad \mathbf{x}_5 = \begin{pmatrix} 7 \\ 2 \end{pmatrix}.$$

□

Eine Verbindung von zwei Punkten nennt man *Kante*. Eine Menge von Kanten und Punkten heißt *Graph*. hmcounterend. (fortgesetzt)

Example 23. Abbildung 5.7 zeigt einen Graphen.

□

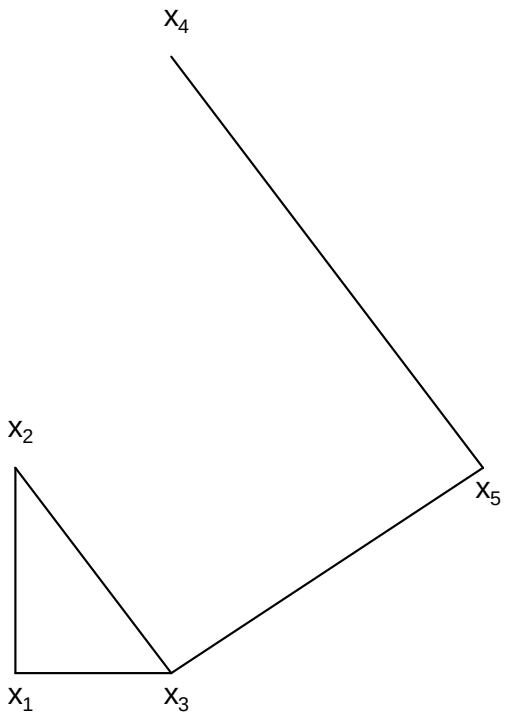


Fig. 5.7. Ein Beispiel eines Graphen

Ein Graph heißt *zusammenhängend*, wenn man von jedem Punkt jeden anderen Punkt erreichen kann, indem man eine Folge von Kanten durchläuft. hmcounterend. (fortgesetzt)

Example 23. Der Graph in Abbildung 5.7 ist zusammenhängend. □

Ein *Kreis* ist eine Folge von unterscheidbaren Kanten, die im gleichen Punkt beginnt und endet, ohne eine Kante mehrfach zu durchlaufen. hmcounterend. (fortgesetzt)

Example 23. In Abbildung 5.6 bilden die Kanten, die die Punkte $\mathbf{x}_1$, $\mathbf{x}_2$ und $\mathbf{x}_3$ verbinden, einen Kreis. □

Ein zusammenhängender Graph ohne Kreis heißt *spannender Baum*. hmcounterend. (fortgesetzt)

Example 23. Abbildung 5.8 zeigt einen spannenden Baum.

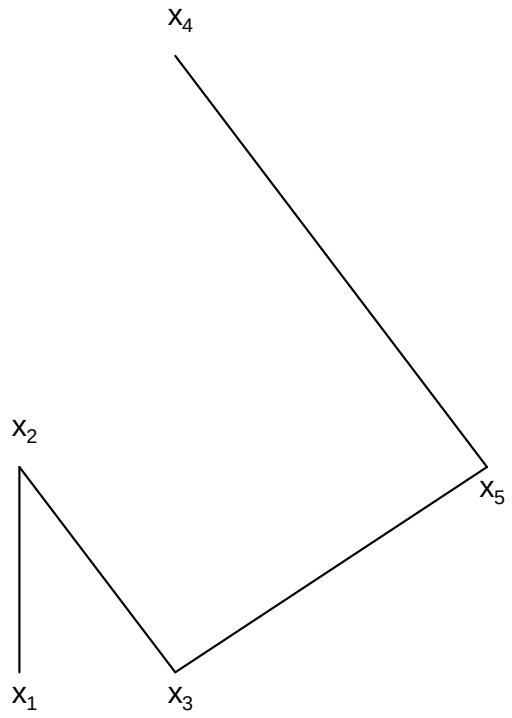


Fig. 5.8. Ein Beispiel eines spannenden Baumes

□

Bei vielen Anwendungen werden die Kanten bewertet. hmcounterend. (fortgesetzt)

Example 23. Wir bestimmen die euklidischen Distanzen zwischen den Punkten. Diese sind in der folgenden Distanzmatrix zu finden:

$$\mathbf{D} = \begin{pmatrix} 0 & 1.0 & 2.0 & 3.6 & 6.1 \\ 1.0 & 0 & 2.2 & 2.8 & 6.0 \\ 2.0 & 2.2 & 0 & 3.0 & 4.1 \\ 3.6 & 2.8 & 3.0 & 0 & 4.5 \\ 6.1 & 6.0 & 4.1 & 4.5 & 0 \end{pmatrix}.$$

□

Der spannende Baum mit kleinster Summe der Kantengewichte heißt *minimal spannender Baum*. Ein Algorithmus zur Bestimmung des minimal spannenden Baumes wurde von Kruskal (1956) vorgeschlagen. Bei diesem Algorithmus wird zur Konstruktion der Reihe nach die kürzeste Kante ausgewählt und dem Baum hinzugefügt, wenn durch ihre Hinzunahme kein Kreis entsteht. Der Algorithmus endet, wenn ein spannender Baum gefunden wurde. Dies ist dann auch der minimal spannende Baum. hmcounterend. (fortgesetzt)

Example 23. Um den minimal spannenden Baum zu finden, suchen wir die kleinste Distanz aus. Diese ist 1. Somit werden die Punkte $\mathbf{x}_1$ und $\mathbf{x}_2$ mit einer Kante verbunden. Die kleinste Distanz unter den restlichen Punkten ist 2. Es werden also die Punkte $\mathbf{x}_1$ und $\mathbf{x}_3$ mit einer Kante verbunden. Die kleinste Distanz unter den restlichen Punkten ist 2.2. Dies ist der Abstand der Punkte $\mathbf{x}_2$ und $\mathbf{x}_3$. Die Punkte $\mathbf{x}_2$ und $\mathbf{x}_3$ werden aber nicht durch eine Kante verbunden, da hierdurch ein Kreis entsteht. Die kleinste Distanz unter den restlichen Punkten ist 2.8. Es werden also die Punkte $\mathbf{x}_2$ und $\mathbf{x}_4$ mit einer Kante verbunden. Die kleinste Distanz unter den restlichen Punkten ist 3. Dies ist der Abstand der Punkte $\mathbf{x}_3$ und $\mathbf{x}_4$. Die Punkte $\mathbf{x}_3$ und $\mathbf{x}_4$ werden aber nicht durch eine Kante verbunden, da hierdurch ein Kreis entsteht. Die kleinste Distanz unter den restlichen Punkten ist 3.6. Dies ist der Abstand der Punkte $\mathbf{x}_1$ und $\mathbf{x}_4$. Die Punkte 1 und 4 werden aber nicht durch eine Kante verbunden, da hierdurch ein Kreis entsteht. Die kleinste Distanz unter den restlichen Punkten ist 4.1. Es werden also die Punkte $\mathbf{x}_3$ und $\mathbf{x}_5$ mit einer Kante verbunden. Abbildung 5.9 zeigt den minimal spannenden Baum.

□

Man kann einen spannenden Baum verwenden, um die Güte einer zweidimensionalen Darstellung zu überprüfen. Hierzu führt man eine Hauptkomponentenanalyse durch und zeichnet die Punkte bezüglich der ersten beiden Hauptkomponenten in ein Koordinatensystem. Anschließend bestimmt man den minimal spannenden Baum für die Originaldaten und zeichnet diesen Baum in das Streudiagramm. hmcounterend. (fortgesetzt)

Example 22. Für den Datensatz der Noten erhalten wir die Abbildung 5.10. Wir sehen, dass die zweidimensionale Darstellung die Lage der Studenten im vierdimensionalen Raum gut wiedergibt. Nur die relative Lage der Studenten 4, 13 und 17 wird im zweidimensionalen Raum nicht richtig wiedergegeben.

□

5.4.3 Analyse auf Basis der Varianz-Kovarianz-Matrix oder auf Basis der Korrelationsmatrix

Bei einer Hauptkomponentenanalyse führt man eine Spektralzerlegung der empirischen Varianz-Kovarianz-Matrix durch. Diese kann man, wie wir in

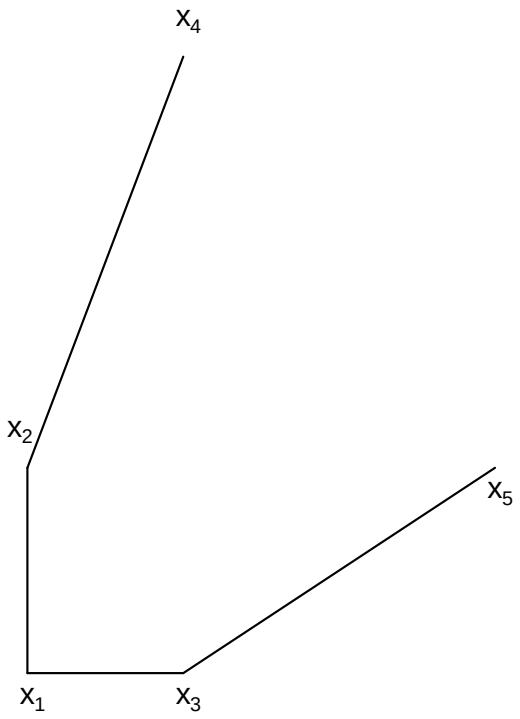


Fig. 5.9. Ein Beispiel eines minimal spannenden Baumes

(2.17) gesehen haben, folgendermaßen aus der Matrix $\tilde{\mathbf{X}}$ der zentrierten Merkmale gewinnen:

$$\mathbf{S} = \frac{1}{n-1}\,\tilde{\mathbf{X}}'\tilde{\mathbf{X}}.$$

Geht man hingegen von der Matrix $\mathbf{X}^*$ der standardisierten Merkmale aus, so erhält man über die gleiche Vorgehensweise die empirische Korrelationsmatrix $\mathbf{R}$:

$$\mathbf{R} = \frac{1}{n-1}\,\mathbf{X}^{*\prime}\mathbf{X}^*.$$

Dies wird auf Seite 39 gezeigt. Eine Spektralzerlegung der Korrelationsmatrix liefert also Hauptkomponenten für die standardisierten Merkmale. Welche der beiden Vorgehensweisen ist vorzuziehen? Schauen wir uns dies anhand eines Beispiels an.

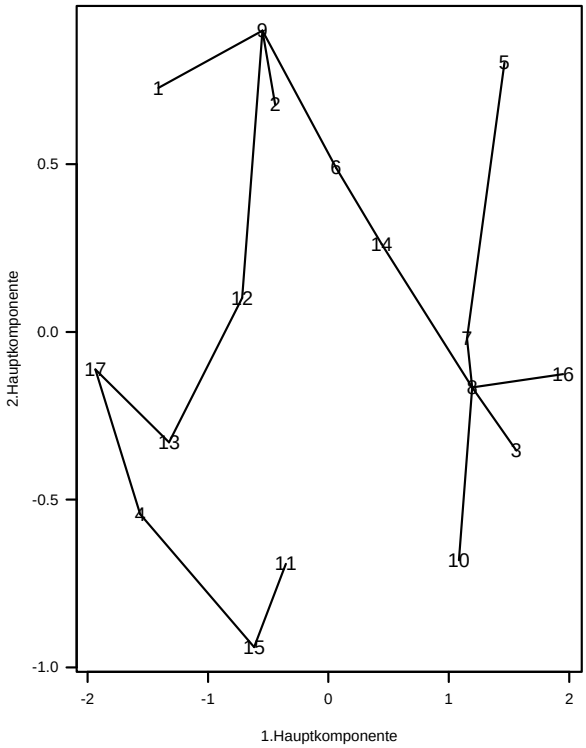


Fig. 5.10. Darstellung der Noten der 17 Studenten bezüglich der ersten beiden Hauptkomponenten mit dem minimal spannenden Baum der Originaldaten

Example 24. Wir betrachten das Beispiel 12 auf Seite 11. Die empirische Varianz-Kovarianz-Matrix lautet:

$$\mathbf{S} = \begin{pmatrix} 637438 & 120648 & -9904 & -6093 & -738 & 162259 \\ 120648 & 29988 & -3532 & -1992 & -591 & 31184 \\ -9904 & -3532 & 1017 & 422 & 140 & -4445 \\ -6093 & -1992 & 422 & 315 & 95 & -4219 \\ -738 & -591 & 140 & 95 & 37 & -847 \\ 162259 & 31184 & -4445 & -4219 & -847 & 96473 \end{pmatrix}.$$

Schauen wir uns die Varianzen an:

$$s_1^2 = 637438,\ s_2^2 = 29988,\ s_3^2 = 1017,\ s_4^2 = 315,\ s_5^2 = 37,\ s_6^2 = 96473.$$

Wir sehen, dass sich die Varianzen stark unterscheiden. Dies führt dazu, dass die Hauptkomponenten von den Merkmalen dominiert werden, die große Varianzen besitzen. Abbildung 5.11 zeigt das. Führt man die Analyse auf Basis

der empirischen Korrelationsmatrix durch, so erhält man Abbildung 5.12. Die Hauptkomponenten auf Basis der empirischen Varianz-Kovarianz-Matrix unterscheiden sich stark von den Hauptkomponenten auf Basis der empirischen Korrelationsmatrix. In Abbildung 5.11 wird alle Struktur in den Hauptkomponenten durch die große Varianz des ersten Merkmals überdeckt. Eine sinnvolle Interpretation ist nur in Abbildung 5.12 möglich. Unterscheiden sich die Varianzen der Merkmale, so sollte man eine Hauptkomponentenanalyse auf Basis der Korrelationsmatrix durchführen. □

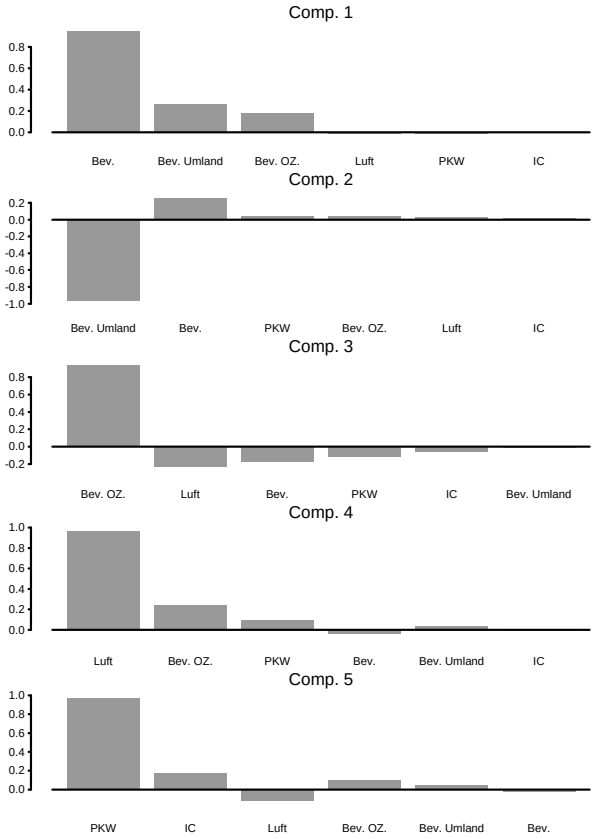


Fig. 5.11. Hauptkomponenten auf Basis der empirischen Varianz-Kovarianz-Matrix

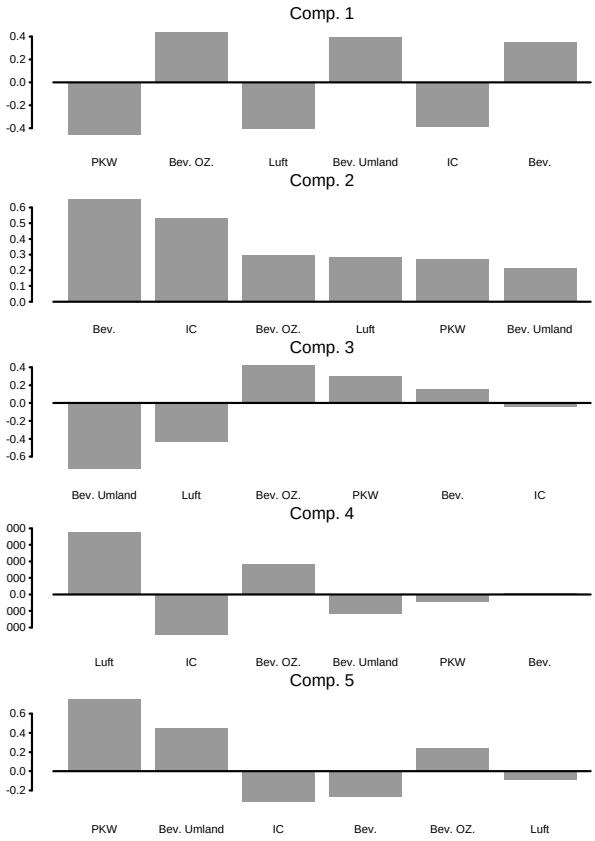


Fig. 5.12. Hauptkomponenten auf Basis der empirischen Korrelationsmatrix

5.5 Hauptkomponentenanalyse der Ergebnisse der PISA-Studie

Im Beispiel 1 auf Seite 3 wurden in 31 Ländern die Fähigkeiten fünfzehnjähriger Schüler in den Bereichen `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` ermittelt. Wir wollen in diesem Abschnitt mit Hilfe der Hauptkomponentenanalyse die Gesamtleistung in den drei Bereichen bestimmen. Außerdem wollen wir überprüfen, ob die Daten durch eine Dimension ausreichend beschrieben werden können. Zunächst stellt sich die Frage, ob wir eine Hauptkomponentenanalyse auf Basis der Originaldaten oder der standardisierten Daten durchführen sollen. Die empirische Varianz-Kovarianz-Matrix der Daten lautet

$$\mathbf{S} = \begin{pmatrix} 1109.4 & 1428.3 & 1195.6 \\ 1428.3 & 2192.9 & 1644.0 \\ 1195.6 & 1644.0 & 1419.0 \end{pmatrix}.$$

Keine der Varianzen dominiert die anderen, sodass wir eine Hauptkomponentenanalyse auf Basis der empirischen Varianz-Kovarianz-Matrix durchführen. Schauen wir uns zuerst in Abbildung 5.13 die Hauptkomponenten an.

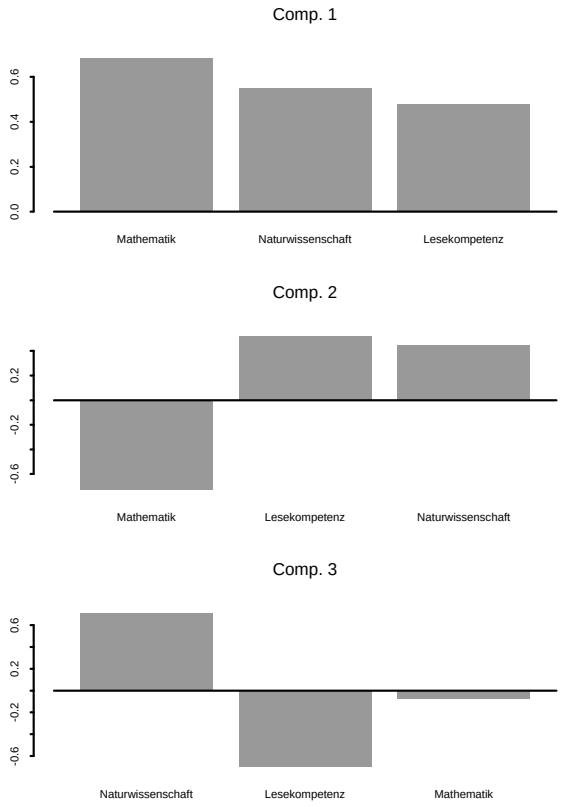


Fig. 5.13. Hauptkomponenten bei allen Ländern der PISA-Studie

Die erste Hauptkomponente ist eine Art Mittelwert der Merkmale, wobei das Merkmal `Mathematische Grundbildung` das stärkste Gewicht erhält. Die zweite Hauptkomponente ist ein Kontrast aus dem Merkmal `Mathematische Grundbildung` und den beiden anderen Merkmalen, während die dritte Hauptkomponente ein Kontrast aus dem Merkmal `Naturwissenschaftliche Grundbildung` und dem Merkmal `Lesekompetenz` ist. Die Eigenwerte sind

$$\lambda_1 = 4513.53, \quad \lambda_2 = 150.03, \quad \lambda_3 = 57.673.$$

Die Eigenwerte deuten darauf hin, dass eine Hauptkomponente zur Beschreibung der Daten ausreicht. Dies bestätigen auch der Screeplot in Abbil-

dung 5.14 und das Kaiser-Kriterium, da der Mittelwert der Eigenwerte gleich 1573.747 ist.

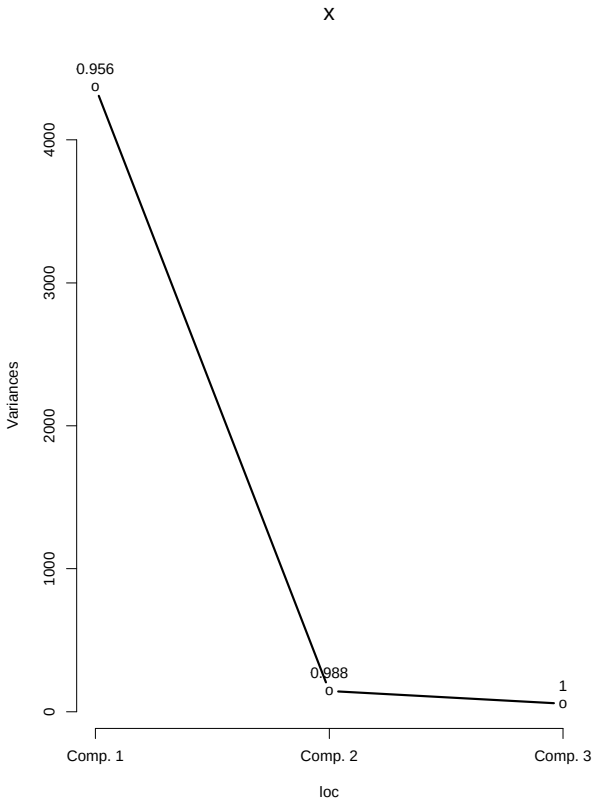


Fig. 5.14. Screeplot bei allen Ländern der PISA-Studie

Tabelle 5.6 gibt die Scores der einzelnen Länder wieder, wobei bei der Berechnung die zentrierten Merkmale berücksichtigt wurden.

Table 5.6. Scores der zentrierten Merkmale bei einer Hauptkomponentenanalyse der Merkmale Lesekompetenz, Mathematische Grundbildung und Naturwissenschaftliche Grundbildung in 31 Ländern im Rahmen der PISA-Studie

Land	Score	Land	Score	Land	Score
Japan	88.9	Belgien	26.6	Polen	-28.1
Korea	84.6	Frankreich	25.9	Italien	-35.9
Finnland	79.5	Island	2.6	Russland	-43.4
Kanada	66.7	Norwegen	13.6	Lettland	-55.6
Neuseeland	66.5	Tschechien	12.7	Portugal	-56.5
Australien	63.3	Dänemark	9.6	Griechenland	-58.3
Großbritannien	60.3	USA	8.5	Luxemburg	-84.7
Österreich	35.9	Liechtenstein	0.1	Mexiko	-145.7
Irland	34.0	Ungarn	-8.1	Brasilien	-220.3
Schweden	33.0	Deutschland	-9.8		
Schweiz	26.7	Spanien	-12.8		

5.6 Hauptkomponentenanalyse in S-PLUS

Wir wollen nun das Beispiel 22 auf Seite 130 in `S-PLUS` nachvollziehen. Hierzu verwenden wir die Daten aus Beispiel 4 auf Seite 5. Diese mögen in der Matrix `note` stehen. Schauen wir uns diese an:

```
> note
    Mathe  BWL   VWL Methoden
 1 1.325 1.00 1.825     1.75
 2 2.000 1.25 2.675     1.75
 3 3.000 3.25 3.000     2.75
 4 1.075 2.00 1.675     1.00
 5 3.425 2.00 3.250     2.75
 6 1.900 2.00 2.400     2.75
 7 3.325 2.50 3.000     2.00
 8 3.000 2.75 3.075     2.25
 9 2.075 1.25 2.000     2.25
10 2.500 3.25 3.075     2.25
11 1.675 2.50 2.675     1.25
12 2.075 1.75 1.900     1.50
13 1.750 2.00 1.150     1.25
14 2.500 2.25 2.425     2.50
15 1.675 2.75 2.000     1.25
16 3.675 3.00 3.325     2.50
17 1.250 1.50 1.150     1.00
```

Mit der Funktion `princomp` kann man in `S-PLUS` eine Hauptkomponentenanalyse durchführen.

Der Aufruf von `princomp` ist

```
princomp(x, data=NULL, covlist=NULL, weights=NULL,
         scores=T,cor=F,na.action=na.fail, subset=T)
```

Betrachten wir die Argumente, die sich auf Charakteristika beziehen, mit denen wir uns beschäftigt haben. Liegen die Daten in Form einer Datenmatrix vor, so weisen wir diese beim Aufruf dem Argument `x` zu. Soll die Hauptkomponentenanalyse auf Basis der standardisierten Merkmale durchgeführt werden, so setzt man das Argument `cor` auf den Wert `T`. Standardmäßig steht dieser auf `F`, sodass die Originaldaten verwendet werden. Das Argument `covlist` bietet die Möglichkeit, die Daten in Form der empirischen Varianz-Kovarianz-Matrix oder der empirischen Korrelationsmatrix zu übergeben. Standardmäßig werden die Scores berechnet. Dies sieht man am Argument `scores`, das auf `T` gesetzt ist.

Das Ergebnis der Funktion `princomp` ist eine Liste. Schauen wir uns die relevanten Komponenten am Beispiel an. Wir rufen die Funktion `princomp` mit dem Argument `note` auf und weisen das Ergebnis der Variablen `e` zu:

```
> e<-princomp(note)
```

Der Aufruf

```
> e$sdev
```

liefert die Wurzeln der Eigenwerte:

```
   Comp. 1   Comp. 2   Comp. 3   Comp. 4
 1.187007 0.5512975 0.3118719 0.2883909
```

Um das Kriterium von Kaiser anwenden zu können, benötigen wir die Eigenwerte. Wir bilden also

```
> eig<-e$sdev^2
```

und erhalten als Ergebnis

```
> eig
Comp. 1   Comp. 2    Comp. 3    Comp. 4
 1.408985 0.3039289 0.09726408 0.08316932
```

Diese Werte stimmen nicht mit den Werten in Gleichung 5.15 auf Seite 131 überein. Dies liegt daran, dass `S-PLUS` die Eigenwerte und Eigenvektoren der Matrix $(n-1)/n\,\mathbf{S}$ und nicht der Matrix $\mathbf{S}$ bestimmt. Dies führt zu keinen Problemen bei der Analyse, da die Eigenvektoren der Matrizen $\mathbf{S}$ und $(n-1)/n\,\mathbf{S}$ identisch sind. Außerdem ist $n/(n-1)\,\lambda$ Eigenwert von $\mathbf{S}$, wenn λ Eigenwert von $(n-1)/n\,\mathbf{S}$ ist. Da die Entscheidungen über die Anzahl der Hauptkomponenten nur von den Verhältnissen der Eigenwerte abhängen, können wir die Eigenwerte und Eigenvektoren der Matrix $\mathbf{S}$ oder der Matrix $(n-1)/n\,\mathbf{S}$ betrachten. Um die gleichen Zahlen und Graphiken wie im Text zu erhalten, bilden wir

```
> e$sdev<-sqrt(17/16)*e$sdev
> eig<-e$sdev^2
```

Die Eigenwerte sind

```
> eig
  Comp. 1   Comp. 2   Comp. 3   Comp. 4
 1.497047 0.3229245 0.1033431 0.0883674
```

Der Mittelwert der Eigenwerte ist

```
> mean(eig)
[1] 0.5029205
```

Nach dem Kriterium von Kaiser benötigen wir zur Beschreibung der Daten eine Hauptkomponente. Um das Kriterium von Jolliffe anwenden zu können, multiplizieren wir den Mittelwert der Eigenwerte mit 0.7:

```
> 0.7*mean(eig)
[1] 0.3520443
```

Auch bei diesem Kriterium benötigen wir zur Beschreibung der Daten nur eine Hauptkomponente. Die Abbildung 5.6 des Screeplots auf Seite 134 erhält man mit der Funktion `screeplot` folgendermaßen:

```
> screeplot(e,style="l")
```

Die Anteile der Gesamtstreuung, die durch die einzelnen Hauptkomponenten erklärt werden, und den kumulierten Anteil der Gesamtstreuung erhält man durch

```
> summary(e)
Importance of components:
                          Comp.1   Comp.2    Comp.3    Comp.4
    Standard deviation 1.223538 0.568264 0.3214701 0.2972665
Proportion of Variance 0.744176 0.160524 0.0513714 0.0439271
 Cumulative Proportion 0.744176 0.904701 0.9560728 1.0000000
>
```

Wir sehen, dass durch die ersten beiden Hauptkomponenten rund 90 Prozent der Gesamtstreuung erklärt werden. Schauen wir uns die Hauptkomponenten an. Diese erhält man durch

```
> loadings(e)
         Comp. 1 Comp. 2 Comp. 3 Comp. 4
   Mathe  0.617   0.177   0.602   0.475
     BWL  0.397  -0.855  -0.289   0.169
     VWL  0.536                  -0.837
Methoden  0.417   0.485  -0.738   0.213
```

Die leeren Positionen der Matrix charakterisieren Werte, die in der Nähe von Null liegen. Die Abbildung 5.4 der Hauptkomponenten auf Seite 131 liefert der folgende Befehl:

```
> plot(loadings(e))
```

Die Scores der Studenten liefert

```
> e$scores
```

Das Streudiagramm der Scores der Studenten bezüglich der ersten beiden Hauptkomponenten in Abbildung 5.5 erhält man durch

```
> plot(e$scores[,1:2],xlab="1.Hauptkomponente",
     ylab="2.Hauptkomponente",type="n")
> text(e$scores[,1:2],1:17)
```

Einen Eindruck, inwieweit die zweidimensionale Darstellung die ursprüngliche Lage der Punkte gut wiedergibt, erhalten wir dadurch, dass wir den minimal spannenden Baum der ursprünglichen Daten in die Darstellung der beiden Hauptkomponenten legen. Die Funktion `mstree` liefert hierzu die nötigen Informationen. Der Aufruf von `mstree` ist

```
> mstree(x, plane=T)
```

Dabei stehen in den Zeilen der Matrix `x` die Koordinaten der Punkte, für die der minimal spannende Baum bestimmt werden soll. Das Argument `plane` ist hier nicht relevant. Wir müssen es auf `F` setzen. Das Ergebnis der Funktion `mstree` ist ein Vektor. Die i-te Komponente dieses Vektors gibt den Index des Punktes an, mit dem der i-te Punkt verbunden werden soll. Im Beispiel erhalten wir

```
> mst<-mstree(note,plane=F)
> mst
 [1]  9  9  8 17  7  9  8 14 12  8 15 13 17  6  4  8
```

Wir sehen, dass der erste Punkt mit dem neunten Punkt verbunden werden soll, und dass `mst` nur aus 16 Komponenten besteht. Da es sich um einen spannenden Baum handelt, führt zur 17-ten Beobachtung auf jeden Fall von einem der anderen Punkte eine Gerade. Um die Abbildung 5.10 zu erhalten, erstellen wir zunächst das Streudiagramm der Scores:

```
> x<-e$scores[,1]
> y<-e$scores[,2]
> plot(x,y,xlab="1.Hauptkomponente",
     ylab="2.Hauptkomponente",type="n")
> text(x,y,1:17)
```

Dann rufen wir `mst` auf:

```
> mst<-mstree(note,plane=F)
```

Um die Linien zu zeichnen, benötigen wir nun noch die Funktion `segments`. Diese wird aufgerufen durch

```
segments(x1,y1,x2,y2)
```

Dabei sind `x1`, `y1`, `x2` und `y2` Vektoren der Länge n. Durch den Aufruf der Funktion `segments` werden die Punkte (`x1`$[i]$,`y1`$[i]$) und (`x2`$[i]$,`y2`$[i]$) für jeden Wert von $i = 1, \ldots, n$ durch eine Linie verbunden. Wir müssen also nur noch eingeben:

```
> segments(x[1:length(mst)],y[1:length(mst)],x[mst],y[mst])
```

Im Beispiel 8 auf Seite 8 sind nur die Korrelationen gegeben. Wir erzeugen mit diesen die Matrix `rnutzen`:

```
> rnutzen
          Fehler Kunden Angebot Qualitaet  Zeit Kosten
   Fehler  1.000  0.223   0.133     0.625 0.506  0.500
   Kunden  0.223  1.000   0.544     0.365 0.320  0.361
  Angebot  0.133  0.544   1.000     0.248 0.179  0.288
Qualitaet  0.625  0.365   0.248     1.000 0.624  0.630
     Zeit  0.506  0.320   0.179     0.624 1.000  0.625
   Kosten  0.500  0.361   0.288     0.630 0.625  1.000
```

Um eine Hauptkomponentenanalyse auf Basis von `rnutzen` duchführen zu können, erzeugen wir eine Liste mit den Komponenten `cov` und `center`. Die Komponente `cov` enthält die Matrix `rnutzen`. Die Komponente `center` enthält einen Vektor, dessen Länge gleich der Anzahl der Merkmale ist, und der aus Nullen besteht:

```
> cov.obj<-list(cov=rnutzen,center=rep(0,dim(rnutzen)[1]))
```

Nach dem Aufruf

```
> e<-princomp(covlist=cov.obj)
```

stehen in `e` die Informationen über die Hauptkomponentenanalyse. In diesem Fall können natürlich keine Scores bestimmt werden.

5.7 Ergänzungen und weiterführende Literatur

Wir haben in diesem Kapitel eine Einführung in die Hauptkomponentenanalyse gegeben. Dabei haben wir eine Reihe von Aspekten nicht berücksichtigt. Eine Vielzahl weiterer Aspekte sind bei Jackson (1991) und Jolliffe (1986) zu finden. So sind die Ergebnisse der Hauptkomponentenanalyse sehr ausreißerempfindlich. Eine Möglichkeit einer robusten Schätzung der Hauptkomponenten besteht in einer Spektralanalyse einer robusten Schätzung der Varianz-Kovarianz-Matrix. Dieses und weitere Verfahren sind bei Jackson

(1991), Seber (1984) und Jolliffe (1986) beschrieben. Bei der Hauptkomponentenanalyse werden Linearkombinationen mit großer Varianz gesucht. Friedman & Tukey (1974) haben ein Verfahren vorgeschlagen, bei dem andere Kriterien als die Varianz der Linearkombination bei der Suche nach interessanten Projektionen berücksichtigt werden. Sie nennen dieses Verfahren *Projection Pursuit*. Einen Überblick über Projection Pursuit geben Jones & Sibson (1987) und Huber (1985).

5.8 Übungen

Exercise 6. Bestimmen Sie für den Fall $p = 2$ die Eigenwerte und Eigenvektoren der Korrelationsmatrix in Abhängigkeit vom Wert des Korrelationskoeffizienten r. Betrachten Sie die Spezialfälle $r = 0$ und $r = 1$.

Exercise 7. Gegeben sei die Datenmatrix

$$\mathbf{X} = \begin{pmatrix} 60 & 64 \\ 58 & 62 \\ 64 & 68 \\ 56 & 52 \end{pmatrix}.$$

Die Matrix der euklidischen Abstände lautet

$$\mathbf{D} = \begin{pmatrix} 0 & 2.83 & 5.66 & 12.65 \\ 2.83 & 0 & 8.49 & 10.2 \\ 5.66 & 8.49 & 0 & 17.89 \\ 12.65 & 10.2 & 17.89 & 0 \end{pmatrix}.$$

Stellen Sie die Punkte graphisch dar und bestimmen Sie den minimal spannenden Baum.

Exercise 8. In Kapitel 5.4.3 haben wir uns mit dem Beispiel 12 auf Seite 11 unter dem Aspekt beschäftigt, ob man die Hauptkomponentenanalyse auf Basis der Varianz-Kovarianz-Matrix oder der Korrelationsmatrix durchführen soll. Führen Sie eine vollständige Hauptkomponentenanalyse der Daten mit `S-PLUS` durch.

Exercise 9. Vollziehen Sie die Hauptkomponentenanalyse der Ergebnisse der PISA-Studie im Kapitel 5.5 mit `S-PLUS` nach.

Exercise 10. In der PISA-Studie wurden von den einzelnen Ländern nicht nur die Mittelwerte der Punkte in den Bereichen `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung`, sondern auch ausgewählte Quantile angegeben. Tabelle 5.7 enthält die 0.95-Quantile in den drei Bereichen.

Table 5.7. 0.95-Quantil der Punkte in den Bereichen Lesekompetenz, Mathematische Grundbildung und Naturwissenschaftliche Grundbildung im Rahmen der PISA-Studie, vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S. 107, 173, 229

Land	Lesekompetenz	Mathematische Grundbildung	Naturwissenschaftliche Grundbildung
Australien	685	679	675
Belgien	659	672	656
Brasilien	539	499	531
Dänemark	645	649	645
Deutschland	650	649	649
Finnland	681	664	674
Frankreich	645	656	663
Griechenland	625	617	616
Großbritannien	682	676	687
Irland	669	630	661
Island	647	649	635
Italien	627	600	633
Japan	650	688	688
Kanada	681	668	670
Korea	629	676	674
Lettland	617	625	620
Liechtenstein	625	665	629
Luxemburg	592	588	593
Mexiko	565	527	554
Neuseeland	692	689	683
Norwegen	660	643	649
Österreich	648	661	659
Polen	630	632	639
Portugal	620	596	604
Russland	608	648	625
Schweden	657	656	660
Schweiz	651	682	656
Spanien	620	621	643
Tschechien	638	655	663
USA	669	652	658
Ungarn	626	648	659

Führen Sie für die Daten in Tabelle 5.7 eine Hauptkomponentenanalyse durch. Gehen Sie hierbei auf folgende Aspekte ein:

1. Soll eine Hauptkomponentenanalyse der empirischen Varianz-Kovarianz-Matrix oder der empirischen Korrelationsmatrix durchgeführt werden?
2. Wie viele Hauptkomponenten benötigt man?
3. Weist die Graphik des minimal spannenden Baumes auf schlechte Anpassung hin?

Exercise 11. Führen Sie für die Daten in Übung 1 auf Seite 68 eine Hauptkomponentenanalyse mit `S-PLUS` durch. Gehen Sie hierbei auf folgende Aspekte ein:

1. Soll eine Hauptkomponentenanalyse der empirischen Varianz-Kovarianz-Matrix oder der empirischen Korrelationsmatrix durchgeführt werden?
2. Wie viele Hauptkomponenten benötigt man?
3. Weist die Graphik des minimal spannenden Baumes auf schlechte Anpassung hin?

6 Mehrdimensionale Skalierung

6.1 Problemstellung

Bisher haben wir Datensätze analysiert, bei denen die Daten in Form von Datenmatrizen anfielen. Bei jedem von n Objekten wurden p Merkmale erhoben. Sind alle Merkmale quantitativ, so ist mit Hilfe der Hauptkomponentenanalyse eine approximative Darstellung der Objekte im $\mathbb{R}^2$ möglich. In der Praxis sind nicht alle Merkmale in einer Datenmatrix quantitativ. Im Beispiel 3 auf Seite 5 sind nur die Merkmale `Alter`, `Größe` und `Gewicht` quantitativ. Mit Hilfe der mehrdimensionalen Skalierung ist es aber auch hier möglich, eine zweidimensionale Darstellung der Studenten unter Berücksichtigung aller Merkmale zu erhalten. Hierzu muss man zunächst Distanzen zwischen allen Paaren von Studenten bestimmen. Im Beispiel 18 auf Seite 103 haben wir den Gower-Koeffizienten bestimmt. Die Distanzen sind in der Matrix $\mathbf{D}$ in Gleichung (4.13) auf Seite 103 zu finden. Nimmt man die Distanzmatrix als Ausgangspunkt, so kann man aus dieser eine Konfiguration der Objekte im $\mathbb{R}^2$ bestimmen, die die Distanzen zwischen den Objekten möglichst gut wiedergibt. Abbildung 6.1 zeigt die approximative Konfiguration der Punkte im $\mathbb{R}^2$.
Wir werden im nächsten Abschnitt sehen, wie man die Konfiguration aus der Distanzmatrix gewinnt. Sehr oft ist die Distanzmatrix der Ausgangspunkt der Analyse. In Tabelle 1.5 auf Seite 7 sind die Entfernungen zwischen deutschen Städten zu finden. Abbildung 6.2 beinhaltet eine Darstellung der Städte in der Ebene, die mit Hilfe eines Verfahrens der mehrdimensionalen Skalierung gewonnen wurde.
Wie wir sehen, entspricht dies nicht der gewohnten Ausrichtung der Konfiguration. Nach einer Drehung der Konfiguration um 90 Grad im Gegenzeigersinn erhalten wir das gewohnte Bild, das in Abbildung 6.3 zu finden ist.
Eine mit Hilfe einer mehrdimensionalen Skalierung gewonnene Konfiguration ist also bezüglich ihrer Lage und Ausrichtung nicht eindeutig. Verschiebungen der Konfiguration und Drehungen um den Nullpunkt ändern die Distanzen zwischen den Punkten nicht. Bezüglich der Lage wird die Konfiguration in der Regel dadurch eindeutig, dass man ihr Zentrum in den Ursprung legt. Anschließende Drehungen erleichtern unter Umständen die Interpretation.

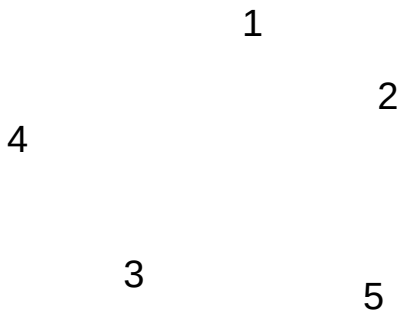


Fig. 6.1. Graphische Darstellung von 5 Studenten

Wir werden uns im Folgenden mit zwei Verfahren der mehrdimensionalen Skalierung beschäftigen. Die *metrische mehrdimensionale Skalierung* geht von einer Distanzmatrix aus. Man sucht eine Konfiguration von Punkten, sodass die Distanzen zwischen den Punkten der Konfiguration möglichst gut die Distanzen in der Distanzmatrix wiedergeben. Bei der *nichtmetrischen mehrdimensionalen Skalierung* ist man nicht an den Distanzen selbst, sondern an der Reihenfolge der Distanzen interessiert. Man sucht also eine Konfiguration von Punkten, sodass die Reihenfolge der Distanzen zwischen den Punkten der Konfiguration der Reihenfolge der Distanzen in der Distanzmatrix entspricht.

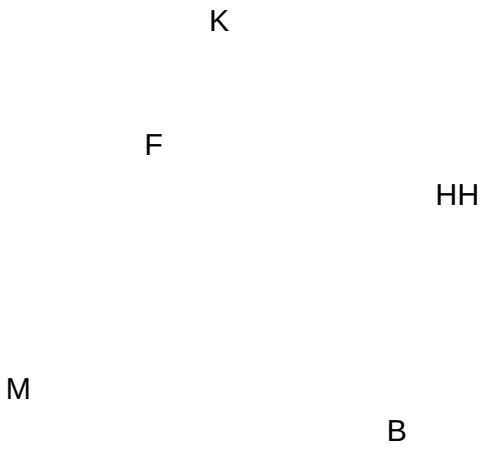


Fig. 6.2. Konfiguration von 5 deutschen Städten, die mit Hilfe der mehrdimensionalen Skalierung gewonnen wurde

6.2 Metrische mehrdimensionale Skalierung

6.2.1 Theorie

Sind n Punkte $\mathbf{x}_1, \ldots, \mathbf{x}_n$ gegeben, so können wir mit der Gleichung (4.1) auf Seite 94 die euklidischen Distanzen zwischen den Punkten bestimmen:

Example 25. Gegeben seien die folgenden Punkte

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}.$$

Abbildung 6.4 zeigt die Punkte.

Die euklidischen Distanzen zwischen diesen Punkten sind

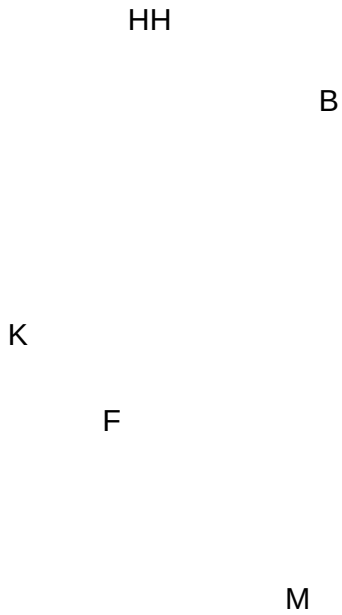


Fig. 6.3. Konfiguration von 5 deutschen Städten, die mit Hilfe der mehrdimensionalen Skalierung gewonnen wurde, nach Drehung

$$d_{12} = \sqrt{(1-5)^2 + (1-1)^2} = 4,$$
$$d_{13} = \sqrt{(1-1)^2 + (1-4)^2} = 3,$$
$$d_{23} = \sqrt{(5-1)^2 + (1-4)^2} = 5.$$

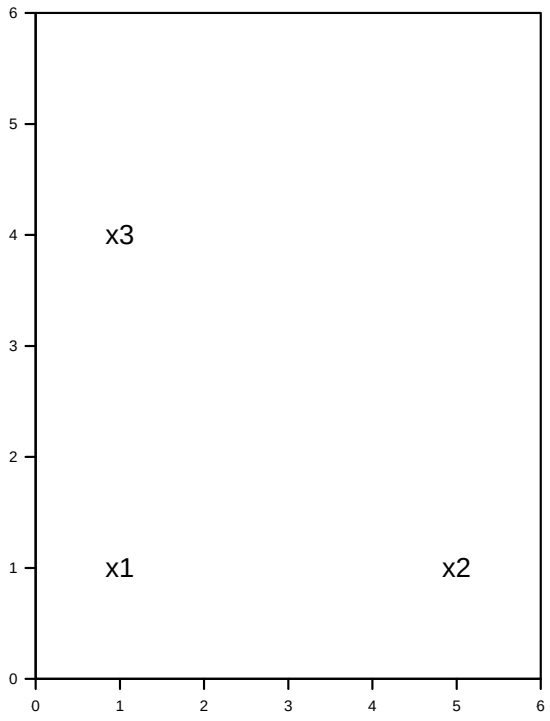


Fig. 6.4. Graphische Darstellung von 3 Punkten

□

Nun kehren wir die Fragestellung um. Wir gehen aus von einer Distanzmatrix **D** und suchen eine Konfiguration von Punkten im $\mathbb{R}^2$, die genau diese Distanzmatrix besitzt. hmcounterend. (fortgesetzt)

Example 25. Für die Distanzmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 3 \\ 4 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix} \tag{6.1}$$

kennen wir eine Lösung. □

Die im Beispiel gewonnene Lösung ist nicht eindeutig. Die Konfiguration kann verschoben oder um den Nullpunkt gedreht werden, ohne dass sich die Distanzen zwischen den Punkten ändern. Schauen wir uns dies für das Beispiel an. hmcounterend. (fortgesetzt)

Example 25. Gegeben sei folgende Konfiguration:

$$\mathbf{y}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \quad \mathbf{y}_3 = \begin{pmatrix} 2 \\ 5 \end{pmatrix}.$$

Es gilt

$$d_{12} = 4, \qquad d_{13} = 3, \qquad d_{23} = 5.$$

Diese Distanzen sind identisch mit den Distanzen in der Matrix $\mathbf{D}$ in (6.1). Abbildung 6.5 zeigt, dass die Konfiguration der Punkte $\mathbf{y}_1$, $\mathbf{y}_2$ und $\mathbf{y}_3$ durch Verschieben aus der Konfiguration der Punkte $\mathbf{x}_1$, $\mathbf{x}_2$ und $\mathbf{x}_3$ gewonnen wurde.

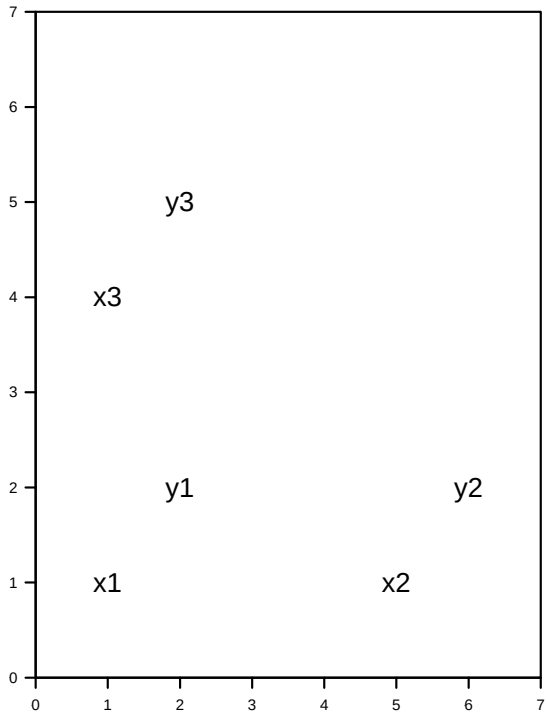


Fig. 6.5. Graphische Darstellung von Punkten

□

hmcounterend. (fortgesetzt)

Example 25. Gegeben sei folgende Konfiguration:

$$\mathbf{z}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \mathbf{z}_2 = \begin{pmatrix} -1 \\ 5 \end{pmatrix}, \quad \mathbf{z}_3 = \begin{pmatrix} -4 \\ 1 \end{pmatrix}.$$

Es gilt

$$d_{12} = 4, \qquad d_{13} = 3, \qquad d_{23} = 5.$$

Die Distanzen zwischen diesen Punkten sind also auch in der Matrix $\mathbf{D}$ in (6.1) zu finden. Abbildung 6.6 zeigt, dass die Konfiguration der Punkte $\mathbf{z}_1$, $\mathbf{z}_2$ und $\mathbf{z}_3$ durch Drehung im Gegenzeigersinn um den Nullpunkt um 90 Grad aus der Konfiguration der Punkte $\mathbf{x}_1$, $\mathbf{x}_2$ und $\mathbf{x}_3$ gewonnen wurde.

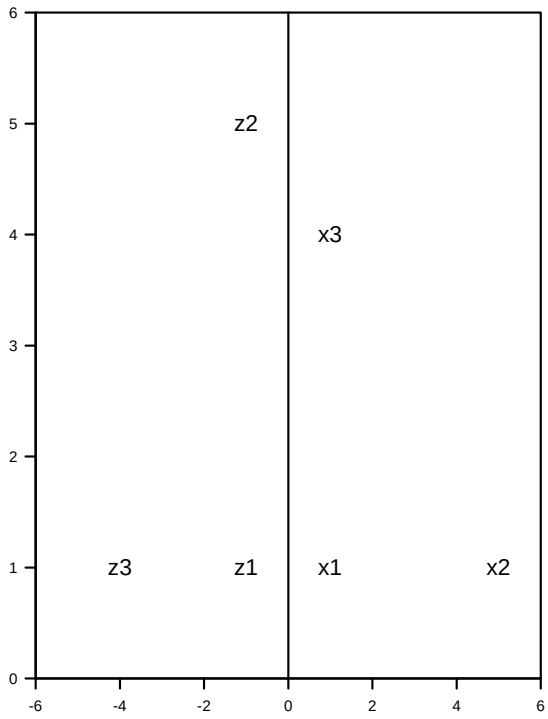


Fig. 6.6. Graphische Darstellung von Punkten

□

Wir wollen uns nun anschauen, wie man aus einer $(3, 3)$-Distanzmatrix eine Konfiguration im $\mathbb{R}^2$ graphisch konstruieren kann. Wir wählen hierzu die

Distanz d_{12} aus der Distanzmatrix aus und zeichnen zwei Punkte $\mathbf{x}_1$ und $\mathbf{x}_2$, deren Abstand gleich d_{12} ist. Den Punkt $\mathbf{x}_3$ erhalten wir dadurch, dass wir einen Kreis mit Mittelpunkt $\mathbf{x}_1$ und Radius d_{13} und einen Kreis mit Mittelpunkt $\mathbf{x}_2$ und Radius d_{23} zeichnen. Schneiden sich die beiden Kreise in zwei Punkten, so gibt es zwei Lösungen. hmcounterend. (fortgesetzt)

Example 25. Wir gehen also aus von

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 3 \\ 4 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix}.$$

Die Distanz der Punkte

$$\mathbf{x}_1 = \begin{pmatrix} 4 \\ 0 \end{pmatrix} \tag{6.2}$$

und

$$\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{6.3}$$

beträgt 4. Abbildung 6.7 zeigt die Konstruktion des dritten Punktes.

□

Es kann aber auch passieren, dass sich die beiden Kreise in genau einem Punkt schneiden. In diesem Fall können wir die Punkte im $\mathbb{R}^1$ darstellen.

Example 26. Wir betrachten die Distanzmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 1 \\ 4 & 0 & 3 \\ 1 & 3 & 0 \end{pmatrix}. \tag{6.4}$$

Wir wählen wieder die Punkte $\mathbf{x}_1$ und $\mathbf{x}_2$ in (6.2) und (6.3) und erhalten die Abbildung 6.8.

□

Möglich ist aber auch, dass die beiden Kreise sich gar nicht schneiden. In diesem Fall gibt es keine Konfiguration von Punkten im $\mathbb{R}^2$, deren Distanzen mit denen in der Distanzmatrix übereinstimmen.

Example 27. Sei

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 1 \\ 4 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}. \tag{6.5}$$

Es gilt

$$d_{12} > d_{13} + d_{23}.$$

Somit ist die Dreiecksungleichung verletzt, und es gibt keine Konfiguration im $\mathbb{R}^2$, die die Distanzen reproduziert. Wir wählen wieder die Punkte $\mathbf{x}_1$ und $\mathbf{x}_2$ in (6.2) und (6.3) und erhalten die Abbildung 6.9.

□

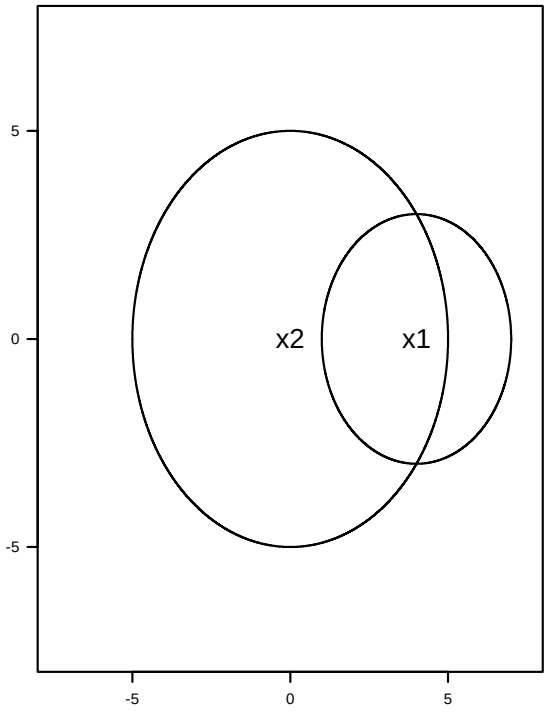


Fig. 6.7. Konstruktion einer Graphik aus der Distanzmatrix, wobei zwei Lösungen existieren

Bei einer $(3, 3)$-Distanzmatrix gibt es also drei Möglichkeiten:

1. Es können drei Punkte im $\mathbb{R}^2$ gefunden werden, sodass die euklidischen Distanzen zwischen diesen Punkten mit denen in der Distanzmatrix übereinstimmen.
2. Es können drei Punkte im $\mathbb{R}^1$ gefunden werden, sodass die Abstände dieser Punkte mit denen in der Distanzmatrix übereinstimmen.
3. Es kann keine Konfiguration im $\mathbb{R}^2$ oder $\mathbb{R}^1$ gefunden werden.

Es stellen sich zwei Fragen:

1. Wie kann man herausfinden, ob und, wenn ja, in welchem Raum eine Darstellung der Distanzen durch eine Punktekonfiguration möglich ist?
2. Wie kann man bei Distanzmatrizen eine Konfiguration von Punkten im $\mathbb{R}^k$ finden, sodass die Abstände zwischen den Punkten mit denen in der Distanzmatrix übereinstimmen?

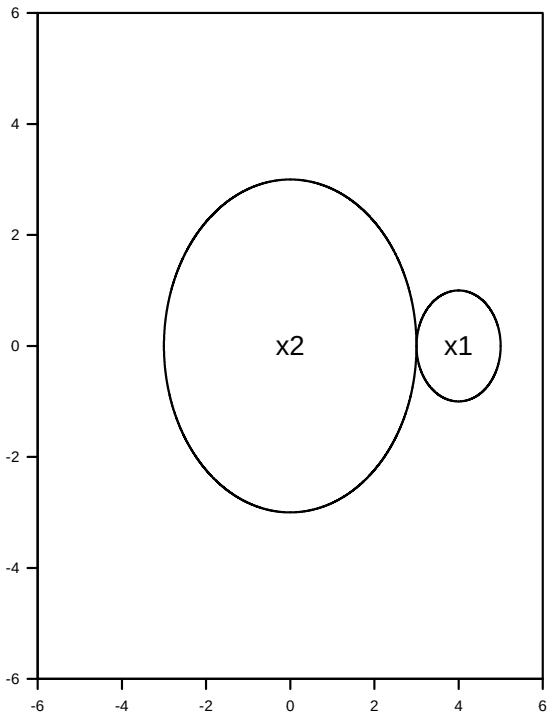


Fig. 6.8. Konstruktion einer Graphik aus der Distanzmatrix, wobei genau eine Lösung existiert

Im Folgenden werden wir diese Fragen beantworten. Dabei gehen wir von einer Distanzmatrix $\mathbf{D} = (d_{rs}), r = 1, \ldots, n, s = 1, \ldots, n$ aus. Gesucht ist eine Konfiguration im $I\!R^k$ mit den in $\mathbf{D}$ angegebenen Distanzen. Zur Bestimmung der Lösung drehen wir zunächst die Fragestellung um. Wir gehen aus von einer Datenmatrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

und bestimmen die quadrierten euklidischen Distanzen d_{rs}^2 zwischen den Zeilenvektoren, d.h. zwischen den Objekten.

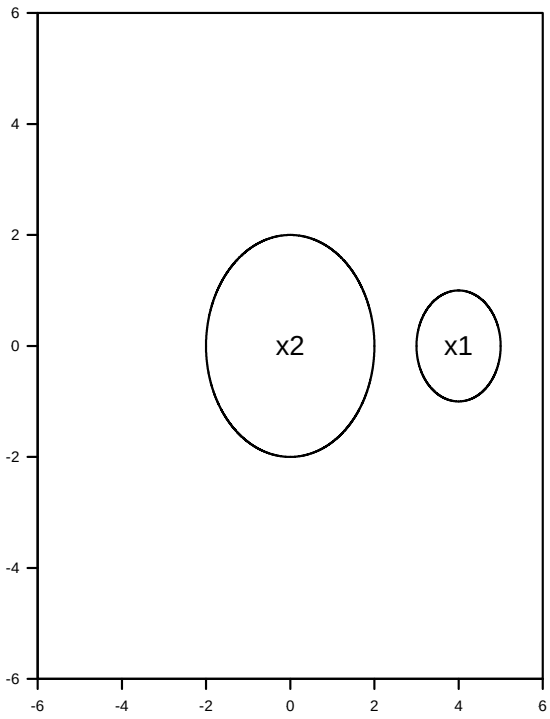


Fig. 6.9. Konstruktion einer Graphik aus der Distanzmatrix, wobei keine Lösung existiert

Es gilt

$$d_{rs}^2 = b_{rr} + b_{ss} - 2\,b_{rs} \tag{6.6}$$

mit

$$b_{rr} = \sum_{j=1}^{p} x_{rj}^2, \tag{6.7}$$

$$b_{ss} = \sum_{j=1}^{p} x_{sj}^2, \tag{6.8}$$

$$b_{rs} = \sum_{j=1}^{p} x_{rj} x_{sj}. \tag{6.9}$$

Dies sieht man folgendermaßen:

$$d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2 = \sum_{j=1}^{p} x_{rj}^2 + \sum_{j=1}^{p} x_{sj}^2 - 2 \sum_{j=1}^{p} x_{rj} x_{sj}$$

$$= b_{rr} + b_{ss} - 2b_{rs}.$$

Mit

$$\mathbf{x}_r = \begin{pmatrix} x_{r1} \\ \vdots \\ x_{rp} \end{pmatrix}$$

und

$$\mathbf{x}_s = \begin{pmatrix} x_{s1} \\ \vdots \\ x_{sp} \end{pmatrix}$$

können wir dies auch vektoriell folgendermaßen darstellen:

$$b_{rr} = \mathbf{x}_r' \mathbf{x}_r,$$

$$b_{ss} = \mathbf{x}_s' \mathbf{x}_s,$$

$$b_{rs} = \mathbf{x}_r' \mathbf{x}_s.$$

In matrizieller Form können wir die Matrix $\mathbf{B} = (b_{rs})$ folgendermaßen schreiben:

$$\begin{aligned}\mathbf{B} &= \mathbf{X}\mathbf{X}' \\ &= \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{x}_1'\mathbf{x}_1 & \dots & \mathbf{x}_1'\mathbf{x}_n \\ \vdots & \ddots & \vdots \\ \mathbf{x}_n'\mathbf{x}_1 & \dots & \mathbf{x}_n'\mathbf{x}_n \end{pmatrix}.\end{aligned}$$

Example 28. Wir schauen uns dies für die Punkte

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

an. Die Datenmatrix lautet

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 5 & 1 \\ 1 & 4 \end{pmatrix}.$$

Wir bestimmen die Matrix $\mathbf{B}$. Es gilt

$$\mathbf{B} = \mathbf{X}\mathbf{X}' = \begin{pmatrix} 2 & 6 & 5 \\ 6 & 26 & 9 \\ 5 & 9 & 17 \end{pmatrix}. \tag{6.10}$$

Mit (6.6) folgt

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 3 \\ 4 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix}.$$

Schauen wir uns dies exemplarisch für d_{12} an. Es gilt

$$d_{12}^2 = b_{11} + b_{22} - 2\,b_{12} = 2 + 26 - 2 \cdot 6 = 16.$$

Also gilt

$$d_{12} = 4.$$

□

Ist $\mathbf{B}$ bekannt, so lässt sich $\mathbf{D}$ also bestimmen. Könnte man aber von $\mathbf{D}$ auf $\mathbf{B}$ schließen, so könnte man die Konfiguration $\mathbf{X}$ folgendermaßen ermitteln: Man führt eine Spektralzerlegung von $\mathbf{B}$ durch und erhält

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}', \tag{6.11}$$

wobei die Matrix $\mathbf{U}$ eine orthogonale Matrix ist, deren Spaltenvektoren die Eigenvektoren $\mathbf{u}_1, \ldots, \mathbf{u}_n$ von $\mathbf{B}$ sind, und $\boldsymbol{\Lambda}$ eine Diagonalmatrix ist, deren Hauptdiagonalelemente $\lambda_1, \ldots, \lambda_n$ die Eigenwerte von $\mathbf{B}$ sind, siehe dazu Kapitel A.1.9 auf Seite 478. Sind die Eigenwerte von $\mathbf{B}$ alle nichtnegativ, so können wir die Diagonalmatrix $\boldsymbol{\Lambda}^{0.5}$ mit

$$\boldsymbol{\Lambda}^{0.5} = \begin{pmatrix} \sqrt{\lambda_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\lambda_n} \end{pmatrix}$$

bilden. Es gilt

$$\boldsymbol{\Lambda}^{0.5}\boldsymbol{\Lambda}^{0.5} = \boldsymbol{\Lambda}. \tag{6.12}$$

Mit (6.12) können wir (6.11) folgendermaßen umformen:

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}' = \mathbf{U}\boldsymbol{\Lambda}^{0.5}\boldsymbol{\Lambda}^{0.5}\mathbf{U}' = \mathbf{U}\boldsymbol{\Lambda}^{0.5}(\mathbf{U}\boldsymbol{\Lambda}^{0.5})'.$$

Mit

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{0.5} \tag{6.13}$$

gilt also

$$\mathbf{B} = \mathbf{X}\mathbf{X}'. \tag{6.14}$$

Wir können also aus $\mathbf{B}$ die Matrix $\mathbf{X}$ der Konfiguration bestimmen. Notwendig hierfür ist jedoch, dass alle Eigenwerte von $\mathbf{B}$ nichtnegativ sind. Sind in diesem Fall einige Eigenwerte gleich 0, so ist eine Darstellung in einem Raum niedriger Dimension möglich. Sind Eigenwerte negativ, so ist die Zerlegung (6.14) mit (6.13) nicht möglich. Hier kann man aber die Eigenvektoren zu positiven Eigenwerten auswählen, um eine Konfiguration von Punkten zu finden. Diese wird die Abstände nur approximativ wiedergeben. hmcounterend. (fortgesetzt)

Example 28. Die Eigenwerte der Matrix $\mathbf{B}$ sind $\lambda_1 = 33.47$, $\lambda_2 = 11.53$ und $\lambda_3 = 0$ und die Eigenvektoren sind

$$\mathbf{u}_1 = \begin{pmatrix} 0.239 \\ 0.820 \\ 0.521 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0.087 \\ -0.552 \\ 0.829 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0.967 \\ -0.153 \\ -0.204 \end{pmatrix}.$$

Es gilt also

$$\mathbf{U} = \begin{pmatrix} 0.239 & 0.087 & 0.967 \\ 0.820 & -0.552 & -0.153 \\ 0.521 & 0.829 & -0.204 \end{pmatrix}$$

und

$$\boldsymbol{\Lambda} = \begin{pmatrix} 33.47 & 0 & 0 \\ 0 & 11.53 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Somit gilt

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{0.5} = \begin{pmatrix} 1.383 & 0.297 & 0 \\ 4.742 & -1.875 & 0 \\ 3.012 & 2.816 & 0 \end{pmatrix}.$$

Abbildung 6.10 zeigt die Konfiguration der Punkte.

□

Nun müssen wir nur noch einen Weg finden, um **B** aus **D** zu gewinnen. Wir haben in (6.6) gesehen, dass gilt

$$d_{rs}^2 = b_{rr} + b_{ss} - 2b_{rs}.$$

Wir lösen diese Gleichung nach b_{rs} auf und erhalten

$$b_{rs} = -0.5 \left(d_{rs}^2 - b_{rr} - b_{ss} \right). \tag{6.15}$$

Nun müssen wir noch b_{ss} und b_{rr} in Abhängigkeit von d_{rs}^2 darstellen. Da die Konfiguration bezüglich der Lage nicht eindeutig ist, legen wir ihren Schwerpunkt in den Ursprung. Wir nehmen also an

$$\sum_{r=1}^{n} x_{rj} = 0 \quad \text{für} \quad j = 1, \ldots, p.$$

Hieraus folgt

$$\sum_{r=1}^{n} b_{rs} = 0 \quad \text{für} \quad s = 1, \ldots, n. \tag{6.16}$$

Mit (6.9) sieht man dies folgendermaßen:

$$\sum_{r=1}^{n} b_{rs} = \sum_{r=1}^{n} \sum_{j=1}^{p} x_{rj}\, x_{sj} = \sum_{j=1}^{p} \sum_{r=1}^{n} x_{sj}\, x_{rj} = \sum_{j=1}^{p} x_{sj} \sum_{r=1}^{n} x_{rj} = 0.$$

Analog erhalten wir

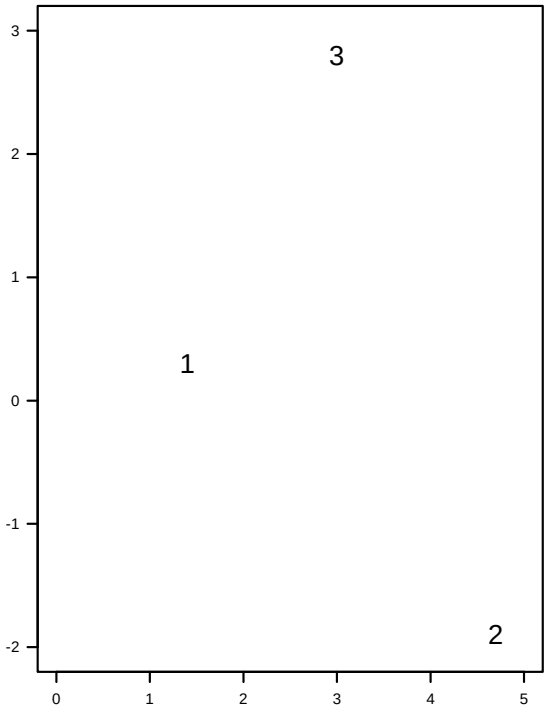


Fig. 6.10. Graphische Darstellung von 3 Punkten

$$\sum_{s=1}^{n} b_{rs} = 0 \quad \text{für} \quad r = 1, \ldots, n. \tag{6.17}$$

Summieren wir (6.6) über r, so folgt mit (6.16)

$$\sum_{r=1}^{n} d_{rs}^2 = \sum_{r=1}^{n} b_{rr} + \sum_{r=1}^{n} b_{ss} - 2\sum_{r=1}^{n} b_{rs} = \sum_{r=1}^{n} b_{rr} + n\, b_{ss}.$$

Mit

$$T = \sum_{r=1}^{n} b_{rr} \tag{6.18}$$

gilt also

$$\sum_{r=1}^{n} d_{rs}^2 = T + n\, b_{ss}\,. \tag{6.19}$$

Somit gilt

$$b_{ss} = \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{T}{n}. \tag{6.20}$$

Analog erhalten wir

$$b_{rr} = \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{T}{n}. \tag{6.21}$$

Setzen wir in (6.15) für b_{ss} die rechte Seite von (6.20) und für b_{rr} die rechte Seite von (6.21) ein, so erhalten wir

$$\begin{aligned} b_{rs} &= -0.5\left(d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{T}{n} - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{T}{n}\right) \\ &= -0.5\left(d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{2T}{n}\right). \end{aligned} \tag{6.22}$$

Nun müssen wir nur noch T in Abhängigkeit von d_{rs} darstellen. Summiert man (6.19) über s und berücksichtigt (6.18), so gilt:

$$\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = \sum_{s=1}^{n} T + \sum_{s=1}^{n} n\, b_{ss} = n\,T + n\,T = 2\,n\,T. \tag{6.23}$$

Aus (6.23) folgt

$$T = \frac{1}{2n}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2. \tag{6.24}$$

Setzen wir für T in (6.22) die rechte Seite von (6.24) ein, so erhalten wir

$$\begin{aligned} b_{rs} &= -0.5\left(d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right) \\ &= -0.5\,d_{rs}^2 + \frac{1}{n}\sum_{r=1}^{n} 0.5\,d_{rs}^2 + \frac{1}{n}\sum_{s=1}^{n} 0.5\,d_{rs}^2 - \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} 0.5\,d_{rs}^2. \end{aligned}$$

Wir haben also eine Möglichkeit gefunden, die Matrix $\mathbf{B}$ aus der Matrix $\mathbf{D}$ zu gewinnen. Wir können diese Beziehung noch übersichtlicher gestalten. Mit

$$a_{rs} = -0.5\,d_{rs}^2$$

erhalten wir

$$b_{rs} = a_{rs} - \frac{1}{n}\sum_{r=1}^{n} a_{rs} - \frac{1}{n}\sum_{s=1}^{n} a_{rs} + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} a_{rs}.$$

Mit

$$\bar{a}_{r.} = \frac{1}{n} \sum_{s=1}^{n} a_{rs},$$

$$\bar{a}_{.s} = \frac{1}{n} \sum_{r=1}^{n} a_{rs},$$

$$\bar{a}_{..} = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} a_{rs}$$

gilt also

$$b_{rs} = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..}. \tag{6.25}$$

Im Folgenden ist $\mathbf{A} = (a_{rs})$ mit

$$a_{rs} = -0.5\, d_{rs}^2.$$

Die Transformation

$$b_{rs} = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..}$$

beinhaltet eine doppelte Zentrierung der Matrix $\mathbf{A}$ in dem Sinne, dass zuerst die Spalten der Matrix $\mathbf{A}$ zentriert werden, und anschließend in dieser Matrix die Zeilen zentriert werden. Denn zentrieren wir zuerst die Spalten von $\mathbf{A}$, so erhalten wir die Matrix $\tilde{\mathbf{A}} = (\tilde{a}_{rs})$ mit

$$\tilde{a}_{rs} = a_{rs} - \bar{a}_{.s}\,.$$

Es gilt

$$\bar{\tilde{a}}_{r.} = \bar{a}_{r.} - \bar{a}_{..}\,.$$

Dies sieht man folgendermaßen:

$$\bar{\tilde{a}}_{r.} = \frac{1}{n} \sum_{s=1}^{n} \tilde{a}_{rs} = \frac{1}{n} \sum_{s=1}^{n} (a_{rs} - \bar{a}_{.s})$$

$$= \frac{1}{n} \sum_{s=1}^{n} a_{rs} - \frac{1}{n} \sum_{s=1}^{n} \frac{1}{n} \sum_{r=1}^{n} a_{rs} = \bar{a}_{r.} - \bar{a}_{..}\,.$$

Zentrieren wir nun die Zeilen der Matrix $\tilde{\mathbf{A}}$, so erhalten wir

$$\tilde{a}_{rs} - \bar{\tilde{a}}_{r.} = a_{rs} - \bar{a}_{.s} - (\bar{a}_{r.} - \bar{a}_{..}) = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..}.$$

Somit ist gezeigt, dass die Matrix $\mathbf{B}$ aus der Matrix $\mathbf{A}$ durch doppelte Zentrierung hervorgeht.

Wir haben in Gleichung (2.11) auf Seite 26 die Zentrierungmatrix $\mathbf{M}$ betrachtet:

$$\mathbf{M} = \mathbf{I}_n - \frac{1}{n}\,\mathbf{1}\mathbf{1}'.$$

Es gilt also

$$\mathbf{B} = \mathbf{MAM}\,. \tag{6.26}$$

Mit Hilfe von (6.26) können wir zeigen, dass die Dimension des Raumes, in dem eine exakte Darstellung einer (n,n)-Distanzmatrix möglich ist, höchstens n-1 ist. Mindestens einer der Eigenwerte von $\mathbf{B}$ ist 0. Wegen

$$\mathbf{M1} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{1} = \mathbf{I}_n\mathbf{1} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{1} = \mathbf{1} - \frac{1}{n}\mathbf{1}\,n = \mathbf{0}$$

gilt

$$\mathbf{B1} = \mathbf{MAM1} = \mathbf{0} = 0\,\mathbf{1}\,. \tag{6.27}$$

Gleichung (6.27) zeigt, dass $\mathbf{1}$ ein Eigenvektor von $\mathbf{B}$ zum Eigenwert 0 ist.

Fassen wir zusammen. Will man aus einer Distanzmatrix $\mathbf{D}$ eine Konfiguration von Punkten im $\mathbb{R}^2$ bestimmen, deren Abstände die Distanzmatrix gut approximieren, so sollte man folgendermaßen vorgehen:

1. Bilde die Matrix $\mathbf{A} = (a_{rs})$ mit
$$a_{rs} = -0.5\,d_{rs}^2.$$
2. Bilde die Matrix $\mathbf{B} = (b_{rs})$ mit
$$b_{rs} = a_{rs} - \bar{a}_{r\cdot} - \bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}.$$
3. Führe eine Spektralzerlegung von $\mathbf{B}$ durch:
$$\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'.$$
4. Bilde die Diagonalmatrix $\boldsymbol{\Lambda}_1$ mit den beiden größten Eigenwerten λ_1 und λ_2 von $\mathbf{B}$ und die Matrix $\mathbf{U}_1$ mit den zu λ_1 und λ_2 gehörenden normierten Eigenvektoren. Die Konfiguration bilden dann die Zeilenvektoren von
$$\mathbf{X}_1 = \mathbf{U}_1\,\boldsymbol{\Lambda}_1^{0.5}\,.$$
Sind die beiden größten Eigenwerte positiv und alle anderen Eigenwerte gleich 0, so ist die Darstellung im $\mathbb{R}^2$ exakt.

Schauen wir uns die drei Beispiele vom Anfang dieses Kapitels an. hmcounterend. (fortgesetzt)

Example 25. Es gilt

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 3 \\ 4 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix}.$$

Hieraus folgt

$$\mathbf{A} = \begin{pmatrix} 0 & -8 & -4.5 \\ -8 & 0 & -12.5 \\ -4.5 & -12.5 & 0 \end{pmatrix}$$

und

$$\mathbf{B} = \begin{pmatrix} 2.78 & -2.56 & -0.22 \\ -2.56 & 8.11 & -5.56 \\ -0.22 & -5.56 & 5.78 \end{pmatrix}.$$

Die Eigenwerte der Matrix $\mathbf{B}$ lauten $\lambda_1 = 12.9$, $\lambda_2 = 3.7$ und $\lambda_3 = 0$. Da genau zwei Eigenwerte von $\mathbf{B}$ positiv sind, existiert eine exakte Darstellung im $\mathbb{R}^2$. □

hmcounterend. (fortgesetzt)

Example 26. Es gilt

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 1 \\ 4 & 0 & 3 \\ 1 & 3 & 0 \end{pmatrix}.$$

Hieraus folgt

$$\mathbf{A} = \begin{pmatrix} 0 & -8 & -0.5 \\ -8 & 0 & -4.5 \\ -0.5 & -4.5 & 0 \end{pmatrix}$$

und

$$\mathbf{B} = \begin{pmatrix} 2.78 & -3.89 & 1.11 \\ -3.89 & 5.44 & -1.56 \\ 1.11 & -1.56 & 0.44 \end{pmatrix}.$$

Die Eigenwerte der Matrix $\mathbf{B}$ lauten $\lambda_1 = 8.67$, $\lambda_2 = 0$ und $\lambda_3 = 0$. Somit ist eine exakte Darstellung im $\mathbb{R}^1$ möglich. □

hmcounterend. (fortgesetzt)

Example 27. Es gilt

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 1 \\ 4 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}.$$

Es gilt

$$\mathbf{A} = \begin{pmatrix} 0 & -8 & -0.5 \\ -8 & 0 & -2 \\ -0.5 & -2 & 0 \end{pmatrix}$$

und

$$\mathbf{B} = \begin{pmatrix} 3.33 & -4.17 & 0.83 \\ -4.17 & 4.33 & -0.17 \\ 0.83 & -0.17 & -0.67 \end{pmatrix}.$$

Die Eigenwerte der Matrix $\mathbf{B}$ lauten $\lambda_1 = 8.0826$, $\lambda_2 = 0$ und $\lambda_3 = -1.0826$. Da einer der Eigenwerte negativ ist, ist keine exakte Darstellung im $\mathbb{R}^1$ oder $\mathbb{R}^2$ möglich. □

6.2.2 Praktische Aspekte

Wahl der Dimension Wir haben bisher immer eine Darstellung der Distanzmatrix im $\mathbb{R}^2$ gesucht. Es stellt sich natürlich die Frage, wie gut die Darstellung im $\mathbb{R}^2$ die Distanzen reproduziert. Wie schon bei der Hauptkomponentenanalyse können wir diese Frage mit Hilfe von Eigenwerten beantworten. Es liegt nahe, die Summe der Distanzen als Ausgangspunkt zu wählen, aus technischen Gründen ist es aber sinnvoller, die Summe der quadrierten Distanzen zu betrachten. Mit (6.18) und (6.23) gilt

$$\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = 2\,n \sum_{r=1}^{n} b_{rr}.$$

Es gilt

$$\sum_{r}^{n} b_{rr} = tr(\mathbf{B}).$$

Da die Spur einer Matrix gleich der Summe der Eigenwerte ist, und einer der Eigenwerte von $\mathbf{B}$ gleich 0 ist, gilt

$$\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = 2\,n \sum_{i=1}^{n-1} \lambda_i.$$

Wir können auch hier die bei der Hauptkomponentenanalyse auf Seite 133 betrachteten Kriterien verwenden, müssen aber berücksichtigen, dass Eigenwerte negativ sein können. Mardia (1978) hat folgende Größen für die Wahl der Dimension vorgeschlagen:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n-1} |\lambda_i|} \tag{6.28}$$

und

$$\frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n-1} \lambda_i^2}\,. \tag{6.29}$$

Dabei sollten die ersten k Eigenwerte natürlich positiv sein. Man gibt einen Wert α vor und wählt für die Dimension den kleinsten Wert k, für den (6.28) beziehungsweise (6.29) größer oder gleich α sind. Typische Werte für α sind 0.75, 0.8 und 0.85.

Das Problem der additiven Konstanten Wir haben gesehen, dass es Distanzmatrizen gibt, bei denen es keine exakte Darstellung in einem Raum gibt. Dies zeigt sich daran, dass einer oder mehrere Eigenwerte der Matrix $\mathbf{B}$ negativ sind. Im Beispiel 27 auf Seite 160 haben wir die Matrix

$$\mathbf{D} = \begin{pmatrix} 0 & 4 & 1 \\ 4 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}$$

betrachtet. Bei dieser ist die Dreiecksungleichung verletzt. Es gilt

$$d_{12} > d_{13} + d_{23}. \tag{6.30}$$

Addiert man zu allen Elementen von $\mathbf{D}$ außerhalb der Hauptdiagonalen die gleiche Zahl c, so erhöht sich die rechte Seite von (6.30) um $2c$ und die linke Seite um c. Wählt man c geeignet, so wird aus der Ungleichung eine Gleichung. Im Beispiel ist dies für $c = 1$ der Fall. Es gilt

$$d_{12} + 1 = d_{13} + 1 + d_{32} + 1.$$

In diesem Fall existiert eine exakte Darstellung im $\mathbb{R}^1$. Für jeden größeren Wert von c erhalten wir eine exakte Darstellung im $\mathbb{R}^2$. Man kann zeigen, dass diese Vorgehensweise immer möglich ist. Die Herleitung ist zum Beispiel in Cox & Cox (1994), S.35-37 zu finden. Die Addition einer Konstanten ist auf jeden Fall dann sinnvoll, wenn die Daten intervallskaliert sind.

Ein Zusammenhang zwischen der metrischen mehrdimensionalen Skalierung und der Hauptkomponentenanalyse Sowohl mit der Hauptkomponentenanalyse als auch mit der metrischen mehrdimensionalen Skalierung kann man eine Darstellung von Objekten in einem zweidimensionalen Raum gewinnen. Ausgangspunkt der Hauptkomponentenanalyse ist meistens eine Datenmatrix quantitativer Merkmale, während die mehrdimensionale Skalierung auf einer Distanzmatrix basiert. Nun könnte man natürlich auf der Basis einer Datenmatrix zunächst eine Distanzmatrix bestimmen und für diese eine metrische mehrdimensionale Skalierung durchführen, um eine Darstellung der Objekte zu gewinnen. Wurden beim Übergang von der Datenmatrix zur Distanzmatrix euklidische Distanzen bestimmt, so liefert die metrische mehrdimensionale Skalierung auf Basis der euklidischen Distanzen und die Hauptkomponentenanalyse auf Basis der Datenmatrix die gleiche Konfiguration. Ein Beweis dieser Tatsache ist bei Mardia et al. (1979), S.405-406 zu finden. Wir wollen dies für das Datenbeispiel 4 auf Seite 5 illustrieren. Abbildung 6.11 vergleicht die Darstellungen, die mit Hilfe der Hauptkomponentenanalyse und der metrischen mehrdimensionale Skalierung gewonnen wurde. Dreht man die mit der metrischen mehrdimensionalen Skalierung gewonnene Konfiguration um 180 Grad, so sehen wir, dass die beiden Verfahren identische Konfigurationen liefern. Im nächsten Kapitel werden wir mit der *Procrustes-Analyse* ein Verfahren kennenlernen, mit dem man systematisch zwei Konfigurationen so verschieben, stauchen oder strecken, drehen und spiegeln kann, dass sie sich möglichst ähnlich sind.

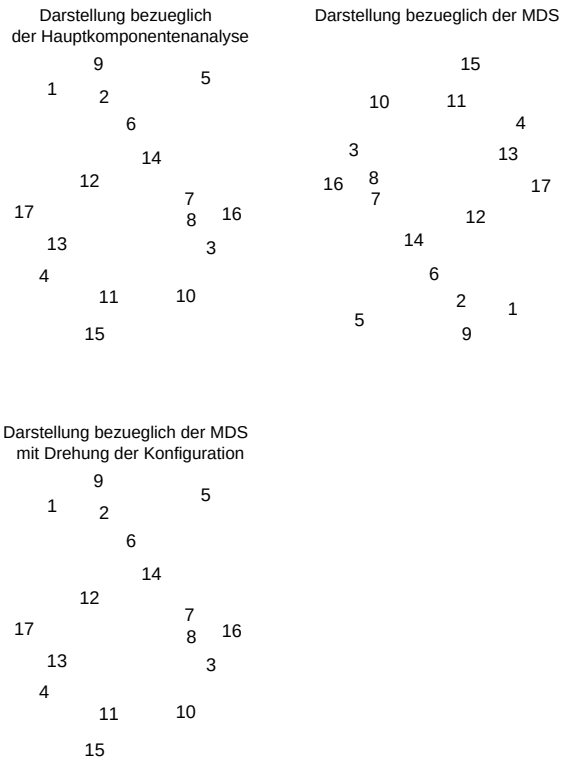


Fig. 6.11. Darstellungen der Noten von 17 Studenten in vier Bereichen. Die erste Graphik zeigt die mit der Hauptkomponentenanalyse gewonnene Darstellung. Die zweite Graphik zeigt die Darstellung, die mit Hilfe der metrischen mehrdimensionalen Skalierung auf Basis einer durch Berechnung euklidischer Distanzen aus der Datenmatrix der Noten bestimmten Distanzmatrix gewonnen wurde. Die dritte Graphik zeigt die zweite Darstellung nach einer Drehung um 180 Grad

.

6.2.3 Metrische mehrdimensionale Skalierung der Rangreihung der Politikerpaare

Wir wollen nun die Vorgehensweise der metrischen mehrdimensionalen Skalierung veranschaulichen. Im Kapitel 4.4 haben wir ein Beispiel betrachtet, bei dem ein Student gebeten wurde, alle Paare von 5 Politikern der Ähnlichkeit nach zu ordnen. Man spricht von Rangreihung. Die Daten sind in Tabelle 4.8 auf Seite 111 zu finden. Wir fassen die Ränge als Distanzen auf und versuchen, eine Darstellung der Politiker zu finden. Hierzu erstellen wir zunächst die Distanzmatrix **D**:

$$\mathbf{D} = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}. \tag{6.31}$$

Dann führen wir eine metrische mehrdimensionale Skalierung durch.

Die Matrix $\mathbf{A}$ lautet

$$\mathbf{A} = \begin{pmatrix} 0 & -40.5 & -8 & -50 & -24.5 \\ -40.5 & 0 & -4.5 & -0.5 & -2 \\ -8 & -4.5 & 0 & -32 & -18 \\ -50 & -0.5 & -32 & 0 & -12.5 \\ -24.5 & -2 & -18 & -12.5 & 0 \end{pmatrix}.$$

Wir zentrieren diese doppelt und erhalten die Matrix $\mathbf{B}$:

$$\mathbf{B} = \begin{pmatrix} 33.8 & -21.8 & 13.7 & -21.8 & -3.9 \\ -21.8 & 3.6 & 2.1 & 12.6 & 3.5 \\ 13.7 & 2.1 & 9.6 & -15.9 & -9.5 \\ -21.8 & 12.6 & -15.9 & 22.6 & 2.5 \\ -3.9 & 3.5 & -9.5 & 2.5 & 7.4 \end{pmatrix}.$$

Die Eigenwerte der Matrix $\mathbf{B}$ sind

$$\lambda_1 = 68.37, \quad \lambda_2 = 15.55, \quad \lambda_3 = 7.23, \quad \lambda_4 = -14.15.$$

Wir sehen, dass eine nur approximative Darstellung im $\mathbb{R}^2$ möglich ist. Es gilt

$$\frac{\sum_{i=1}^{2} \lambda_i}{\sum_{i=1}^{4} |\lambda_i|} = 0.797$$

und

$$\frac{\sum_{i=1}^{2} \lambda_i^2}{\sum_{i=1}^{4} \lambda_i^2} = 0.95\,.$$

Somit ist die Darstellung im $\mathbb{R}^2$ angemessen.

Die Eigenvektoren zu den beiden größten Eigenwerten lauten

$$\mathbf{u}_1 = \begin{pmatrix} -0.6918 \\ 0.3347 \\ -0.3170 \\ 0.5391 \\ 0.1350 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -0.3645 \\ 0.4088 \\ 0.6517 \\ -0.2192 \\ -0.4768 \end{pmatrix}.$$

Die Punkte $\mathbf{x}_1 = \sqrt{\lambda_1}\mathbf{u}_1$ und $\mathbf{x}_2 = \sqrt{\lambda_2}\mathbf{u}_2$ der approximativen zweidimensionalen Darstellung lauten also

$$\mathbf{x}_1 = \begin{pmatrix} -5.720 \\ 2.768 \\ -2.621 \\ 4.458 \\ 1.116 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} -1.437 \\ 1.612 \\ 2.570 \\ -0.864 \\ -1.880 \end{pmatrix}. \tag{6.32}$$

Abbildung 6.12 zeigt die Darstellung im $\mathbb{R}^2$. Die waagerechte Achse kann man sehr schön interpretieren. Sie reflektiert das politische Spektrum. Von links nach rechts findet man Fischer, Schröder, Westerwelle, Merkel und Stoiber. Die zweite Dimension gibt den Wähleranteil der Partei zum Zeitpunkt der Befragung wieder.

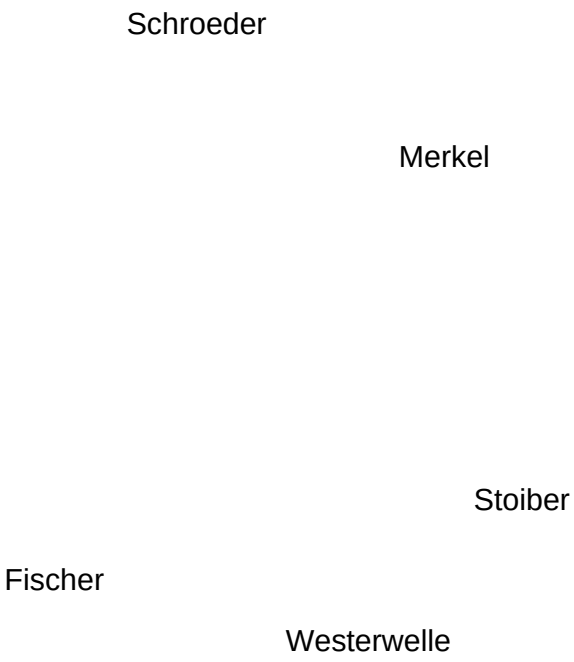


Fig. 6.12. Graphische Darstellung von 5 Politikern mit einer metrischen mehrdimensionalen Skalierung auf der Basis einer Distanzmatrix, die mit der Rangreihung gewonnen wurde

6.2.4 Metrische mehrdimensionale Skalierung in S-PLUS

In `S-PLUS` gibt es eine Funktion `cmdscale`, die eine metrische mehrdimensionale Skalierung durchführt. Der Aufruf von `cmdscale` ist

```
cmdscale(d, k=2, eig=F, add=F)
```

Dabei ist `d` die Distanzmatrix. Diese können wir der Funktion `cmdscale` auf zwei Arten übergeben. Dies schauen wir uns weiter unten an. Die Dimension des Raumes, in dem die Distanzen dargestellt werden sollen, legt man durch das Argument `k` fest. Standardmäßig wird $k = 2$ gewählt. Durch das Argument `eig` kann man festlegen, ob die Eigenwerte ausgegeben werden sollen. Eine additive Konstante kann berücksichtigt werden, wenn das Argument `add` auf das `T` gesetzt wird. Sollen keine Eigenwerte ausgegeben werden, und soll auch keine additive Konstante berücksichtigt werden, so liefert die Funktion `cmdscale` die Koordinaten der Punkte als Ergebnis. Schauen wir uns das Datenbeispiel aus Kapitel 6.2.3 auf Seite 175 in `S-PLUS` an. Die Distanzmatrix ist in Gleichung 6.31 auf Seite 176 zu finden. Wir geben zunächst die Daten ein. Wir erzeugen einen Vektor `v`:

```
> v<-c(9,4,10,7,3,1,2,8,6,5)
```

und weisen dem Attribut `"Size"` den Wert 5 zu:

```
> attr(v,"Size")<-5.
```

Das Attribut von `v` wird beim Aufruf angegeben:

```
> v
  [1]  9  4 10  7  3  1  2  8  6  5
 attr(, "Size"):
 [1] 5
```

Nun können wir mit der Funktion `distfull`, die auf Seite 495 zu finden ist, die Distanzmatrix erzeugen:

```
> dpol<-distfull(v)
> dpol
      [,1] [,2] [,3] [,4] [,5]
 [1,]    0    9    4   10    7
 [2,]    9    0    3    1    2
 [3,]    4    3    0    8    6
 [4,]   10    1    8    0    5
 [5,]    7    2    6    5    0
```

Wir können die Funktion `cmdscale` sowohl mit dem Vektor `v` als auch mit der Matrix `d` aufrufen. Wir wollen eine Darstellung im $\mathbb{R}^2$ ohne Berücksichtigung der additiven Konstante. Es sollen aber die Eigenwerte ausgegeben werden. Wir geben also ein

```
> e<-cmdscale(v,eig=T)
> e
$points:
           [,1]       [,2]
 [1,] -5.720606 -1.4375756
 [2,]  2.767820  1.6121084
```

```
[3,] -2.620944  2.5697830
[4,]  4.457811 -0.8642175
[5,]  1.115923 -1.8800985

$eig:
 [1] 68.37286 15.55094
```

Um die Abbildung 6.12 zu erhalten, erzeugen wir einen Vektor mit den Namen der Politiker:

```
> polnamen<-c("Fischer","Merkel","Schroeder",
              "Stoiber","Westerwelle")
```

und rufen die Funktion `plot` auf:

```
> plot(e,axes=F,xlab="",ylab="",type="n")
> text(e,polnamen)
```

Da wir keine Achsenbeschriftung wünschen, setzen wir `xlab` und `ylab` auf `""`. Das Argument `axes` gibt an, ob die Koordinatenachsen gezeichnet werden sollen, während das Argument `type` angibt, wie die Punkte verbunden werden sollen. Der Wert `"n"` stellt sicher, dass keine Punkte gezeichnet werden. Mit dem zweiten Befehl tragen wir die Namen der Politiker in die Graphik ein.

6.3 Nichtmetrische mehrdimensionale Skalierung

6.3.1 Theorie

Bei der metrischen mehrdimensionalen Skalierung wird eine Konfiguration von Punkten so bestimmt, dass die Distanzen zwischen den Punkten der Konfiguration die entsprechenden Elemente der Distanzmatrix approximieren. Bei vielen Anwendungen ist man aber nicht an den Distanzen, sondern an der Ordnung der Distanzen interessiert. Sehr oft ist sogar nur die Ordnung der Distanzen vorgegeben.

Example 29. Wir betrachten die Tabelle 4.8 auf Seite 111. Wir haben die Merkmale bisher wie quantitative Merkmale behandelt, obwohl sie ordinalskaliert sind. Denn der Student hat keine Distanzen angegeben, sondern nur mitgeteilt, dass das Paar Merkel-Stoiber das ähnlichste Paar, das Paar Merkel-Westerwelle das zweitähnlichste Paar ist, u.s.w.. Die Zahlen 1 bis 10 sind willkürlich. Man hätte auch andere Zahlen vergeben können. Es muss nur sichergestellt sein, dass das ähnlichste Paar die kleinste Zahl, das zweitähnlichste Paar die zweitkleinste Zahl, u.s.w. erhält. □

Um zu verdeutlichen, dass wir nicht an den Distanzen selbst, sondern an der Ordnung der Distanzen interessiert sind, bezeichnen wir die Distanzmatrix im Folgenden mit $\boldsymbol{\Delta}$. Es gilt also

$$\boldsymbol{\Delta} = \begin{pmatrix} \delta_{11} & \dots & \delta_{1n} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \dots & \delta_{nn} \end{pmatrix}.$$

hmcounterend. (fortgesetzt)

Example 29. Es gilt

$$\boldsymbol{\Delta} = \begin{pmatrix} 0 & 9 & 4 & 10 & 7 \\ 9 & 0 & 3 & 1 & 2 \\ 4 & 3 & 0 & 8 & 6 \\ 10 & 1 & 8 & 0 & 5 \\ 7 & 2 & 6 & 5 & 0 \end{pmatrix}.$$

□

Von Kruskal (1964) wurde ein Verfahren vorgeschlagen, bei dem man eine Konfiguration von Punkten im $\mathbb{R}^k$ so bestimmt, dass die euklidischen Distanzen zwischen den Punkten die gleiche Ordnung wie in der Matrix $\boldsymbol{\Delta}$ besitzen. Die Ordnung der Elemente unterhalb der Hauptdiagonalen in der Matrix $\boldsymbol{\Delta}$ bezeichnen wir als *Monotoniebedingung*. hmcounterend. (fortgesetzt)

Example 29. Die Monotoniebedingung lautet

$$\delta_{24} < \delta_{25} < \delta_{23} < \delta_{13} < \delta_{45} < \delta_{35} < \delta_{15} < \delta_{34} < \delta_{12} < \delta_{14}. \tag{6.33}$$

Die kleinste Distanz soll also zwischen den Punkten 2 und 4 sein, die zweitkleinste Distanz zwischen den Punkten 2 und 5, u.s.w.. □

Das Verfahren von Kruskal (1964) beginnt mit einer *Startkonfiguration*. In der Regel wählt man das Ergebnis der metrischen mehrdimensionalen Skalierung als Startkonfiguration $\mathbf{X}$. hmcounterend. (fortgesetzt)

Example 29. Die Konfiguration der metrischen mehrdimensionalen Skalierung ist in (6.32) zu finden. Wir betrachten aus Gründen der Übersichtlichkeit die auf eine Stelle nach dem Komma gerundeten Werte:

$$\mathbf{X} = \begin{pmatrix} -5.7 & -1.4 \\ 2.8 & 1.6 \\ -2.6 & 2.6 \\ 4.5 & -0.9 \\ 1.1 & -1.9 \end{pmatrix}. \tag{6.34}$$

□

Um zu sehen, wie gut die Monotoniebedingung durch die Punkte der Startkonfiguration erfüllt ist, bestimmt man die euklidischen Distanzen d_{ij} zwischen den Punkten der Startkonfiguration und bringt sie in die Reihenfolge, in der die δ_{ij} sind. Wir wollen die aus der Konfiguration gewonnene Distanzmatrix der euklidischen Distanzen mit $\mathbf{D}$ bezeichnen. hmcounterend. (fortgesetzt)

Example 29. Es gilt

$$\mathbf{D} = \begin{pmatrix} 0 & 9.0 & 5.1 & 10.2 & 6.8 \\ 9.0 & 0 & 5.5 & 3.0 & 3.9 \\ 5.1 & 5.5 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 3.5 \\ 6.8 & 3.9 & 5.8 & 3.5 & 0 \end{pmatrix}.$$

Auch hier haben wir die Werte aus Gründen der Übersichtlichkeit auf eine Stelle nach dem Komma gerundet. Wir ordnen die d_{ij} so an wie die δ_{ij} in (6.33):

$$3 \quad 3.9 \quad 5.5 \quad 5.1 \quad 3.5 \quad 5.8 \quad 6.8 \quad 7.9 \quad 9 \quad 10.2.$$

Wir sehen, dass die d_{ij} die Monotoniebedingung nicht erfüllen. □

Dass die Distanzen d_{ij} die Monotoniebedingung nicht erfüllen, ist der Regelfall. Dies bedeutet, dass einige oder alle Punkte verschoben werden müssen. Wie weit die Punkte verschoben werden sollen, sollte davon abhängen, wie stark die Monotoniebedingung verletzt ist. Um das herauszufinden, führen wir eine *monotone Regression* durch. Wir bestimmen sogenannte *Disparitäten* $\hat{d}_{ij}$, die möglichst nahe an den d_{ij} liegen und die Monotoniebedingung erfüllen. hm-counterend. (fortgesetzt)

Example 29. Es muss also gelten

$$\hat{d}_{24} \leq \hat{d}_{25} \leq \hat{d}_{23} \leq \hat{d}_{13} \leq \hat{d}_{45} \leq \hat{d}_{35} \leq \hat{d}_{15} \leq \hat{d}_{34} \leq \hat{d}_{12} \leq \hat{d}_{14}.$$

□

Minimiert man

$$\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2$$

unter der Nebenbedingung, dass die $\hat{d}_{ij}$ die Monotoniebedingung erfüllen, so kann man die $\hat{d}_{ij}$ mit Hilfe des *PAV-Algorithmus* (Pool Adjacent Violators-Algorithmus) bestimmen. Bei diesem durchwandern wir die Sequenz von links nach rechts, bis wir einen Block aufeinanderfolgender Distanzen finden, in dem die Monotoniebedingung verletzt ist. Die Beobachtungen in diesem Block ersetzen wir dann durch ihren Mittelwert. Danach gehen wir wieder an den Anfang und suchen wieder einen Block aufeinanderfolgender Beobachtungen, in dem die Monotoniebedingung verletzt ist, und ersetzen die Beobachtungen dieses Blocks durch deren Mittelwert. Dies machen wir so lange, bis alle Distanzen die Monotoniebedingung erfüllen. Die Disparitäten bilden die Elemente der *Disparitätenmatrix* $\widehat{\mathbf{D}}$:

$$\widehat{\mathbf{D}} = \begin{pmatrix} \hat{d}_{11} & \hat{d}_{12} & \dots & \hat{d}_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{d}_{n1} & \hat{d}_{n2} & \dots & \hat{d}_{nn} \end{pmatrix}.$$

hmcounterend. (fortgesetzt)

Example 29. Wir beginnen mit

$$3 \quad 3.9 \quad 5.5 \quad 5.1 \quad 3.5 \quad 5.8 \quad 6.8 \quad 7.9 \quad 9 \quad 10.2.$$

Im Block

$$5.5 \quad 5.1 \quad 3.5$$

ist die Monotoniebedingung verletzt. Wir ersetzen diese Zahlen durch ihren Mittelwert 4.7 und erhalten nachstehende Folge, die bereits die Monotoniebedingung erfüllt:

$$3 \quad 3.9 \quad 4.7 \quad 4.7 \quad 4.7 \quad 5.8 \quad 6.8 \quad 7.9 \quad 9 \quad 10.2.$$

Abbildung 6.13 verdeutlicht die Vorgehensweise.

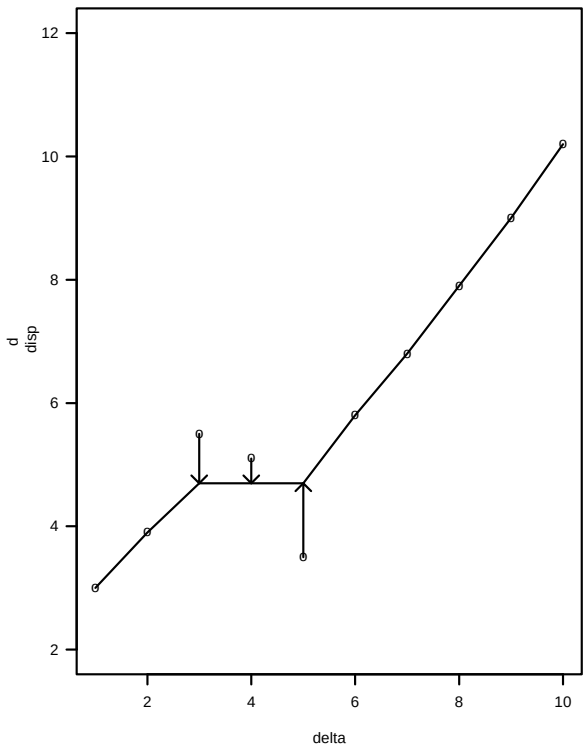


Fig. 6.13. Veranschaulichung der Vorgehensweise der monotonen Regression

Wir erhalten die Disparitätenmatrix

$$\widehat{\mathbf{D}} = \begin{pmatrix} 0 & 9.0 & 4.7 & 10.2 & 6.8 \\ 9.0 & 0 & 4.7 & 3.0 & 3.9 \\ 4.7 & 4.7 & 0 & 7.9 & 5.8 \\ 10.2 & 3.0 & 7.9 & 0 & 4.7 \\ 6.8 & 3.9 & 5.8 & 4.7 & 0 \end{pmatrix}.$$

□

Nun stellt sich die Frage, wie stark die Monotoniebedingung verletzt ist. Es liegt nahe als Maß zu wählen

$$\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2. \tag{6.35}$$

hmcounterend. (fortgesetzt)

Example 29. Es gilt

$$\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2 = 2.24.$$

□

Da (6.35) nicht normiert ist, sagt der Wert wenig aus. Kruskal (1964) hat folgende Größe vorgeschlagen:

$$\text{STRESS1} = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}}. \tag{6.36}$$

Tabelle 6.1 zeigt, wie eine Konfiguration in Abhängigkeit von STRESS1 bewertet werden sollte.

Table 6.1. Bewertung einer Konfiguration anhand von STRESS1

Wert von STRESS1	Güte der Konfiguration
$0.00 \leq STRESS1 < 0.05$	hervorragend
$0.05 \leq STRESS1 < 0.10$	gut
$0.10 \leq STRESS1 < 0.15$	zufriedenstellend
$0.15 \leq STRESS1$	nicht gut

hmcounterend. (fortgesetzt)

Example 29. Es gilt STRESS1 = 0.073. Also wurde eine gute Konfiguration gefunden. □

Ist man mit dem Wert von STRESS1 nicht zufrieden, so sollte man die Konfiguration verbessern. Schauen wir uns dies für zwei Punkte $\mathbf{x}_i$ und $\mathbf{x}_j$ an. Es möge gelten $d_{ij} > \hat{d}_{ij}$. Der beobachtete Abstand d_{ij} ist also zu groß. Die beiden Punkte müssen näher beieinander liegen. Um das zu erreichen, müssen wir sie verschieben. Schauen wir uns dies aus Sicht des Punktes $\mathbf{x}_j$ an. Wir müssen den Punkt $\mathbf{x}_j$ zum Punkt $\mathbf{x}_i$ hinbewegen. hmcounterend. (fortgesetzt)

Example 29. Wir betrachten die Punkte

$$\mathbf{x}_2 = \begin{pmatrix} 2.8 \\ 1.6 \end{pmatrix}, \qquad \mathbf{x}_3 = \begin{pmatrix} -2.6 \\ 2.6 \end{pmatrix}.$$

Es gilt $d_{23} = 5.5$ und $\hat{d}_{23} = 4.7$. Wir müssen den Punkt $\mathbf{x}_3$ zum Punkt $\mathbf{x}_2$ hinbewegen. Abbildung 6.14 veranschaulicht das.

□

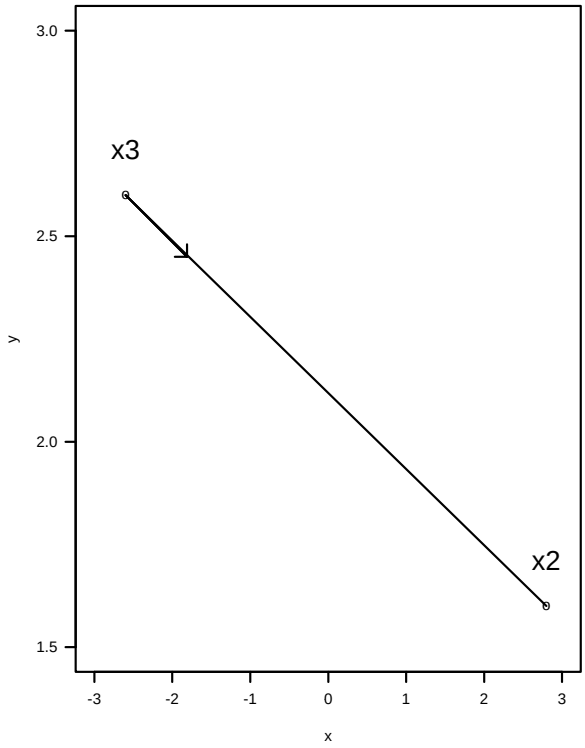


Fig. 6.14. Verschiebung des Punktes $\mathbf{x}_3$ in Richtung des Punktes $\mathbf{x}_2$

Wie weit soll man $\mathbf{x}_j$ in Richtung $\mathbf{x}_i$ verschieben? Es liegt nahe, $\mathbf{x}_j$ so weit in Richtung $\mathbf{x}_i$ zu verschieben, bis der Abstand zwischen beiden Punkten genau $\hat{d}_{ij}$ beträgt. Bezeichnen wir mit $\mathbf{x}^*_{j(i)}$ die neuen Koordinaten des Punkts $\mathbf{x}_j$ bezüglich $\mathbf{x}_i$, so muss für die erste Koordinate $x^*_{j1(i)}$ gelten:

$$x^*_{j1(i)} = x_{j1} + \frac{d_{ij} - \hat{d}_{ij}}{d_{ij}} (x_{i1} - x_{j1}).$$

Aufgrund des Strahlensatzes gilt nämlich

$$\frac{x^*_{j1(i)} - x_{j1}}{x_{i1} - x_{j1}} = \frac{d_{ij} - \hat{d}_{ij}}{d_{ij}},$$

wie man sich anhand der Abbildung 6.15 klar machen kann.
Eine analoge Beziehung gilt für die zweite Koordinate. Die neuen Koordinaten von $\mathbf{x}_j$ sind somit

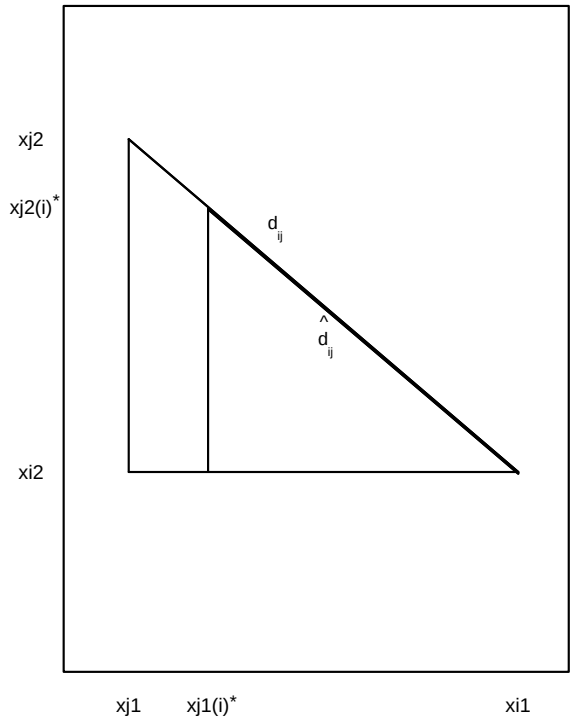


Fig. 6.15. Bestimmung des Umfangs der Verschiebung des Punktes $\mathbf{x}_j$ in Richtung des Punktes $\mathbf{x}_i$

$$\mathbf{x}^*_{j(i)} = \mathbf{x}_j + \frac{d_{ij} - \hat{d}_{ij}}{d_{ij}} (\mathbf{x}_i - \mathbf{x}_j). \tag{6.37}$$

hmcounterend. (fortgesetzt)

Example 29. Es gilt

$$\mathbf{x}^*_{3(2)} = \begin{pmatrix} -2.6 \\ 2.6 \end{pmatrix} + \frac{5.5 - 4.7}{5.5} \left(\begin{pmatrix} 2.8 \\ 1.6 \end{pmatrix} - \begin{pmatrix} -2.6 \\ 2.6 \end{pmatrix} \right) = \begin{pmatrix} -1.81 \\ 2.45 \end{pmatrix}.$$

□

Nachdem wir wissen, wie wir den Punkt $\mathbf{x}_j$ bezüglich des Punktes $\mathbf{x}_i$ verschieben müssen, bestimmen wir analog die Koordinaten von $\mathbf{x}_j$ bezüglich der anderen Punkte. hmcounterend. (fortgesetzt)

Example 29. Es gilt $d_{31} = 5.1$ und $\hat{d}_{31} = 4.7$, $d_{34} = 7.9$ und $\hat{d}_{34} = 7.9$ und $d_{35} = 5.8$ und $\hat{d}_{35} = 5.8$. Also müssen wir die Koordinaten von $\mathbf{x}_3$ nicht

bezüglich $\mathbf{x}_4$ und $\mathbf{x}_5$, sondern nur bezüglich $\mathbf{x}_1$ ändern. Es ergibt sich

$$\mathbf{x}^*_{3(1)} = \begin{pmatrix} -2.84 \\ 2.29 \end{pmatrix}, \qquad \mathbf{x}^*_{3(4)} = \begin{pmatrix} -2.6 \\ 2.6 \end{pmatrix}, \qquad \mathbf{x}^*_{3(5)} = \begin{pmatrix} -2.6 \\ 2.6 \end{pmatrix}.$$

□

Die neuen Koordinaten von $\mathbf{x}_j$ erhalten wir, indem wir den Mittelwert der Koordinaten bezüglich aller anderen Punkte bestimmen:

$$\mathbf{x}^*_j = \mathbf{x}_j + \frac{1}{n-1} \sum_{k \neq j} \frac{d_{jk} - \hat{d}_{jk}}{d_{jk}} (\mathbf{x}_j - \mathbf{x}_k).$$

hmcounterend. (fortgesetzt)

Example 29. Es gilt

$$\mathbf{x}^*_3 = \begin{pmatrix} -2.46 \\ 2.49 \end{pmatrix}.$$

Abbildung 6.16 zeigt die Verschiebungen in Bezug auf die anderen Punkte und die Gesamtverschiebung. Dabei ist die neue Position von $\mathbf{x}_3$ durch ein Dreieck gekennzeichnet.

□

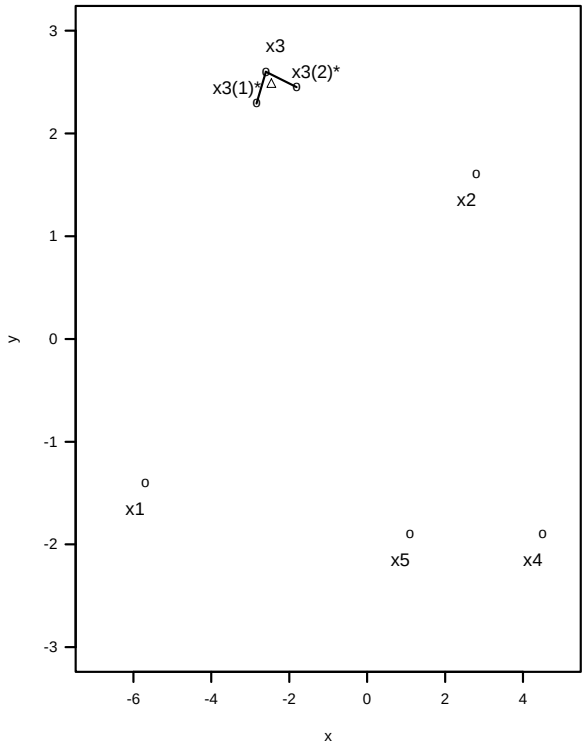


Fig. 6.16. Verschiebung eines Punktes in Bezug auf die anderen Punkte und die Gesamtverschiebung, die durch das Dreieck gekennzeichnet ist

Analog erhalten wir für alle anderen Punkte neue Koordinaten. hmcounterend. (fortgesetzt)

Example 29. Die neue Konfiguration lautet

$$\mathbf{X}^* = \begin{pmatrix} -5.64 & -1.33 \\ 2.61 & 1.64 \\ -2.46 & 2.49 \\ 4.78 & -0.82 \\ 0.82 & -1.98 \end{pmatrix}.$$

Abbildung 6.17 zeigt die Startkonfiguration und die Konfiguration, die sich nach der ersten Iteration ergibt. Dabei steht ein `a` für einen Punkt der alten Konfiguration. Der daneben mit `n` bezeichnete Punkt gehört zur neuen Konfiguration.

□

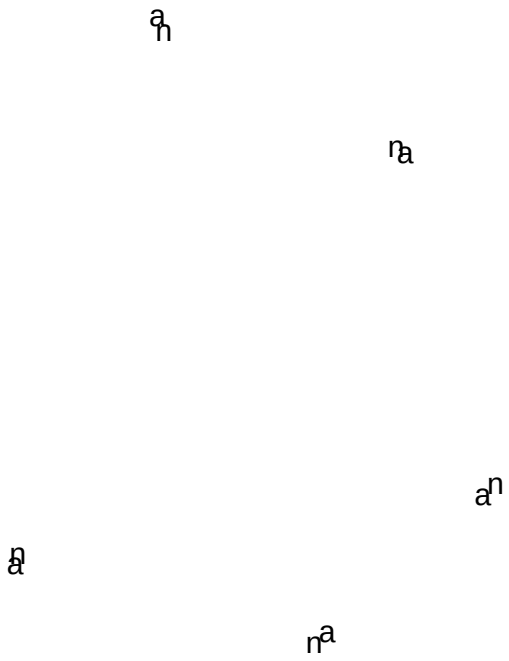


Fig. 6.17. Startkonfiguration und die sich nach der ersten Iteration ergebende Konfiguration

Für die entstandene Konfiguration können wir den Wert von STRESS1 berechnen. Sind wir mit diesem noch nicht zufrieden, so bestimmen wir eine neue Konfiguration. Diese Schritte wiederholen wir so lange, bis wir eine Konfiguration gefunden haben, die einen akzeptablen Wert von STRESS1 besitzt. Man kann diesen Algorithmus noch modifizieren. Schauen wir uns dazu noch einmal (6.37) an:

$$\mathbf{x}^*_{j(i)} = \mathbf{x}_j + \frac{d_{ij} - \hat{d}_{ij}}{d_{ij}} (\mathbf{x}_i - \mathbf{x}_j).$$

Wir können diese Formel auch so interpretieren, dass wir von $\mathbf{x}_j$ in Richtung $\mathbf{x}_i$ gehen, wobei die Schrittweite vom Unterschied zwischen beobachteter und gewünschter Distanz abhängt. In (6.37) ist die Schrittweite so gewählt, dass die gewünschte Distanz genau erreicht wird. Die Modifikation besteht

nun darin, nicht diese Schrittweite zu wählen, sondern ein Vielfaches dieser Schrittweite.

Man bildet also

$$\mathbf{x}^*_{j(i)} = \mathbf{x}_j + \alpha \frac{d_{ij} - \hat{d}_{ij}}{d_{ij}} (\mathbf{x}_i - \mathbf{x}_j),$$

wobei α eine positive reelle Zahl ist. Die neuen Koordinaten eines Punktes erhält man also durch

$$\mathbf{x}^*_i = \mathbf{x}_i + \alpha \frac{1}{n-1} \sum_{j \neq i} \frac{d_{ij} - \hat{d}_{ij}}{d_{ij}} (\mathbf{x}_j - \mathbf{x}_i). \tag{6.38}$$

Es gibt noch eine Reihe weiterer Algorithmen zur Bestimmung der Konfiguration. Der *SMACOF-Algorithmus* ist bei Borg & Groenen (1997) und Cox & Cox (1994) detailliert beschrieben. Kearsley et al. (1998) untersuchen das *Newton-Verfahren*. Trippel (2001) hat in seiner Diplomarbeit diese Verfahren in S-PLUS implementiert und verglichen.

6.3.2 Nichtmetrische mehrdimensionale Skalierung in S-PLUS

In S-PLUS sind keine Funktionen zur nichtmetrischen mehrdimensionalen Skalierung implementiert. Wir wollen im Folgenden die Vorgehensweise des Algorithmus von Kruskal programmieren. Schauen wir uns die Daten in Tabelle 4.8 auf Seite 111 an. Wir haben auf Seite 179 die Distanzmatrix `dpol` in S-PLUS erzeugt. Wir weisen die Werte, die unterhalb der Hauptdiagonalen von `dpol` stehen, einem Vektor `delta` zu. Hierzu verwenden wir die Funktion `lower.tri`:

```
> delta<-dpol[lower.tri(dpol)]
> delta
 [1]  9  4 10  7  3  1  2  8  6  5
```

Wir bestimmen die Startkonfiguration mit Hilfe einer metrischen mehrdimensionalen Skalierung und runden die Zahlen auf eine Stelle nach dem Komma, um die gleichen Zahlen wie im Text zu benutzen:

```
> X<-round(cmdscale(dpol),1)
> X
      [,1] [,2]
 [1,] -5.7 -1.4
 [2,]  2.8  1.6
 [3,] -2.6  2.6
 [4,]  4.5 -0.9
 [5,]  1.1 -1.9
```

Nun bestimmen wir die euklidischen Distanzen zwischen den Zeilen der Matrix `X` und runden auch sie auf eine Stelle nach dem Komma:

```
> d<-round(dist(X),1)
> d
  [1]  9.0  5.1 10.2  6.8  5.5  3.0  3.9  7.9  5.8  3.5
 attr(, "Size"):
 [1] 5
```

Nun müssen wir die Elemente des Vektors `d` so anordnen, dass sie mit der Anordnung im Vektor `delta` übereinstimmen. Hierzu müssen wir zuerst bestimmen, an welcher Stelle in `delta` das kleinste, das zweitkleinste,... Element steht. Dies leistet die Funktion `order`:

```
> order(delta)
 [1]  6  7  5  2 10  9  4  8  1  3
```

Die kleinste Zahl im Vektor `delta` steht also an der sechsten Stelle, die zweitkleinste an der siebten Stelle, u.s.w.. Wir indizieren `d` mit diesem Vektor:

```
> d[order(delta)]
 [1]  3.0  3.9  5.5  5.1  3.5  5.8  6.8  7.9  9.0 10.2
```

Die Monotoniebedingung ist verletzt. Um die Disparitäten und die Werte von STRESS1 bestimmen zu können, müssen wir eine monotone Regression durchführen. Im Anhang B ist auf Seite 496 eine Funktion `monreg` zu finden, die eine monotone Regression durchführt. Wir rufen diese Funktion auf:

```
> disp<-monreg(d[order(delta)])
> disp
 [1]  3.0  3.9  4.7  4.7  4.7  5.8  6.8  7.9  9.0 10.2
```

Nun müssen wir nur noch die Disparitäten in die richtige Reihenfolge bringen:

```
> disp<-disp[delta]
> attr(disp,"Size")<-dim(X)[1]
```

Im Anhang B ist auf Seite 497 die Funktion `stress1` zu finden, die STRESS1 bestimmt. Wir rufen die Funktion `stress1` auf:

```
> stress1(d,disp)
[1] 0.07302533
```

Bei der Bestimmung der neuen Konfiguration berücksichtigen wir, dass `S-PLUS` eine matrizenorientierte Sprache ist. Wir erzeugen uns eine Matrix `dm` mit den Distanzen zwischen den Punkten der aktuellen Konfiguration und eine Matrix `dispm` mit den Disparitäten. Hierzu benutzen wir die Funktion `distfull` auf Seite 495:

```
> dm<-distfull(d)
> dispm<-distfull(disp)
```

Mit Hilfe dieser Matrizen können wir die neuen Koordinaten der Punkte schnell bestimmen. Schauen wir uns dies exemplarisch für den dritten Punkt an. Wir müssen zunächst den Quotienten

$$\frac{d_{3j} - \hat{d}_{3j}}{d_{3j}}$$

für jedes $j \neq 3$ bestimmen. Wir benötigen von den Matrizen `dm` und `dispm` die Elemente in der dritten Zeile ohne das dritte Element. Indiziert man mit einem negativen Skalar oder einem Vektor negativer Zahlen, so werden die entsprechenden Komponenten nicht ausgewählt. Wir geben also ein

```
> (dm[3,-3]-dispm[3,-3])/dm[3,-3]
[1] 0.07843136 0.14545455 0.00000000 0.00000000
```

Nun benötigen wir noch die Differenz $\mathbf{x}_i - \mathbf{x}_3$ für $i = 1, 2, 4, 5$. Hierzu erzeugen wir eine Matrix, die viermal die dritte Zeile von `X` enthält, und subtrahieren sie von der Matrix, die wir erhalten, wenn wir die dritte Zeile von `X` streichen:

```
> matrix(X[3,],4,2,b=T)-X[-3,]
      [,1] [,2]
 [1,] -3.1 -4.0
 [2,]  5.4 -1.0
 [3,]  7.1 -3.5
 [4,]  3.7 -4.5
```

Nun müssen wir noch die beiden Größen zusammenbringen. Auch dieses geht wiederum matriziell:

```
> m<-matrix(X[3,],4,2,b=T)+matrix((dm[3,-3]-dispm[3,-3])/
     dm[3,-3],4,2)*(X[-3,]-matrix(X[3,],4,2,b=T))
> m
           [,1]     [,2]
 [1,] -2.843137 2.286275
 [2,] -1.814545 2.454545
 [3,] -2.600000 2.600000
 [4,] -2.600000 2.600000
```

Jetzt müssen nur noch die Mittelwerte der Spalten bestimmt werden, um die neuen Koordinaten von $\mathbf{x}_3$ zu gewinnen:

```
> apply(m,2,mean)
[1] -2.464421  2.485205
```

Wir erhalten das bereits manuell bestimmte Ergebnis. Die neue Konfiguration aller Punkte bestimmen wir dann mit einer Iteration. Da die gleiche Befehlsfolge 5-mal mit unterschiedlichen Werten ausgeführt werden soll, bilden wir eine `for`-Schleife:

```
> xneu<-matrix(0,5,2)
> for(i in 1:5){
> m<-matrix(X[i,],4,2,b=T)+matrix((dm[i,-i]-dispm[i,-i])/
     dm[i,-i],4,2)*(X[-i,]-matrix(X[i,],4,2,b=T))
> xneu[i,]<-apply(m,2,mean)}
> xneu
            [,1]       [,2]
 [1,] -5.6392157 -1.3215686
 [2,]  2.6036364  1.6363636
 [3,] -2.4644207  2.4852050
 [4,]  4.7914286 -0.8142857
 [5,]  0.8085714 -1.9857143
```

Im Anhang B ist auf Seite 497 eine Funktion `Neuekon` zu finden, die die neue Konfiguration in Abhängigkeit von der Startkonfiguration `X` bestimmt.

6.4 Ergänzungen und weiterführende Literatur

Wir haben in diesem Kapitel beschrieben, wie man die Distanzen einer Distanzmatrix in einem niedrigdimensionalen Raum darstellt. Oft werden mehrere Personen gebeten, die Distanzen zwischen mehreren Objekten anzugeben. Gesucht ist in diesem Fall ebenfalls eine Darstellung der Objekte in einem niedrigdimensionalen Raum. Um diese zu erreichen, kann man die mittlere Distanz jedes Objektpaars bestimmen und eine mehrdimensionale Skalierung dieser mittleren Distanzen durchführen. Hierbei werden aber die Unterschiede zwischen den einzelnen Personen, die die Bewertung abgeben, nicht in Betracht gezogen. Das Verfahren INDSCAL von Carroll & Chang (1970) berücksichtigt diesen Aspekt. Es ist bei Borg & Groenen (1997), Cox & Cox (1994) und Davison (1983) detailliert beschrieben.

6.5 Übungen

Exercise 12. Im Beispiel 6 auf Seite 7 wurden Reisezeiten verglichen, die mit unterschiedlichen Verkehrsmitteln innerhalb Deutschlands benötigt werden. Führen sie für die Reisezeiten der Pkws und für die Reisezeiten der Bahn jeweils eine metrische und eine nichtmetrische mehrdimensionale Skalierung durch.

Exercise 13. Führen Sie für die Beispiele 25, 26 und 27 auf den Seiten 155, 160 und 160 jeweils eine metrische mehrdimensionale Skalierung in `S-PLUS` durch.

Exercise 14. Ein Student wurde gebeten, 10 Paare von Lebensmittelmärkten der Ähnlichkeit nach zu ordnen, sodass das ähnlichste Paar den Wert 1, das zweitähnlichste den Wert 2, u.s.w. erhält. Tabelle 6.2 zeigt die Ergebnisse.

Table 6.2. Vergleich aller Paare aus einer Menge von 5 Lebensmittelmärkten mit dem Verfahren der Rangreihung

1. Lebensmittelmarkt	2. Lebensmittelmarkt	Rang
ALDI	LIDL	1
ALDI	MARKTKAUF	6
ALDI	REAL	8
ALDI	EDEKA	10
LIDL	MARKTKAUF	5
LIDL	REAL	9
LIDL	EDEKA	7
MARKTKAUF	REAL	2
MARKTKAUF	EDEKA	4
REAL	EDEKA	3

Der Student will eine zweidimensionale Darstellung mit Hilfe einer metrischen mehrdimensionalen Skalierung gewinnen.

1. Führen Sie eine metrische mehrdimensionale Skalierung der Daten durch.
2. Die Koordinaten der Punkte der metrischen mehrdimensionalen Skalierung sind in Tabelle 6.3 zu finden.

Table 6.3. Koordinaten von 5 Punkten

Geschäft	1. Koordinate	2. Koordinate
ALDI	5.0	-1.7
LIDL	3.9	2.3
MARKTKAUF	-1.1	-0.6
REAL	-3.8	-2.3
EDEKA	-4.1	2.4

Die Matrix der euklidischen Distanzen zwischen den Punkten ist

$$\mathbf{D} = \begin{pmatrix} 0.0 & 4.1 & 6.2 & 8.8 & 10.0 \\ 4.1 & 0.0 & 5.8 & 9.0 & 8.0 \\ 6.2 & 5.8 & 0.0 & 3.2 & 4.2 \\ 8.8 & 9.0 & 3.2 & 0.0 & 4.7 \\ 10.0 & 8.0 & 4.2 & 4.7 & 0.0 \end{pmatrix}.$$

Wir wollen diese Konfiguration als Ausgangspunkt für eine nichtmetrische mehrdimensionale Skalierung wählen.

a) Führen Sie eine monotone Regression durch.
b) Bestimmen Sie den Wert von STRESS1.
c) Bestimmen Sie die neuen Koordinaten aller Punkte.

Exercise 15. Führen Sie für das Beispiel 25 auf der Seite 155 eine metrische mehrdimensionale Skalierung in `S-PLUS` durch, wobei Sie nicht die Funktion `cmdscale` anwenden, sondern die einzelnen Schritte auf Seite 171 in `S-PLUS` nachvollziehen sollten.

7 Procrustes-Analyse

7.1 Problemstellung und Grundlagen

Da das Ergebnis einer mehrdimensionalen Skalierung nicht eindeutig ist, sind unterschiedliche Konfigurationen nicht leicht zu vergleichen. Wir haben das beim Vergleich der durch eine Hauptkomponentenanalyse und metrische mehrdimensionale Skalierung gewonnenen Konfigurationen auf Seite 175 gesehen. Erst nachdem man eine Konfiguration um 180 Grad gedreht hatte, konnte man erkennen, dass die beiden Konfigurationen identisch sind. Oft reicht eine Drehung nicht aus. Man muss auch verschieben und strecken oder stauchen. Bei der direkten Bestimmung der Distanzen im Kapitel 4.4 auf Seite 110 haben wir zwei unterschiedliche Methoden betrachtet, die Ähnlichkeit zwischen Objekten zu bestimmen. Bei beiden Methoden wurden alle Paare von n Personen betrachtet. Beim Ratingverfahren wurde eine Person gebeten, die Ähnlichkeit jedes Paares auf einer Skala von 1 bis 7 zu bewerten, wobei das Paar den Wert 1 erhält, wenn sich die beiden Personen sehr ähnlich sind, und den Wert 7, wenn sich die Personen sehr unähnlich sind. Die Rangreihung hingegen besteht darin, die 10 Paarvergleiche zwischen 5 Personen nach Ähnlichkeit der Größe nach zu ordnen. Dabei erhält das ähnlichste Paar eine 1 und das unähnlichste eine 10.

Example 30. Bei beiden Verfahren wurden die fünf deutschen Politiker Joschka Fischer, Angela Merkel, Gerhard Schröder, Edmund Stoiber und Guido Westerwelle betrachtet. Die Bewertung dieser Politiker mit dem Ratingverfahren ist in Tabelle 4.6 auf Seite 110 und die Bewertung mit der Rangreihung in Tabelle 4.7 auf Seite 111 zu finden. Für beide Situationen wurde eine metrische mehrdimensionale Skalierung durchgeführt. Die Darstellungen sind in den Abbildungen 7.1 und 7.2 zu finden.

Wir wollen nun die beiden Konfigurationen miteinander vergleichen, um zu sehen, wie konsistent der Student bei der Bewertung ist. Da die Konfigurationen beliebig verschoben, gedreht und gestreckt oder gestaucht werden können, sollte man vor dem Vergleich eine von beiden so verschieben, drehen und strecken oder stauchen, dass sie der anderen ähnelt. Man spricht von *Procrustes-Analyse*. So heißt in der griechischen Sage ein Räuber, der seine Gefangenen in sein Bett legte. Waren sie zu klein, so wurden sie gestreckt, waren sie zu groß, so wurden sie gekürzt. In der römischen Mythologie trägt er

Fig. 7.1. Graphische Darstellung von 5 Politikern mit einer metrischen mehrdimensionalen Skalierung auf der Basis einer Distanzmatrix, die mit der Rangreihung gewonnen wurde

den Namen Damastes. In Abbildung 7.3 sind die Konfigurationen dargestellt, nachdem die Konfiguration der mit dem Ratingverfahren gewonnenen Distanzen der Konfiguration, die aus den Distanzen der Rangreihung gewonnen wurde, mit einer Procrustes-Analyse ähnlich gemacht wurde. Die Namen der Politiker sind beim Ratingverfahren mit Großbuchstaben geschrieben. Wir sehen, dass der Student bei der Bewertung der Politiker im rechten Bereich der Zeichnung konsistent ist. Nur Fischer und Schröder werden bei den beiden Verfahren unterschiedlich bewertet. Dies war beim Vergleich der Abbildungen 7.1 und 7.2 nicht zu erkennen.

□

Es stellt sich die Frage, wie man eine Konfiguration systematisch so verschieben, drehen und strecken oder stauchen kann, dass sie einer anderen

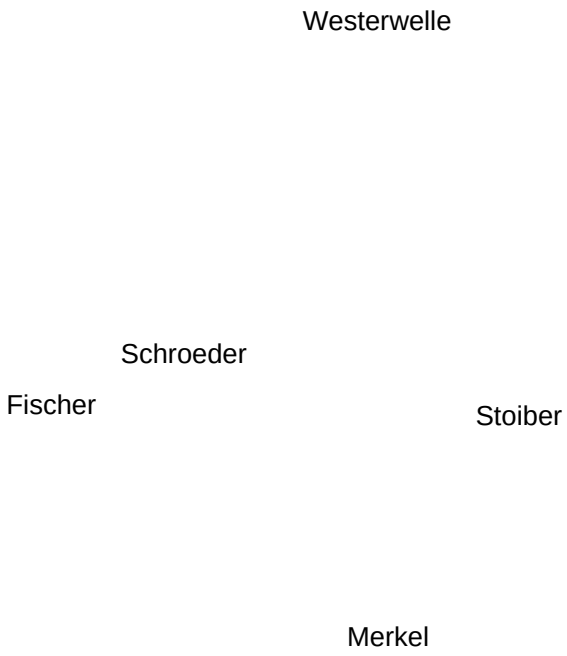


Fig. 7.2. Graphische Darstellung von 5 Politikern mit einer metrischen mehrdimensionalen Skalierung auf der Basis einer Distanzmatrix, die mit dem Ratingverfahren gewonnen wurde

Konfiguration möglichst ähnlich ist. Im nächsten Abschnitt werden wir an einem kleinen Beispiel illustrieren, wie das funktioniert.

7.2 Illustration der Vorgehensweise

Wir gehen aus von zwei Konfigurationen von n Punkten aus dem $\mathbb{R}^k$. Die Punkte der ersten Konfiguration seien $\mathbf{a}_1, \ldots, \mathbf{a}_n$ mit

$$\mathbf{a}_i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ik} \end{pmatrix}$$

für $i = 1, \ldots, n$. Entsprechendes gilt für die Punkte $\mathbf{b}_1, \ldots, \mathbf{b}_n$ der zweiten Konfiguration

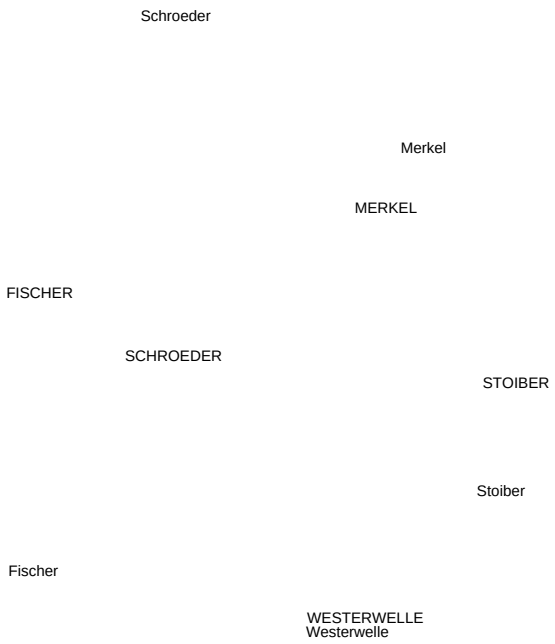


Fig. 7.3. Graphische Darstellung von 5 Politikern mit einer metrischen mehrdimensionalen Skalierung auf der Basis einer Distanzmatrix, die mit dem Verfahren der Rangreihung gewonnen wurde, nach Durchführung einer Procrustes-Analyse

$$\mathbf{b}_i = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{ik} \end{pmatrix}$$

für $i = 1, \ldots, n$. Die Zeilenvektoren $\mathbf{a}'_1, \ldots, \mathbf{a}'_n$ bilden die Zeilen der Matrix $\mathbf{A}$ und die Zeilenvektoren $\mathbf{b}'_1, \ldots, \mathbf{b}'_n$ die Zeilen der Matrix $\mathbf{B}$.

Example 31. Wir gehen aus von zwei Konfigurationen, die aus jeweils drei Punkten im $\mathbb{R}^2$ bestehen.

Die erste Konfiguration besteht aus den Punkten

$$\mathbf{a}_1 = \begin{pmatrix} 10 \\ 2 \end{pmatrix}, \qquad \mathbf{a}_2 = \begin{pmatrix} 10 \\ 8 \end{pmatrix}, \qquad \mathbf{a}_3 = \begin{pmatrix} 4 \\ 8 \end{pmatrix}.$$

Die zweite Konfiguration setzt sich zusammen aus den Punkten

$$\mathbf{b}_1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \qquad \mathbf{b}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \qquad \mathbf{b}_3 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}.$$

Es gilt

$$\mathbf{A} = \begin{pmatrix} 10 & 2 \\ 10 & 8 \\ 4 & 8 \end{pmatrix}$$

und

$$\mathbf{B} = \begin{pmatrix} 2 & 4 \\ 2 & 1 \\ 5 & 1 \end{pmatrix}.$$

Abbildung 7.4 zeigt die beiden Konfigurationen, wobei wir die Zusammengehörigkeit der Punkte einer Konfiguration dadurch hervorheben, dass wir die Punkte durch Strecken miteinander verbinden.
Wir sehen, dass die beiden Konfigurationen ähnlich sind. Beide bilden rechtwinklige Dreiecke. Es sieht auch so aus, als ob das Verhältnis der Längen der Katheten bei beiden Dreiecken sehr ähnlich ist. □

Die beiden Konfigurationen unterscheiden sich durch ihre Lage im Koordinatensystem, ihre Größe und ihre Ausrichtung. Alle drei Aspekte sind aber irrelevant, wenn man an der Lage der Punkte einer Konfiguration zueinander interessiert ist. Wir können eine Konfiguration von Punkten also verschieben, strecken oder stauchen und drehen, ohne dass relevante Information verlorengeht. Wir wollen nun die zweite Konfiguration so verschieben, strecken oder stauchen und drehen, dass sie hinsichtlich Lage, Größe und Ausrichtung der ersten Konfiguration so ähnlich wie möglich ist.
Beginnen wir mit der Verschiebung. Wir betrachten hierzu die Zentren der beiden Konfigurationen. Es liegt nahe die Mittelwerte der beiden Koordinaten zu bilden. Wir bilden also

$$\bar{\mathbf{a}} = \sum_{i=1}^{n} \mathbf{a}_i \tag{7.1}$$

und

$$\bar{\mathbf{b}} = \sum_{i=1}^{n} \mathbf{b}_i. \tag{7.2}$$

hmcounterend. (fortgesetzt)

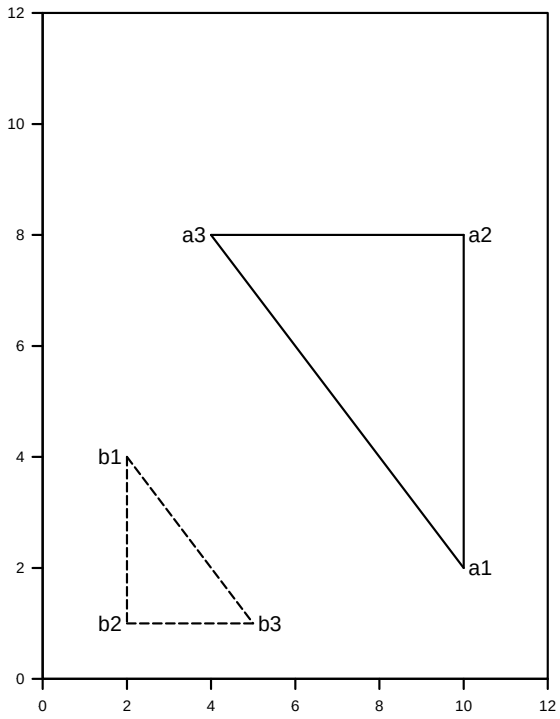


Fig. 7.4. Zwei Konfigurationen, die aus jeweils 3 Punkten bestehen

Example 31. Es gilt

$$\bar{\mathbf{a}} = \frac{1}{3}\left[\begin{pmatrix}10\\2\end{pmatrix} + \begin{pmatrix}10\\8\end{pmatrix} + \begin{pmatrix}4\\8\end{pmatrix}\right] = \begin{pmatrix}8\\6\end{pmatrix}.$$

Entsprechend erhalten wir

$$\bar{\mathbf{b}} = \begin{pmatrix}3\\2\end{pmatrix}.$$

□

Die zweite Konfiguration soll der ersten möglichst ähnlich gemacht werden. Sie sollte also das gleiche Zentrum wie die erste besitzen. Die anderen Operationen sind einfacher zu verstehen, wenn man sie auf Konfigurationen anwendet, deren Zentrum im Nullpunkt liegt. Wir verschieben die beiden Konfigurationen so, dass ihr Zentrum jeweils im Nullpunkt liegt. Wir bilden also

$$\tilde{\mathbf{a}}_i = \mathbf{a}_i - \bar{\mathbf{a}}$$

und

$$\tilde{\mathbf{b}}_i = \mathbf{b}_i - \bar{\mathbf{b}}.$$

Die Koordinaten der zentrierten Punkte mögen die Zeilenvektoren der Matrizen $\tilde{\mathbf{A}}$ und $\tilde{\mathbf{B}}$ bilden. hmcounterend. (fortgesetzt)

Example 31. Es gilt

$$\tilde{\mathbf{a}}_1 = \begin{pmatrix} 2 \\ -4 \end{pmatrix}, \qquad \tilde{\mathbf{a}}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \qquad \tilde{\mathbf{a}}_3 = \begin{pmatrix} -4 \\ 2 \end{pmatrix}$$

und

$$\tilde{\mathbf{b}}_1 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \qquad \tilde{\mathbf{b}}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \qquad \tilde{\mathbf{b}}_3 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Es gilt

$$\tilde{\mathbf{A}} = \begin{pmatrix} 2 & -4 \\ 2 & 2 \\ -4 & 2 \end{pmatrix}, \qquad \tilde{\mathbf{B}} = \begin{pmatrix} -1 & 2 \\ -1 & -1 \\ 2 & -1 \end{pmatrix}.$$

Abbildung 7.5 zeigt die beiden verschobenen Konfigurationen.

□

Wenden wir uns den Drehungen zu. hmcounterend. (fortgesetzt)

Example 31. In Abbildung 7.5 handelt es sich bei beiden Konfigurationen um rechtwinklige Dreiecke. Wir sehen, dass die rechten Winkel sich genau gegenüberliegen, wenn man den Nullpunkt als Bezugspunkt nimmt. Es liegt nahe, die zweite Konfiguration um 180 Grad im Gegenzeigersinn zu drehen. Abbildung 7.6 zeigt, dass dieses Vorgehen richtig ist.

□

Eine Konfiguration von Punkten im $\mathbb{R}^2$, die die Zeilenvektoren der Matrix $\mathbf{C}$ bilden, dreht man um den Winkel α in Gegenzeigersinn, indem man sie von rechts mit der *Rotationsmatrix*

$$\mathbf{T} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

multipliziert. Man bildet also

$$\mathbf{C}\,\mathbf{T}.$$

Eine Begründung dieses Sachverhalts ist bei Zurmühl & Falk (1997), S. 7 zu finden. hmcounterend. (fortgesetzt)

Example 31. Im Beispiel ist α gleich π. Wegen $\cos\pi = -1$ und $\sin\pi = 0$ gilt also

$$\mathbf{T} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

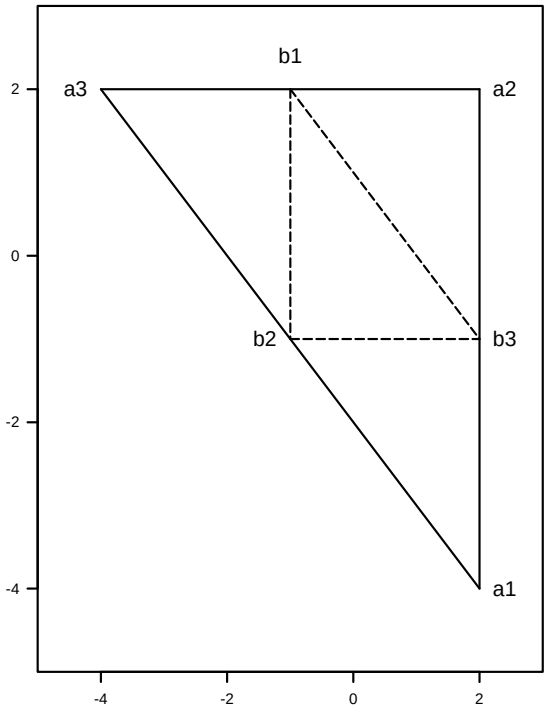


Fig. 7.5. Zwei Konfigurationen, deren Zentrum der Nullpunkt ist

Wir wenden die Matrix $\mathbf{T}$ auf die Matrix $\tilde{\mathbf{B}}$ an:

$$\tilde{\mathbf{B}}_d = \tilde{\mathbf{B}}\,\mathbf{T} = \begin{pmatrix} 1 & -2 \\ 1 & 1 \\ -2 & 1 \end{pmatrix}.$$

□

Jetzt müssen wir nur noch die Größe der Konfiguration verändern. Hierzu multiplizieren wir alle Punkte mit der Zahl c und erhalten die Matrix

$$\tilde{\mathbf{B}}_{dm} = c\,\tilde{\mathbf{B}}_d.$$

hmcounterend. (fortgesetzt)

Example 31. Schauen wir uns die Koordinaten aller Punkte aus Abbildung 7.6 an. Die Koordinaten der Punkte der zu verändernden Konfiguration sind die Zeilenvektoren der Matrix

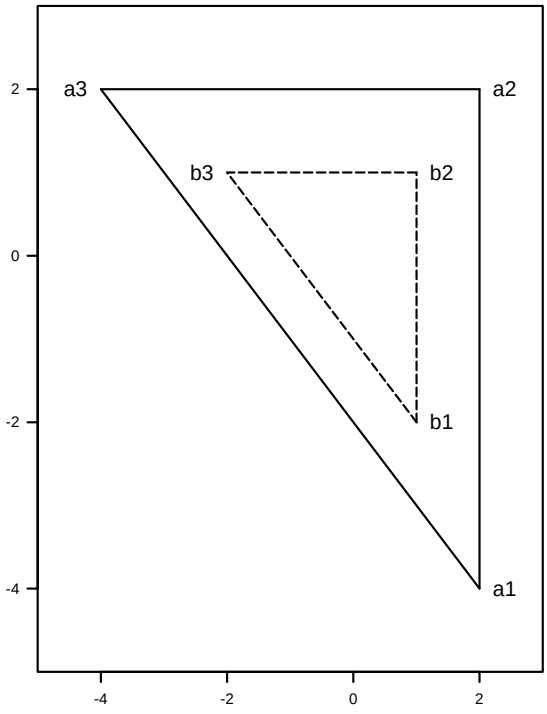


Fig. 7.6. Zwei Konfigurationen nach Verschiebung und Drehung

$$\tilde{\mathbf{B}}_d = \begin{pmatrix} 1 & -2 \\ 1 & 1 \\ -2 & 1 \end{pmatrix}.$$

Die Koordinaten der Punkte der anderen Konfiguration bilden die Zeilenvektoren der Matrix

$$\tilde{\mathbf{A}} = \begin{pmatrix} 2 & -4 \\ 2 & 2 \\ -4 & 2 \end{pmatrix}.$$

Wir sehen, dass wir die Matrix $\tilde{\mathbf{B}}_d$ nur mit 2 multiplizieren müssen, um die Matrix $\tilde{\mathbf{A}}$ zu erhalten:

$$\tilde{\mathbf{B}}_{dm} = 2\,\tilde{\mathbf{B}}_d = 2 \begin{pmatrix} 1 & -2 \\ 1 & 1 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -4 \\ 2 & 2 \\ -4 & 2 \end{pmatrix}.$$

□

Nun müssen wir nur noch die Lage der beiden Konfigurationen so verändern, dass beide das Zentrum der ersten Konfiguration besitzen. Hierzu addieren wir zu den Zeilenvektoren der Matrix $\tilde{\mathbf{A}}$ und den Zeilenvektoren der Matrix $\tilde{\mathbf{B}}_{dm}$ den Vektor $\bar{\mathbf{a}}'$.

7.3 Theorie

Im Beispiel 31 ist es möglich, die zweite Konfiguration so zu verschieben, zu drehen und zu strecken, dass sie mit der ersten Konfiguration zusammenfällt. Außerdem konnte man durch Hinschauen erkennen, wie man die zweite Konfiguration drehen und strecken mußte. Bei Konfigurationen, die mit Hilfe einer mehrdimensionalen Skalierung gewonnen wurden, wird es in der Regel nicht möglich sein, die beiden Konfigurationen vollständig zur Deckung zu bringen. In diesem Fall wird man fordern, dass sie sich sehr ähnlich sind. Ein Maß für die Ähnlichkeit von zwei Konfigurationen im $\mathbb{R}^k$, die aus jeweils n Punkten bestehen und die die Zeilenvektoren der Matrizen $\mathbf{A}$ und $\mathbf{B}$ bilden, ist

$$\sum_{i=1}^{n} \sum_{j=1}^{k} (a_{ij} - b_{ij})^2.$$

In Seber (1984), S. 253-256 wird hergeleitet, wie man eine Konfiguration verschieben, drehen und strecken oder stauchen muss, um sie einer anderen Konfiguration hinsichtlich dieses Kriteriums möglichst ähnlich zu machen. Wir wollen auf die Herleitung der zugrunde liegenden Beziehungen nicht eingehen, sondern nur zeigen, wie man vorgehen muss.

Ausgangspunkt sind die (n, k)-Matrizen $\mathbf{A}$ und $\mathbf{B}$. Die Zeilenvektoren dieser Matrizen bilden Konfigurationen im $\mathbb{R}^k$. Die Konfiguration in $\mathbf{B}$ soll der Konfiguration in $\mathbf{A}$ möglichst ähnlich gemacht werden.

Example 32. Wir gehen wieder aus von den Matrizen

$$\mathbf{A} = \begin{pmatrix} 10 & 2 \\ 10 & 8 \\ 4 & 8 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 2 & 4 \\ 2 & 1 \\ 5 & 1 \end{pmatrix}.$$

□

Um die Konfigurationen bezüglich der Lage möglichst ähnlich zu machen, legt man das Zentrum jeder Konfiguration in den Nullpunkt. Man bildet also entsprechend Gleichung (2.10) die zentrierten Matrizen $\tilde{\mathbf{A}}$ und $\tilde{\mathbf{B}}$. hmcounterend. (fortgesetzt)

Example 32. Es gilt

$$\tilde{\mathbf{A}} = \begin{pmatrix} 2 & -4 \\ 2 & 2 \\ -4 & 2 \end{pmatrix}, \qquad \tilde{\mathbf{B}} = \begin{pmatrix} -1 & 2 \\ -1 & -1 \\ 2 & -1 \end{pmatrix}.$$

□

Die Rotationsmatrix $\mathbf{T}$ und den Streckungs- beziehungsweise Stauchungsfaktor c erhält man durch eine Singulärwertzerlegung der Matrix $\tilde{\mathbf{A}}'\tilde{\mathbf{B}}$. Man bildet also folgende Zerlegung:

$$\tilde{\mathbf{A}}'\tilde{\mathbf{B}} = \mathbf{UDV}'. \tag{7.3}$$

Die Singulärwertzerlegung wird in Kapitel A.1.10 auf Seite 480 besprochen.

Die Rotationsmatrix $\mathbf{T}$ erhält man durch

$$\mathbf{T} = \mathbf{VU}'$$

und den Streckungs- beziehungsweise Stauchungsfaktor c durch

$$c = \frac{tr(\mathbf{D})}{tr(\tilde{\mathbf{B}}\tilde{\mathbf{B}}')}.$$

hmcounterend. (fortgesetzt)

Example 32. Es gilt

$$\tilde{\mathbf{A}}'\tilde{\mathbf{B}} = \mathbf{UDV}$$

mit

$$\mathbf{U} = \frac{1}{2}\begin{pmatrix} -\sqrt{2} & \sqrt{2} \\ \sqrt{2} & \sqrt{2} \end{pmatrix}, \quad \mathbf{V} = \frac{1}{2}\begin{pmatrix} \sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & -\sqrt{2} \end{pmatrix}$$

und

$$\mathbf{D} = \begin{pmatrix} 18 & 0 \\ 0 & 6 \end{pmatrix}.$$

Es gilt

$$\mathbf{T} = \mathbf{V}\mathbf{U}' = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Mit

$$\tilde{\mathbf{B}}\tilde{\mathbf{B}}' = \begin{pmatrix} 5 & -1 & -4 \\ -1 & 2 & -1 \\ -4 & -1 & 5 \end{pmatrix}$$

folgt

$$c = \frac{tr(\mathbf{D})}{tr(\tilde{\mathbf{B}}\tilde{\mathbf{B}}')} = \frac{24}{12} = 2.$$

□

7.4 Procrustes-Analyse der Reisezeiten

Wir wollen eine Procrustes-Analyse der Daten im Beispiel 6 auf Seite 7 durchführen. Hierzu führen wir zuerst eine metrische mehrdimensionale Skalierung der Reisezeiten mit dem Pkw durch. Die Daten sind in Tabelle 1.6 auf Seite 7 zu finden. Abbildung 7.7 zeigt die Darstellung der Städte, die man mit der metrischen mehrdimensionalen Skalierung erhält.
Nun schauen wir uns die Darstellung an, die man für die Reisezeiten mit der Bahn mit der metrischen mehrdimensionalen Skalierung gewinnt. Die Daten sind in Tabelle 1.7 auf Seite 7 zu finden. Abbildung 7.8 zeigt die Darstellung der Städte, die man mit der metrischen mehrdimensionalen Skalierung erhält.
Nun führen wir eine Procrustes-Analyse durch, wobei wir die Darstellung der Reisezeiten mit der Bahn der Darstellung der Reisezeiten mit dem Pkw ähnlich machen. Das Ergebnis ist in Abbildung 7.9 zu finden. Dabei wurden Großbuchstaben für die Städte beim Pkw und Kleinbuchstaben bei der Bahn gewählt. Wir sehen, dass die Reise zwischen Frankfurt und Köln mit dem Pkw viel schneller geht, während eine Fahrt von Frankfurt nach München mit dem Pkw ungefähr so lange dauert wie mit der Bahn.

7.5 Procrustes-Analyse in S-PLUS

Eine Funktion, mit der man eine Konfiguration so verschieben, drehen und strecken kann, dass sie einer anderen Konfiguration möglichst ähnlich ist, ist die Funktion `procrustes`. Sie wird aufgerufen durch

```
procrustes(amat,target,orthogonal=F,translate=F, magnify=F)
```

Dabei ist `amat` die Konfiguration, die verändert werden soll. Die Konfiguration, der die Konfiguration in der Matrix `amat` ähnlich gemacht werden

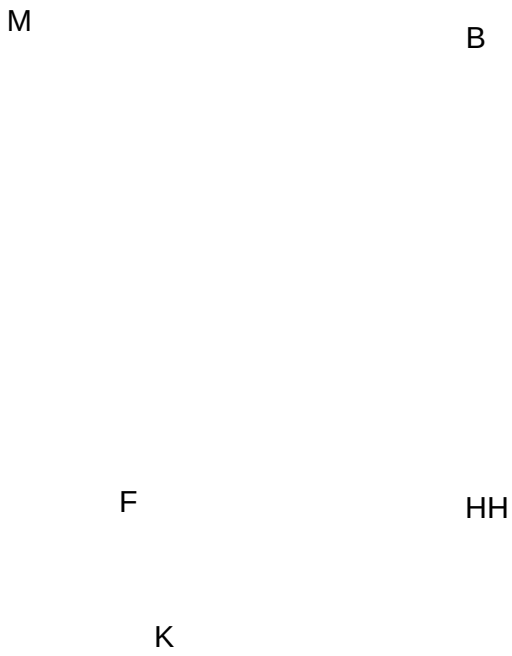


Fig. 7.7. Graphische Darstellung von 5 deutschen Städten mit einer metrischen mehrdimensionalen Skalierung auf der Basis einer Distanzmatrix, die aus den Reisezeiten mit dem Pkw gewonnen wurde

soll, steht in der Matrix `target`. Dabei kann man die Transformationen getrennt spezifizieren. Soll eine Drehung durchgeführt werden, so gibt man `orthogonal=T` ein. Soll die Konfiguration verschoben werden, so setzt man das Argument `translate` auf den Wert `T`. Eine Vergrößerung beziehungsweise Verkleinerung der Konfiguration bewirkt das Argument `magnify`.

Schauen wir uns die Procrustes-Analyse in `S-PLUS` für das Beispiel 31 auf Seite 202 an. Die Konfigurationen mögen in den Matrizen `A` und `B` stehen:

```
> A
      [,1] [,2]
 [1,]   10    2
 [2,]   10    8
 [3,]    4    8
> B
```

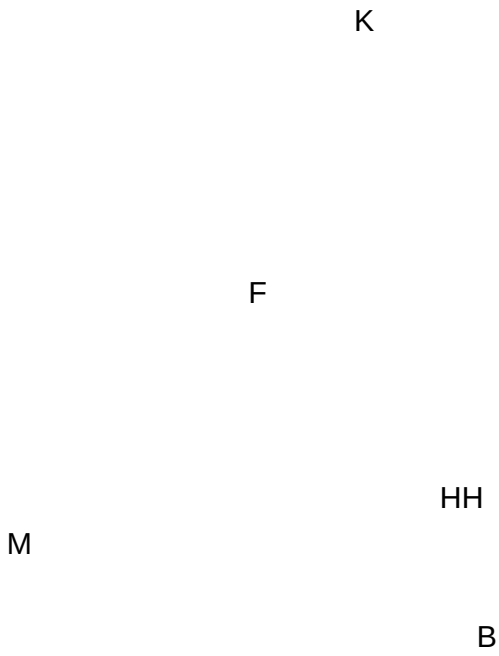


Fig. 7.8. Graphische Darstellung von 5 deutschen Städten mit einer metrischen mehrdimensionalen Skalierung auf der Basis einer Distanzmatrix, die aus den Reisezeiten mit der Bahn gewonnen wurde

```
     [,1] [,2]
[1,]    2    4
[2,]    2    1
[3,]    5    1
```

Die Zielkonfiguration steht in der Matrix `A`, die zu ändernde Konfiguration in der Matrix `B`. Wir rufen die Funktion `procrustes` auf:

```
> e<-procrustes(B, A, orthogonal=T, translate=T, magnify=T)
```

Das Ergebnis der Funktion `procrustes` ist ein Liste, in der die erste Komponente die neue Konfiguration enthält:

```
> e[[1]]
      [,1] [,2]
 [1,]   10    2
```

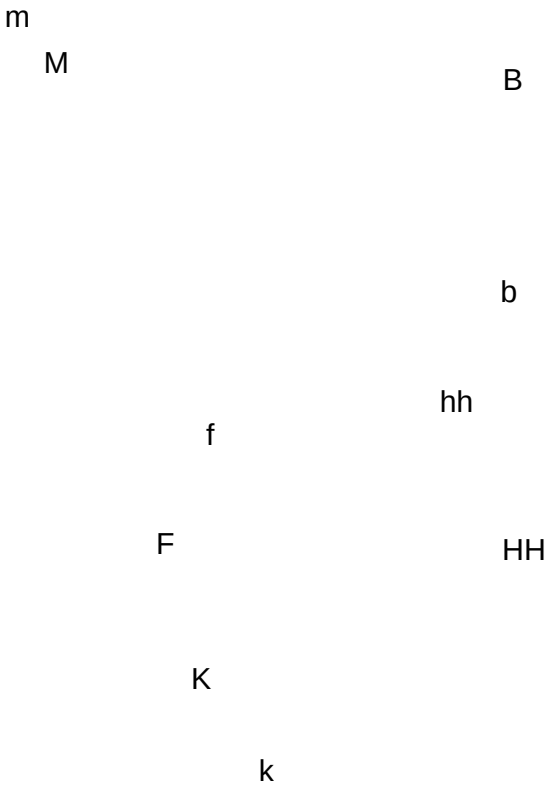


Fig. 7.9. Graphische Darstellung von 5 deutschen Städten mit einer metrischen mehrdimensionalen Skalierung auf der Basis von Distanzmatrizen, die aus den Reisezeiten mit dem Pkw und mit der Bahn gewonnen wurden, nach einer Procrustes-Analyse

```
 [2,]   10    8
 [3,]    4    8
```

Wir können noch überprüfen, wie gut die Anpassung ist:

```
 \index{S-PLUS Funktionen!sum}
> sum((A-e[[1]])^2)
[1] 3.944305e-030
```

Wir können eine Procrustes-Analyse auch ohne die Funktion `procrustes` durchführen. Zunächst zentrieren wir die beiden Matrizen:

```
> Az<-scale(A,scale=F)
> Az
      [,1] [,2]
 [1,]    2   -4
 [2,]    2    2
 [3,]   -4    2
> Bz<-scale(B,scale=F)
> Bz
      [,1] [,2]
 [1,]   -1    2
 [2,]   -1   -1
 [3,]    2   -1
```

Dann führen wir eine Singulärwertzerlegung der Matrix $\tilde{\mathbf{A}}'\tilde{\mathbf{B}}$ mit der Funktion `svd` durch:

```
> e<-svd(t(Az)%*%Bz)
> e
 $d:
 [1] 18  6

 $v:
            [,1]       [,2]
 [1,]  0.7071068 -0.7071068
 [2,] -0.7071068 -0.7071068

 $u:
            [,1]      [,2]
 [1,] -0.7071068 0.7071068
 [2,]  0.7071068 0.7071068
```

Die Rotationsmatrix $\mathbf{T}$ erhält man durch

```
> e$v%*%t(e$u)
                [,1]           [,2]
 [1,] -1.000000e+000 -1.110223e-016
 [2,]  1.110223e-016 -1.000000e+000
```

und den Streckungs- beziehungsweise Stauchungsfaktor c durch

```
> sum(e$d)/sum(diag(Bz%*%t(Bz)))
[1] 2
```

Wir wollen nun noch die Procrustes-Analyse der Reisezeiten aus Kapitel 7.4 auf Seite 210 in `S-PLUS` nachvollziehen.

Die Distanzen mögen in den Variablen

```
> reisezeitenpkw
     HH   B   K   F   M
 HH   0 192 271 314 454
  B 192   0 381 365 386
  K 271 381   0 134 295
  F 314 365 134   0 251
  M 454 386 295 251   0
```

und

```
> reisezeitenbahn
     HH   B   K   F   M
 HH   0 184 247 254 409
  B 184   0 297 263 433
  K 247 297   0 175 385
  F 254 263 175   0 257
  M 409 433 385 257   0
```

stehen. Wir führen jeweils eine metrische mehrdimensionale Skalierung:

```
> A<-cmdscale(reisezeitenpkw)
> B<-cmdscale(reisezeitenbahn)
```

und anschließend die Procrustes-Analyse durch:

```
> e<-procrustes(B,A, orthogonal=T, translate=T, magnify=T).
```

Die Abbildung 7.9 auf Seite 213 erhalten wir durch folgende Befehle:

```
plot(A,xlab="",ylab="",axes=F,type="n")
text(A,dimnames(reisezeitenpkw)[[1]])
text(e$rmat,c("hh","b","k","f","m"))
```

7.6 Ergänzungen und weiterführende Literatur

Wir haben hier gezeigt, wie man vorzugehen hat, wenn eine von zwei Konfigurationen der anderen durch Verschieben, Drehen, Strecken oder Stauchen ähnlich gemacht werden soll. Gower (1975) beschreibt, wie man mehr als zwei Konfigurationen simultan ähnlich machen kann. Weitere Aspekte der Procrustes-Analyse sind bei Cox & Cox (1994), S.98-104 zu finden.

7.7 Übungen

Exercise 16. Im Beispiel 30 auf Seite 199 sind die Ergebnisse einer Procrustes-Analyse zu finden. Vollziehen Sie diese Ergebnisse in `S-PLUS` nach.

Exercise 17. In Tabelle 1.5 auf Seite 7 sind die Luftlinienentfernungen zwischen deutschen Städten in Kilometern angegeben.

1. Führen Sie eine metrische mehrdimensionale Skalierung dieser Daten durch.
2. In Tabelle 1.6 auf Seite 7 sind die Reisezeiten zu finden, die zwischen 5 Städten mit dem Pkw benötigt werden. Führen Sie eine metrische mehrdimensionale Skalierung durch und vergleichen Sie die Ergebnisse mit den unter 1. gewonnenen mit Hilfe einer Procrustes-Analyse.
3. In Tabelle 1.7 auf Seite 7 sind die Reisezeiten zu finden, die zwischen 5 Städten mit der Bahn benötigt werden. Führen Sie eine metrische mehrdimensionale Skalierung durch und vergleichen Sie die Ergebnisse mit den unter 1. gewonnenen mit Hilfe einer Procrustes-Analyse.

Exercise 18. In Übung 10 auf Seite 150 wurde eine Hauptkomponentenanalyse der 0.95-Quantile der Punktezahlen in den Bereichen `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` durchgeführt. Vergleichen Sie das Ergebnis der durch die Hauptkomponentenanalyse gewonnenen zweidimensionalen Darstellung mit der durch die Hauptkomponentenanalyse in Kapitel 5.5 gewonnenen zweidimensionalen Darstellung der Mittelwerte.

Part III

Abhängigkeitsstrukturen

8 Lineare Regression

8.1 Problemstellung und Modell

Oft ist man daran interessiert, die Abhängigkeit einer Variablen Y von einer Variablen x_1 oder mehreren Variablen $x_1, \ldots, x_p$ zu modellieren. Dabei bezeichnen wir Y als *zu erklärende Variable* und $x_1, \ldots, x_p$ als *erklärende Variablen.*

Example 33. Im Juli 1999 möchte ein Dozent gerne seinen VW-Golf 3 verkaufen, der 50000 Kilometer gefahren wurde. Er fragt sich, zu welchem Preis y er ihn anbieten soll. Um dies entscheiden zu können, möchte er wissen, wie der Preis von den gefahrenen Kilometern x_1 abhängt. □

Wir wollen den Zusammenhang zwischen Y und $x_1, \ldots, x_p$ durch eine Funktion beschreiben. In der Regel wird Y noch von anderen Variablen abhängen, sodass der Zusammenhang zwischen Y und $x_1, \ldots, x_p$ nicht exakt ist. Wir nehmen aber an, dass der Einfluss der anderen Variablen gering ist. Wir fassen die anderen Variablen zur *Störgröße* ϵ zusammen. Dabei ist ϵ eine Zufallsvariable. Wir unterstellen außerdem, dass die Störgröße ϵ additiv wirkt. Somit gehen wir von folgendem Modell aus:

$$Y = f(x_1, \ldots, x_p) + \epsilon. \tag{8.1}$$

Man nennt (8.1) ein *Regressionsmodell.* Wird $f(x_1, \ldots, x_p)$ nicht näher spezifiziert, so ist (8.1) ein *nichtparametrisches Regressionsmodell.* Meist unterstellt man, dass $f(x_1, \ldots, x_p)$ linear in den Parametern $\beta_0, \beta_1, \ldots, \beta_p$ ist. Es gilt also

$$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Im Normalfall sind die *Parameter* $\beta_0, \beta_1, \ldots, \beta_p$ unbekannt. Um sie zu schätzen, werden die Daten $x_{i1}, \ldots, x_{ip}, y_i$ für die Merkmalsträger oder Zeitpunkte $i = 1, \ldots, n$ erhoben. hmcounterend. (fortgesetzt)

Example 33. Der Dozent studiert den Anzeigenteil der Süddeutschen Zeitung und findet 33 VW-Golf 3. In Tabelle 1.8 auf Seite 8 sind die Merkmale `Alter`, `Gefahrene Kilometer` und `Angebotspreis` der Autos zu finden. □

Wird nur eine erklärende Variable betrachtet, so spricht man von *Einfachregression*, ansonsten von *multipler Regression*. Bei Einfachregression gibt ein Streudiagramm Hinweise auf den Zusammenhang zwischen den beiden Variablen. hmcounterend. (fortgesetzt)

Example 33. Abbildung 8.1 zeigt das Streudiagramm der Daten.

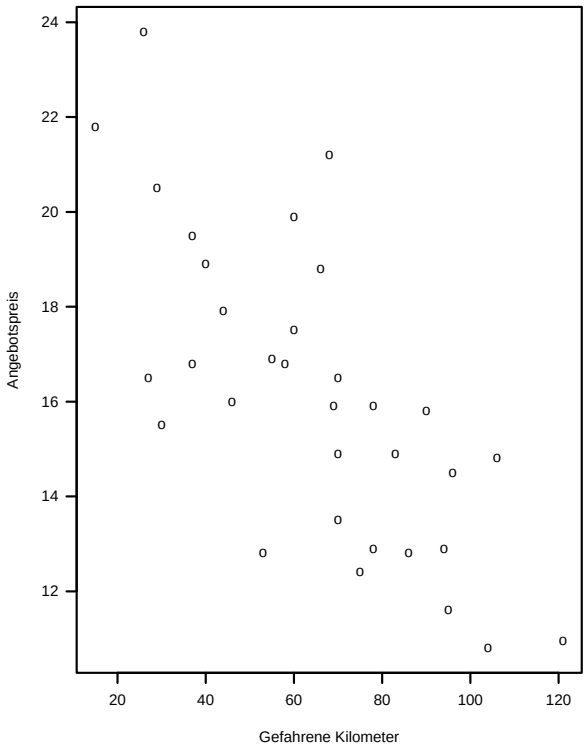


Fig. 8.1. Streudiagramm der Merkmale Gefahrene Kilometer und Angebotspreis (in 1000 DM)

Wie wir erwartet haben, nimmt der Angebotspreis mit zunehmender Anzahl gefahrener Kilometer ab. Wir sehen auch, dass der Zusammenhang tendenziell linear ist. □

Wir werden uns im Folgenden mit linearer Regression beschäftigen. Ausgangspunkt ist hier für $i = 1, \ldots, n$ das Modell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i. \tag{8.2}$$

Dabei nimmt man an, dass gilt

$$E(\epsilon_i) = 0 \tag{8.3}$$

und

$$Var(\epsilon_i) = \sigma^2 \tag{8.4}$$

für $i = 1, \ldots, n$. Annahme (8.3) besagt, dass sich der Einfluss der Störgrößen im Mittel aufhebt. Alle anderen Variablen üben also keinen systematischen Einfluss auf Yaus. Wenn die Annahme (8.4) erfüllt ist, so ist die Varianz der Störgrößen konstant. Man spricht von *Homoskedastie*. Ist die Varianz der Störgrößen nicht konstant, so liegt *Heteroskedastie* vor. Neben (8.3) und (8.4) wird unterstellt, dass die Störgrößen unkorreliert sind. Für $i \neq j$ muss also gelten

$$Cov(\epsilon_i, \epsilon_j) = 0. \tag{8.5}$$

Wir können das Modell (8.2) folgendermaßen in Matrixform schreiben:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{8.6}$$

Dabei sind

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

und

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Die Annahmen (8.3), (8.4) und (8.5) über die Störgrößen lauten in matrizieller Form

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \tag{8.7}$$

und

$$Var(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n. \tag{8.8}$$

Dabei ist $\mathbf{0}$ der Nullvektor und $\mathbf{I}_n$ die (n, n)-Einheitsmatrix. Wir unterstellen im Folgenden noch, dass $\mathbf{X}$ vollen Spaltenrang besitzt, siehe dazu Seite 475.

Die Annahme über den Erwartungswert von $\boldsymbol{\epsilon}$ erlaubt es, das Modell (8.6) folgendermaßen zu schreiben:

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}. \tag{8.9}$$

Die Darstellung in (8.9) ermöglicht Verallgemeinerungen des linearen Modells. Eine dieser Verallgemeinerungen werden wir in Kapitel 12.4 kennenlernen.

8.2 Schätzung der Parameter

Schauen wir uns die Schätzproblematik zunächst anhand der Regression mit nur einer erklärenden Variablen an. Das Modell lautet

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

für $i = 1, \ldots, n$. Um die Parameter β_0 und β_1 zu schätzen, werden in $x_{11}, \ldots, x_{n1}$ Realisationen $y_1, \ldots, y_n$ von $Y_1, \ldots, Y_n$ beobachtet. Diese Punkte stellt man in einem Streudiagramm dar. Schätzung der Parameter β_0 und β_1 bedeutet, eine Gerade durch die Punktewolke zu legen. Gauss hat vorgeschlagen, die Gerade so durch die Punktewolke zu legen, dass die Summe der quadrierten senkrechten Abstände der Punkte von der Geraden minimal ist. Man nennt diese Vorgehensweise die *Kleinste-Quadrate-Methode*. Abbildung 8.2 veranschaulicht sie für drei Punkte.

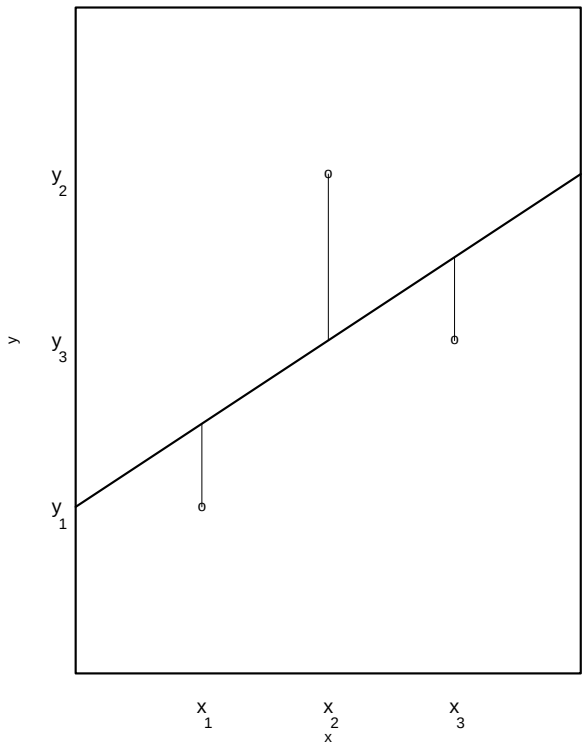


Fig. 8.2. Veranschaulichung der Kleinste-Quadrate-Methode

Gesucht sind also Werte von β_0 und β_1, sodass

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2 \tag{8.10}$$

minimal wird. Mit

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}$$

und

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

können wir (8.10) auch schreiben als

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{8.11}$$

Wir können die Zielfunktion (8.11) auch verwenden, wenn mehr als eine erklärende Variable betrachtet wird. In diesem Fall legen wir eine Hyperebene so durch die Punktewolke, dass die Summe der quadrierten Abstände der Punkte von der Hyperebene minimal ist. Sei

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}' - (\mathbf{X}\boldsymbol{\beta})') (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}' - \boldsymbol{\beta}'\mathbf{X}') (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - (\mathbf{y}'\mathbf{X}\boldsymbol{\beta})' - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{8.12}$$

Gesucht ist der Wert $\hat{\boldsymbol{\beta}}$, für den (8.12) minimal wird. Wir bezeichnen diesen als Kleinste-Quadrate-Schätzer von $\hat{\boldsymbol{\beta}}$. Wir bestimmen den Gradienten

$$\frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2\mathbf{X}'\mathbf{y} + 2\,\mathbf{X}'\mathbf{X}\boldsymbol{\beta},$$

siehe dazu Kapitel A.2.1 auf Seite 483. Der Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}$ erfüllt also die Gleichungen

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \tag{8.13}$$

Man nennt (8.13) auch die *Normalgleichungen*. Da $\mathbf{X}$ vollen Spaltenrang besitzt, existiert $(\mathbf{X}'\mathbf{X})^{-1}$. Ein Beweis hierfür ist auf Seite 475 im Anhang zu finden. Somit gilt

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{8.14}$$

Um zu überprüfen, ob es sich um ein Minimum handelt, müssen wir die Hesse-Matrix bestimmen, wie dies auf Seite 484 beschrieben ist. Es gilt

$$\frac{\partial}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'} S(\boldsymbol{\beta}) = 2\,\mathbf{X}'\mathbf{X}.$$

Wir haben zu zeigen, dass $\mathbf{X}'\mathbf{X}$ positiv definit ist. Für jeden Vektor $\mathbf{z} \neq \mathbf{0}$ muss also gelten

$$\mathbf{z}'\mathbf{X}'\mathbf{X}\mathbf{z} > 0. \tag{8.15}$$

Diese Bedingung ist erfüllt. Es gilt nämlich

$$\mathbf{z}'\mathbf{X}'\mathbf{X}\mathbf{z} = (\mathbf{X}\mathbf{z})'\,\mathbf{X}\mathbf{z} = \mathbf{v}\mathbf{v} = \sum_{i=1}^{n} v_i^2 \geq 0$$

mit

$$\mathbf{v} = \mathbf{X}\,\mathbf{z}.$$

Da $\mathbf{X}$ vollen Spaltenrang besitzt, ist $\mathbf{v}$ genau dann gleich dem Nullvektor, wenn $\mathbf{z}$ gleich dem Nullvektor ist.

Man kann zeigen, dass der Kleinste-Quadrate-Schätzer erwartungstreu ist. Unter allen linearen und erwartungstreuen Schätzern von $\boldsymbol{\beta}$ hat er die kleinste Varianz, wenn die Annahmen des Modells erfüllt sind. Ein Beweis ist bei Seber (1977), S. 49 zu finden. hmcounterend. (fortgesetzt)

Example 33. Sei Y_i die Variable `Angebotspreis` und x_{i1} die Variable `Gefahrene Kilometer` des i-ten VW Golfs, $i = 1, \ldots, n$. Wir unterstellen folgendes Modell:

$$Y_i = \beta_0 + \beta_1\, x_{i1} + \epsilon_i \tag{8.16}$$

für $i = 1, \ldots, n$. Dabei erfüllen die Störgrößen die Annahmen (8.3), (8.4) und (8.5).

Es gilt

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 33 & 2136 \\ 2136 & 160548 \end{pmatrix}.$$

Die Inverse von $\mathbf{X}'\mathbf{X}$ ist

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.218258 & -0.002904 \\ -0.002904 & 0.000045 \end{pmatrix}.$$

Mit

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 532150 \\ 32481750 \end{pmatrix}$$

gilt also

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 21825.53 \\ -88.06 \end{pmatrix}.$$

Der geschätzte `Angebotspreis` eines neuen Golfs beträgt somit 21825.53 DM. Diesen Wert erhalten wir, wenn wir x_1 gleich 0 setzen. Der Angebotspreis vermindert sich für 1000 gefahrene Kilometer um 88.06 DM. Da die Funktion linear ist, ist dieser Wert unabhängig davon, wie viele Kilometer bereits gefahren wurden. Abbildung 8.3 zeigt das Streudiagramm der Daten mit der geschätzten Regressionsgeraden.

□

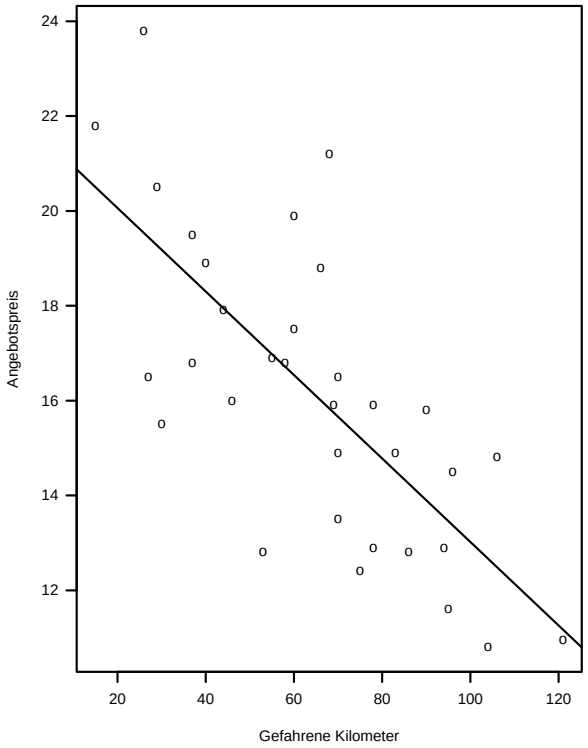


Fig. 8.3. Streudiagramm der Merkmale Gefahrene Kilometer und Angebotspreis (in 1000 DM) mit geschätzter Gerade

Neben dem Parametervektor $\boldsymbol{\beta}$ ist auch die Varianz σ^2 der Störgrößen ϵ_i, $i = 1, \ldots, n$ unbekannt. Die Schätzung der Varianz σ^2 beruht auf den Komponenten des Vektors

$$\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

der *Residuen*. Dieser ist definiert durch

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{8.17}$$

Wegen

$$\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$$

kann man die Residuen als Schätzer der Störgrößen auffassen. Somit liegt es nahe, die Varianz σ^2 durch die Stichprobenvarianz der Residuen zu schätzen:

$$s_e^2 = \frac{1}{n-1} \sum_{i=1}^{n} (e_i - \bar{e})^2 \tag{8.18}$$

mit

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i. \tag{8.19}$$

Man kann (8.18) noch vereinfachen, wenn man berücksichtigt, dass $\bar{e}$ gleich 0 ist. Dies folgt sofort, wenn man die Normalgleichungen (8.13) folgendermaßen schreibt:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{e} = \mathbf{0}. \tag{8.20}$$

Da die erste Spalte von $\mathbf{X}$ der Einservektor ist, folgt aus (8.20)

$$\sum_{i=1}^{n} e_i = 0 \tag{8.21}$$

und somit auch $\bar{e} = 0$.

s_e^2 ist kein erwartungstreuer Schätzer für σ^2. Eine kleine Modifikation von s_e^2 liefert einen erwartungstreuen Schätzer für σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} e_i^2. \tag{8.22}$$

Ein Beweis der Erwartungstreue von $\hat{\sigma}^2$ ist bei Seber (1977), S. 51 zu finden. hmcounterend. (fortgesetzt)

Example 33. Es gilt $\hat{\sigma}^2 = 5090120$. □

Für Tests benötigt man die Varianz-Kovarianz-Matrix von $\hat{\boldsymbol{\beta}}$. Es gilt

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{8.23}$$

Mit (3.31) auf Seite 88 und (A.42) auf Seite 475 gilt nämlich:

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= Var\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\right) \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\, Var(\mathbf{Y}) \left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)' \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\, Var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\mathbf{X} \left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)' \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\, Var(\boldsymbol{\epsilon})\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.
\end{aligned}$$

Die unbekannte Varianz σ^2 schätzen wir durch $\hat{\sigma}^2$ und erhalten folgenden Schätzer von $Var(\hat{\boldsymbol{\beta}})$:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{8.24}$$

hmcounterend. (fortgesetzt)

Example 33. Es gilt

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 1110959.6 & -14780.7 \\ -14780.7 & 228.4 \end{pmatrix}.$$

Es gilt also speziell $\widehat{Var}(\hat{\beta}_1) = 228.4$. □

8.3 Praktische Aspekte

8.3.1 Interpretation der Parameter bei mehreren erklärenden Variablen

hmcounterend. (fortgesetzt)

Example 33. Wir berücksichtigen ab jetzt auch noch das Merkmal `Alter` und gehen also aus vom Modell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{8.25}$$

für $i = 1, \ldots, n$.

Dabei sind Y_i der Angebotpreis, x_{i1} die gefahrenen Kilometer und x_{i2} das Alter des i-ten VW Golfs. Wir bestimmen die Kleinste-Quadrate-Schätzer. Es gilt

$$\hat{\beta}_0 = 24965.06, \quad \hat{\beta}_1 = -36.07, \quad \hat{\beta}_2 = -1421.48.$$

Im Modell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

gilt

$$\hat{\beta}_0 = 21825.53, \quad \hat{\beta}_1 = -88.06.$$

□

Wir sehen, dass sich die Schätzer in beiden Modellen unterscheiden. Es stellen sich zwei Fragen:

1. Woran liegt es, dass sich die Schätzwerte von β_1 in den Modellen

   $$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

   und

   $$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

   unterscheiden?

2. Wie hat man den Schätzwert $\hat{\beta}_1$ im Modell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

zu interpretieren?

Wir wollen die zweite Frage zuerst beantworten. Dafür benötigen wir die sogenannte *Hat-Matrix*. Sei

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{8.26}$$

Setzen wir (8.14) in (8.26) ein, so gilt

$$\hat{\mathbf{y}} = \mathbf{H}_\mathbf{X}\mathbf{y} \tag{8.27}$$

mit

$$\mathbf{H}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{8.28}$$

Tukey nennt $\mathbf{H}_\mathbf{X}$ die Hat-Matrix, da sie dem $\mathbf{y}$ den Hut aufsetzt. Die Hat-Matrix besitzt zwei wichtige Eigenschaften. Sie ist symmetrisch und idempotent. Es gilt also

$$\mathbf{H}_\mathbf{X} = \mathbf{H}'_\mathbf{X} \tag{8.29}$$

und

$$\mathbf{H}_\mathbf{X} = \mathbf{H}^2_\mathbf{X}. \tag{8.30}$$

Gleichung (8.29) ist erfüllt wegen

$$\mathbf{H}'_\mathbf{X} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1})'\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}_\mathbf{X}.$$

Gleichung (8.30) gilt wegen

$$\mathbf{H}^2_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}_\mathbf{X}.$$

Wir können auch den Vektor der Residuen über die Hat-Matrix ausdrücken. Es gilt

$$\mathbf{e} = (\mathbf{I} - \mathbf{H}_\mathbf{X})\mathbf{y}. \tag{8.31}$$

Dies sieht man folgendermaßen:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}_\mathbf{X}\mathbf{y} = (\mathbf{I} - \mathbf{H}_\mathbf{X})\mathbf{y}.$$

Mit

$$\mathbf{M}_\mathbf{X} = \mathbf{I}_n - \mathbf{H}_\mathbf{X} \tag{8.32}$$

gilt dann

$$\mathbf{e} = \mathbf{M}_\mathbf{X}\mathbf{y}. \tag{8.33}$$

Die Matrix $\mathbf{M}_\mathbf{X}$ ist auch symmetrisch und idempotent, denn es gilt

$$\mathbf{M}'_\mathbf{X} = (\mathbf{I}_n - \mathbf{H}_\mathbf{X})' = \mathbf{I}_n - \mathbf{H}'_\mathbf{X} = \mathbf{I}_n - \mathbf{H}_\mathbf{X} = \mathbf{M}_\mathbf{X}$$

und

$$\begin{aligned}\mathbf{M}^2_\mathbf{X} &= \mathbf{M}_\mathbf{X}\mathbf{M}_\mathbf{X} = (\mathbf{I}_n - \mathbf{H}_\mathbf{X})(\mathbf{I}_n - \mathbf{H}_\mathbf{X}) = \mathbf{I}_n - \mathbf{H}_\mathbf{X} - \mathbf{H}_\mathbf{X} + \mathbf{H}^2_\mathbf{X} \\ &= \mathbf{I}_n - \mathbf{H}_\mathbf{X} - \mathbf{H}_\mathbf{X} + \mathbf{H}_\mathbf{X} = \mathbf{I}_n - \mathbf{H}_\mathbf{X} = \mathbf{M}_\mathbf{X}.\end{aligned}$$

Gleichung (8.33) zeigt, dass die Multiplikation der Matrix $\mathbf{M}_\mathbf{X}$ mit dem Vektor $\mathbf{y}$ die Residuen einer linearen Regression von $\mathbf{y}$ auf $\mathbf{X}$ liefert. Wir werden diese Beziehung gleich benötigen.

Wenden wir uns nun der zweiten Frage auf Seite 229 zu. Um sie zu beantworten, zerlegen wir die Matrix der erklärenden Variablen in zwei Teilmatrizen $\mathbf{X}_{(1)}$ und $\mathbf{X}_{(2)}$. Dabei enthält $\mathbf{X}_{(1)}$ k erklärende Variablen und $\mathbf{X}_{(2)}$ die restlichen erklärenden Variablen. Es gilt also

$$\mathbf{X} = \left(\mathbf{X}_{(1)}, \mathbf{X}_{(2)}\right).$$

Entsprechend zerlegen wir den Parametervektor $\boldsymbol{\beta}$ in die Teilvektoren $\boldsymbol{\beta}_{(1)}$ und $\boldsymbol{\beta}_{(2)}$:

$$\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_{(1)} \\ \boldsymbol{\beta}_{(2)}\end{pmatrix}.$$

Das Modell (8.6) lautet also:

$$\mathbf{Y} = \mathbf{X}_{(1)}\boldsymbol{\beta}_{(1)} + \mathbf{X}_{(2)}\boldsymbol{\beta}_{(2)} + \boldsymbol{\epsilon}. \tag{8.34}$$

Wir zerlegen den Kleinste-Quadrate-Schätzer entsprechend:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}_{(1)} \\ \hat{\boldsymbol{\beta}}_{(2)}\end{pmatrix}.$$

Wir zeigen im Folgenden, wie $\hat{\boldsymbol{\beta}}_{(1)}$ von $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$ und $\mathbf{Y}$ abhängt. Schreiben wir die Normalgleichungen (8.13) mit der partitionierten Matrix $\mathbf{X}$ und dem partitionierten Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}$, so gilt

$$\begin{pmatrix}\mathbf{X}'_{(1)} \\ \mathbf{X}'_{(2)}\end{pmatrix} \begin{pmatrix}\mathbf{X}_{(1)} & \mathbf{X}_{(2)}\end{pmatrix} \begin{pmatrix}\hat{\boldsymbol{\beta}}_{(1)} \\ \hat{\boldsymbol{\beta}}_{(2)}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'_{(1)} \\ \mathbf{X}'_{(2)}\end{pmatrix} \mathbf{Y}. \tag{8.35}$$

Hieraus folgt:

$$\mathbf{X}'_{(1)}\,\mathbf{X}_{(1)}\,\hat{\boldsymbol{\beta}}_{(1)} + \mathbf{X}'_{(1)}\,\mathbf{X}_{(2)}\,\hat{\boldsymbol{\beta}}_{(2)} = \boldsymbol{X}'_{(1)}\,\mathbf{Y}\,, \tag{8.36}$$

$$\mathbf{X}'_{(2)}\,\mathbf{X}_{(1)}\,\hat{\boldsymbol{\beta}}_{(1)} + \mathbf{X}'_{(2)}\,\mathbf{X}_{(2)}\,\hat{\boldsymbol{\beta}}_{(2)} = \mathbf{X}'_{(2)}\,\mathbf{Y}. \tag{8.37}$$

Wir lösen (8.37) nach $\hat{\boldsymbol{\beta}}_{(2)}$ auf und erhalten

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(2)} &= (\mathbf{X}'_{(2)}\mathbf{X}_{(2)})^{-1}\,\mathbf{X}'_{(2)}\mathbf{Y} - (\mathbf{X}'_{(2)}\mathbf{X}_{(2)})^{-1}\mathbf{X}'_{(2)}\mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)},\\ &= (\mathbf{X}'_{(2)}\mathbf{X}_{(2)})^{-1}\mathbf{X}'_{(2)}(\mathbf{Y} - \mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)}).\end{aligned} \tag{8.38}$$

Setzen wir (8.38) für $\hat{\boldsymbol{\beta}}_{(2)}$ in (8.36) ein, so ergibt sich

$$\mathbf{X}'_{(1)}\mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)} + \mathbf{X}'_{(1)}\,\mathbf{H}_{X_{(2)}}(\mathbf{Y} - \mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)}) = \mathbf{X}'_{(1)}\mathbf{Y}. \tag{8.39}$$

Dabei ist

$$\mathbf{H}_{\mathbf{X}_{(2)}} = \mathbf{X}_{(2)}(\mathbf{X}'_{(2)}\mathbf{X}_{(2)})^{-1}\mathbf{X}'_{(2)}.$$

Aus (8.39) folgt

$$\mathbf{X}'_{(1)}\mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)} - \mathbf{X}'_{(1)}\,\mathbf{H}_{\mathbf{X}_{(2)}}\mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)} = \mathbf{X}'_{(1)}\mathbf{Y} - \mathbf{X}'_{(1)}\mathbf{H}_{\mathbf{X}_{(2)}}\mathbf{Y}.$$

Dies können wir umformen zu

$$\mathbf{X}'_{(1)}\left(\mathbf{I}_n - \mathbf{H}_{\mathbf{X}_{(2)}}\right)\mathbf{X}_{(1)}\hat{\boldsymbol{\beta}}_{(1)} = \mathbf{X}'_{(1)}\left(\mathbf{I}_n - \mathbf{H}_{\mathbf{X}_{(2)}}\right)\mathbf{Y}. \tag{8.40}$$

Wir setzen

$$\mathbf{M}_{\mathbf{X}_{(2)}} = \mathbf{I}_n - \mathbf{H}_{\mathbf{X}_{(2)}} \tag{8.41}$$

und lösen (8.40) nach $\hat{\boldsymbol{\beta}}_{(1)}$ auf:

$$\hat{\boldsymbol{\beta}}_{(1)} = \left(\mathbf{X}'_{(1)}\mathbf{M}_{\mathbf{X}_{(2)}}\mathbf{X}_{(1)}\right)^{-1}\mathbf{X}'_{(1)}\mathbf{M}_{\mathbf{X}_{(2)}}\mathbf{Y}.$$

Wenn wir noch berücksichtigen, dass $\mathbf{M}_{X_{(2)}}$ idempotent und symmetrisch ist, gilt

$$\hat{\boldsymbol{\beta}}_{(1)} = \left((\mathbf{M}_{\mathbf{X}_{(2)}}\mathbf{X}_{(1)})'\,\mathbf{M}_{\mathbf{X}_{(2)}}\mathbf{X}_{(1)}\right)^{-1}(\mathbf{M}_{\mathbf{X}_{(2)}}\mathbf{X}_{(1)})'\mathbf{M}_{\mathbf{X}_{(2)}}\,\mathbf{Y}. \tag{8.42}$$

Der Vergleich von Gleichung (8.42) mit Gleichung (8.14) zeigt, dass $\hat{\boldsymbol{\beta}}_{(1)}$ der Kleinste-Quadrate-Schätzer einer Regression von $\mathbf{M}_{X_{(2)}}\mathbf{Y}$ auf $\mathbf{M}_{X_{(2)}}\mathbf{X}_{(1)}$ ist. Dabei ist $\mathbf{M}_{X_{(2)}}\mathbf{Y}$ der Vektor der Residuen einer Regression von $\mathbf{Y}$ auf $\mathbf{X}_{(2)}$ und die Spalten von $\mathbf{M}_{X_{(2)}}\mathbf{X}_{(1)}$ sind die Vektoren der Residuen von Regressionen der Spalten von $\mathbf{X}_{(1)}$ auf $\mathbf{X}_{(2)}$. Der Schätzer von $\boldsymbol{\beta}_{(1)}$ im Modell (8.35) ist somit der Kleinste-Quadrate-Schätzer einer Regression von Residuen einer Regression von $\mathbf{Y}$ auf $\mathbf{X}_{(2)}$ auf Residuen einer Regression von $\mathbf{X}_{(1)}$ auf $\mathbf{X}_{(2)}$.

Wir bereinigen also $\mathbf{Y}$ und $\mathbf{X}_{(1)}$ um den linearen Effekt von $\mathbf{X}_{(2)}$. Hierdurch halten wir $\mathbf{X}_{(2)}$ künstlich konstant. Besteht $\hat{\boldsymbol{\beta}}_{(1)}$ nur aus einer Komponente, so gibt $\hat{\boldsymbol{\beta}}_{(1)}$ an, wie sich $\mathbf{Y}$ ändert, wenn sich die zu $\hat{\boldsymbol{\beta}}_{(1)}$ gehörende Variable um eine Einheit erhöht, und alle anderen Variablen konstant sind.

Wenden wir uns Frage 1 auf Seite 228 zu. Wir fragen uns, wann der Kleinste-Quadrate-Schätzer von $\boldsymbol{\beta}_{(1)}$ in den Modellen

$$\mathbf{Y} = \mathbf{X}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon} \tag{8.43}$$

und

$$\mathbf{Y} = \mathbf{X}_{(1)}\boldsymbol{\beta}_{(1)} + \mathbf{X}_{(2)}\boldsymbol{\beta}_{(2)} + \boldsymbol{\epsilon} \tag{8.44}$$

identisch ist. Im Modell (8.43) gilt

$$\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{X}'_{(1)}\mathbf{X}_{(1)})^{-1}\mathbf{X}'_{(1)}\mathbf{Y}. \tag{8.45}$$

Den Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}_{(1)}$ im Modell (8.44) erhalten wir, indem wir Gleichung (8.36) nach $\hat{\boldsymbol{\beta}}_{(1)}$ auflösen:

$$\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{X}'_{(1)}\mathbf{X}_{(1)})^{-1}\mathbf{X}'_{(1)}\mathbf{Y} - (\mathbf{X}'_{(1)}\mathbf{X}_{(1)})^{-1}\mathbf{X}'_{(1)}\mathbf{X}_{(2)}\hat{\boldsymbol{\beta}}_{(2)}.$$

Wir sehen, dass $\hat{\boldsymbol{\beta}}_{(1)}$ in den Modellen (8.43) und (8.44) identisch ist, wenn gilt

$$\mathbf{X}'_{(1)}\mathbf{X}_{(2)} = \mathbf{0}.$$

8.3.2 Die Güte der Anpassung

Das Bestimmtheitsmaß Um ein Maß für die Güte der Anpassung zu erhalten, betrachten wir die Summe $\mathbf{y}'\mathbf{y}$ der quadrierten y_i, $i = 1, \ldots, n$. Aus (8.17) folgt mit (8.26)

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}.$$

Somit gilt

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= (\hat{\mathbf{y}} + \mathbf{e})'(\hat{\mathbf{y}} + \mathbf{e}) = (\hat{\mathbf{y}}' + \mathbf{e}')(\hat{\mathbf{y}} + \mathbf{e}) \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\hat{\mathbf{y}} + \hat{\mathbf{y}}'\mathbf{e} + \mathbf{e}'\mathbf{e}.\end{aligned}$$

Mit (8.20) gilt

$$\hat{\mathbf{y}}'\mathbf{e} = (\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{e} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{e} = \hat{\boldsymbol{\beta}}'\mathbf{0} = \mathbf{0}.$$

Außerdem gilt

$$\mathbf{e}'\hat{\mathbf{y}} = (\mathbf{e}'\hat{\mathbf{y}})' = \hat{\mathbf{y}}'\mathbf{e} = 0. \tag{8.46}$$

Also gilt

$$\mathbf{y}'\,\mathbf{y} = \mathbf{e}'\,\mathbf{e} + \hat{\mathbf{y}}'\,\hat{\mathbf{y}}. \tag{8.47}$$

Wir subtrahieren $n\,\bar{y}^2$ von beiden Seiten von (8.47) und erhalten

$$\mathbf{y}'\,\mathbf{y} - n\,\bar{y}^2 = \mathbf{e}'\,\mathbf{e} + \hat{\mathbf{y}}'\,\hat{\mathbf{y}} - n\,\bar{y}^2. \tag{8.48}$$

Es gilt

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2.$$

Dies sieht man folgendermaßen:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - 2 \sum_{i=1}^{n} y_i \, \bar{y} + \sum_{i=1}^{n} \bar{y}^2$$

$$= \sum_{i=1}^{n} y_i^2 - 2\, n \bar{y}^2 + n\, \bar{y}^2 = \sum_{i=1}^{n} y_i^2 - n\, \bar{y}^2.$$

Also steht auf der linken Seite von Gleichung (8.48)

$$\sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Mit (8.26) lauten die Normalgleichungen (8.13):

$$\mathbf{X}' \, \hat{\mathbf{y}} = \mathbf{X}' \, \mathbf{y}.$$

Da die erste Spalte von $\mathbf{X}$ nur aus Einsen besteht, gilt

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

und somit

$$\bar{y} = \bar{\hat{y}}.$$

Also können wir (8.48) schreiben als

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^{n} e_i^2. \tag{8.49}$$

Auf der linken Seite von Gleichung (8.49) steht die Streuung der y_i. Diese Streuung zerlegen wir gemäß Gleichung (8.49) in zwei Summanden. Der erste Summand

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2$$

ist die Streuung der $\hat{y}_i$, während der zweite Summand

$$\sum_{i=1}^{n} e_i^2$$

die Streuung der Residuen angibt. Das i-te Residuum e_i ist gleich der Differenz aus y_i und $\hat{y}_i$. Je kleiner die Residuen sind, umso besser ist die Anpassung. Umso größer ist dann aber auch der erste Summand auf der rechten Seite in Gleichung (8.49).

Setzen wir also

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2$$

ins Verhältnis zu

$$\sum_{i=1}^{n} (y_i - \bar{y})^2,$$

so erhalten wir ein Maß für die Güte der Anpassung:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \tag{8.50}$$

Man nennt R^2 auch das *Bestimmtheitsmaß* . hmcounterend. (fortgesetzt)

Example 33. Es gilt $R^2 = 0.781$. □

Offensichtlich gilt

$$0 \leq R^2 \leq 1.$$

Aufgrund von Gleichung (8.49) gilt

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Ist R^2 gleich 1, so gilt

$$\sum_{i=1}^{n} e_i^2 = 0.$$

Die Anpassung ist perfekt.

Ist R^2 hingegen gleich 0, so gilt

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} e_i^2.$$

Residuenplot Neben dem Bestimmtheitsmaß zeigt sich die Güte der Anpassung in den Plots der Residuen e_i, $i = 1, \ldots, n$. Es liegt nahe, eine Graphik der Residuen zu erstellen. Sind die Daten zeitlich erhoben worden, so liefert ein Plot der Residuen gegen die Zeit Informationen über Modellverletzungen. Ansonsten sollte man ein Streudiagramm der $\hat{y}_i$ und der Residuen e_i, $i = 1, \ldots, n$ erstellen, da $\mathbf{e}$ und $\hat{\mathbf{y}}$ unkorreliert sind. Wegen (8.21) und (8.46) gilt

$$\sum_{i=1}^{n} (e_i - \bar{e})(\hat{y}_i - \bar{\hat{y}}) = \sum_{i=1}^{n} e_i(\hat{y}_i - \bar{\hat{y}}) = \sum_{i=1}^{n} e_i\,\hat{y}_i - \sum_{i=1}^{n} e_i\,\bar{\hat{y}}$$

$$= \mathbf{e}'\,\hat{\mathbf{y}} - \bar{\hat{y}} \sum_{i=1}^{n} e_i = 0.$$

Eventuell vorhandene Muster im Residuenplot werden nicht durch eine Korrelation zwischen den e_i und $\hat{y}_i$, $i = 1, \ldots, n$ überlagert. Man kann sich also auf das Wesentliche konzentrieren. Schauen wir uns ein typisches Muster in einem Residuenplot an.

Abbildung 8.4 zeigt einen keilförmigen Residuenplot. Wie können wir diesen interpretieren? Da wir die Residuen als Schätzer der Störgrößen und $\hat{y}_i$ für $i = 1, \ldots, n$ als Schätzer von $E(Y_i)$ auffassen können, deutet der Residuenplot darauf hin, dass die Annahme der Homoskedastie verletzt ist. Die Varianz der Störgrößen hängt vom Erwartungswert der zu erklärenden Variablen ab. Im Beispiel wächst die Varianz. Ist die Annahme der Homoskedastie verletzt, so besitzt der Kleinste-Quadrate-Schätzer nicht mehr die kleinste Varianz in der Klasse der erwartungstreuen Schätzer. In diesem Fall sollte man entweder eine gewichtete Regression durchführen oder geeignete Transformationen der Variablen oder des Modells suchen. Eine detaillierte Beschreibung der Verfahren ist bei Carroll & Ruppert (1988) zu finden.

hmcounterend. (fortgesetzt)

Example 33. Abbildung 8.5 zeigt den Residuenplot im Modell

$$Y_i = \beta_0 + \beta_1\,x_{i1} + \beta_2\,x_{i2} + \epsilon_i.$$

Der Residuenplot deutet auf Heteroskedastie hin. □

8.3.3 Tests

Um exakte Tests durchführen zu können, müssen wir annehmen, dass die Störgrößen ϵ_i, $i = 1, \ldots, n$ normalverteilt sind. Wir wollen testen, ob die i-te erklärende Variable im Modell benötigt wird. Wir gehen in diesem Fall aus von den Hypothesen

$$H_0 : \beta_i = 0, \tag{8.51}$$

$$H_1 : \beta_i \neq 0.$$

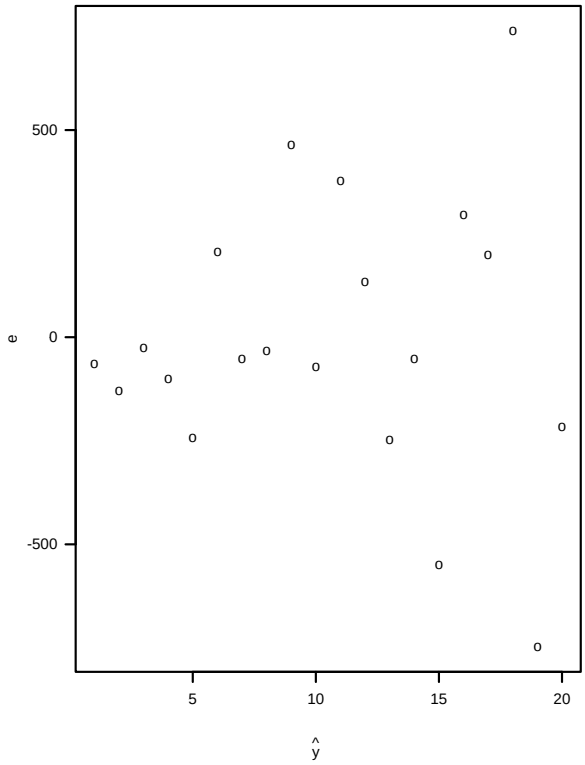


Fig. 8.4. Residuenplot bei Heteroskedastie

Außerdem wollen wir noch überprüfen, ob wir alle erklärenden Variablen gemeinsam benötigen. Wir betrachten also noch folgende Hypothese:

$$H_0 : \beta_1 = \ldots = \beta_p = 0, \tag{8.52}$$

$$H_1 : \text{Mindestens ein } \beta_i \text{ ist ungleich } 0,\ i = 1, \ldots, p.$$

Beginnen wir mit der Hypothese (8.51). Um diese zu überprüfen, bestimmt man die Teststatistik

$$t = \frac{\hat{\beta}_i}{\sqrt{\widehat{Var}(\hat{\beta}_i)}}. \tag{8.53}$$

Die Teststatistik t ist t-verteilt mit $n - p - 1$ Freiheitsgraden, wenn (8.51) zutrifft. Der Beweis ist bei Seber (1977), S.96-98 zu finden.

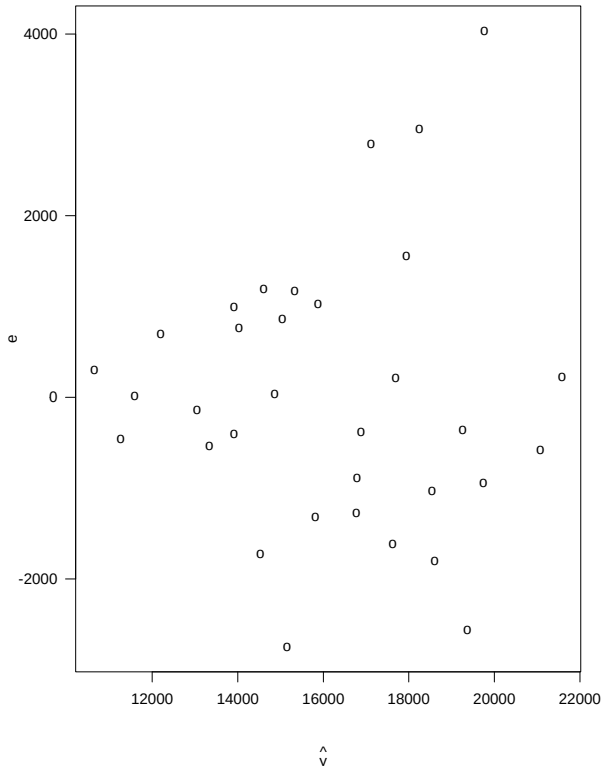


Fig. 8.5. Residuenplot bei der Regression von Angebotspreis auf Gefahrene Kilometer und Alter

Wir lehnen (8.51) ab, wenn $|t|$ größer als der kritische Wert $t_{n-p-1;1-\alpha/2}$ ist, wobei $t_{n-p-1;1-\alpha/2}$ das $1-\alpha/2$-Quantil der t-Verteilung mit $n-p-1$ Freiheitsgraden ist. hmcounterend. (fortgesetzt)

Example 33. Wir gehen aus vom Modell

$$Y_i = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \epsilon_i$$

für $i = 1, \ldots, n$. Es gilt

$$\hat{\beta}_0 = 24965.062, \quad \hat{\beta}_1 = -36.07, \quad \hat{\beta}_2 = -1421.4805$$

und

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 805657.4 & -2394.9 & -126205.8 \\ -2394.9 & 184.7 & -2089.7 \\ -126205.8 & -2089.7 & 57142.1 \end{pmatrix}.$$

Wir testen

$$H_0 : \beta_1 = 0, \tag{8.54}$$

$$H_1 : \beta_1 \neq 0. \tag{8.55}$$

Es gilt

$$t = \frac{-36.07}{\sqrt{184.7}} = -2.654.$$

Der Tabelle C.4 auf Seite 506 entnehmen wir $t_{30,0.975} = 2.0423$. Wir lehnen (8.51) also ab. □

Viele Programmpakete geben die *Überschreitungswahrscheinlichkeit* aus. Diese ist die Wahrscheinlichkeit für das Auftreten von Werten der Teststatistik, die noch extremer sind als der beobachtete Wert. hmcounterend. (fortgesetzt)

Example 33. Die Überschreitungswahrscheinlichkeit beträgt

$$P(|t| > 2.654) = P(t < -2.654) + P(t > 2.654) = 0.0063 + 0.0063 = 0.0126.$$

Abbildung 8.6 verdeutlicht die Berechnung der Überschreitungswahrscheinlichkeit. Diese ist gleich der schraffierten Fläche.

□

Wir lehnen eine Nullhypothese ab, wenn die Überschreitungswahrscheinlichkeit kleiner als das vorgegebene Signifikanzniveau α ist. Die Wahrscheinlichkeit, extremere Werte als den kritischen Wert zu beobachten, beträgt α. Ist die Überschreitungswahrscheinlichkeit also kleiner als α, so muss der beobachtete Wert der Teststatistik extremer sein als der kritische Wert.

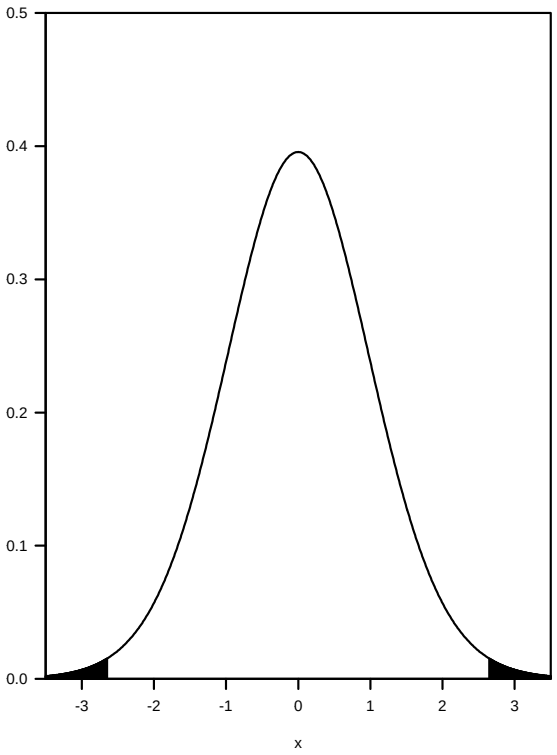


Fig. 8.6. Illustration der Überschreitungswahrscheinlichkeit am Beispiel der t-Verteilung

Schauen wir uns die Hypothese (8.52) an. Diese Hypothese wird mit der folgenden Teststatistik überprüft:

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p}. \tag{8.56}$$

Wenn (8.52) zutrifft, ist diese Teststatistik F-verteilt mit p und $n-p-1$ Freiheitsgraden, siehe dazu Seber (1977), S.97. Wir lehnen (8.52) ab, wenn gilt

$$F > F_{p,n-p-1,1-\alpha}.$$

Dabei ist $F_{p,n-p-1,1-\alpha}$ das $1-\alpha$-Quantil der F-Verteilung mit p und $n-p-1$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 33. Wir gehen aus vom Modell

$$Y_i = \beta_0 + \beta_1\, x_{i1} + \beta_2\, x_{i2} + \epsilon_i$$

für $i = 1, \ldots, n$.

Es soll getestet werden:

$$H_0 : \beta_1 = \beta_2 = 0,$$

$$H_1 : \text{Mindestens ein } \beta_i \text{ ist ungleich 0, } i = 1, 2.$$

Es gilt $R^2 = 0.781$, $n = 33$ und $p = 2$. Somit ergibt sich $F = 53.5$. Der Tabelle C.5 auf Seite 507 entnehmen wir $F_{2,30,0.95} = 3.32$. Wir lehnen H_0 zum Niveau 0.05 also ab. □

8.4 Lineare Regression in S-PLUS

Wir wollen das Beispiel 33 in S-PLUS nachvollziehen. Die Variablen `Alter`, `Gefahrene Kilometer` und `Angebotspreis` mögen in S-PLUS in den Variablen `Alter`, `Kilometer` und `Preis` stehen.

In S-PLUS gibt es eine Funktion `lm`, mit der man unter anderem eine lineare Regression durchführen kann. Sie wird aufgerufen durch

```
lm(formula, data=<<see below>>, weights=<<see below>>,
    subset=<<see below>>, na.action=na.fail, method="qr",
    model=F, x=F, y=F, contrasts=NULL, ...)
```

Mit dem Argument `formula` können wir das Modell durch eine Formel spezifizieren. Schauen wir uns die Vorgehensweise für ein Beispiel an. Wir wollen das Modell

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n$$

schätzen. Dabei stehen $y_1, \ldots, y_n$ in S-PLUS in der Variablen `y`, $x_{11}, \ldots, x_{n1}$ in `x1` und $x_{12}, \ldots, x_{n2}$ in `x2`. Wir spezifizieren die Formel durch

```
y ~ x1 + x2.
```

Auf der linken Seite der Formel steht die zu erklärende Variable. Das Zeichen ~ liest man als 'wird modelliert durch'. Auf der rechten Seite stehen die erklärenden Variablen getrennt durch das Zeichen `+`. Wollen wir also die Variable `Preis` auf die Variablen `Kilometer` und `Alter` regressieren, so geben wir ein

```
> e<-lm(Preis~Kilometer+Alter).
```

Schauen wir uns noch die anderen Argumente von `lm` an, bevor wir auf `e` eingehen. Das Argument `data` erlaubt es, die Daten in Form eines Dataframe zu übergeben. Bei diesem werden mehrere Variablen unter einem Namen zusammengefasst. Wie auf Seite 64 gezeigt wird, müssen wir hierzu

```
> golf<-data.frame(Preis,Kilometer,Alter)
```

eingeben. Existieren nur der Dataframe `golf` und nicht die Variablen `Preis`, `Kilometer` und `Alter`, so rufen wir die Funktion `lm` auf durch

```
> e<-lm(Preis~Kilometer+Alter,data=golf)
```

Das Argument `weights` bietet die Möglichkeit, eine gewichtete Regression durchzuführen. Durch `subset` kann man den Teil der Beobachtungen spezifizieren, die bei der Regression berücksichtigt werden sollen. Auf die anderen Argumente wollen wir nicht eingehen.

Schauen wir uns das Ergebnis der Funktion `lm` an. Der Aufruf

```
> summary(lm(Preis~Kilometer+Alter,data=golf))
```

liefert alle Informationen, die wir kennengelernt haben. Schauen wir uns diese an:

```
Call: lm(formula = Preis ~ Kilometer + Alter)
Residuals:
   Min     1Q Median    3Q  Max
 -2752 -941.2 -145.2 856.1 4037

Coefficients:
                 Value  Std. Error     t value    Pr(>|t|)
(Intercept)  24965.0618    897.5842     27.8136      0.0000
  Kilometer    -36.0737     13.5915     -2.6541      0.0126
      Alter  -1421.4805    239.0441     -5.9465      0.0000

Residual standard error: 1554 on 30 degrees of freedom
Multiple R-Squared: 0.781
F-statistic: 53.48 on 2 and 30 degrees of freedom,
the p-value is 1.282e-010

Correlation of Coefficients:
          (Intercept) Kilometer
Kilometer -0.1963
    Alter -0.5882     -0.6432
```

Als erstes wird der Aufruf der Funktion wiedergegeben. Es folgt Information über die Residuen in Form der Fünf-Zahlen-Zusammenfassung, die auf Seite 19 dargestellt wird.

Unter der Überschrift `Coefficients` schließen sich Angaben über die Kleinste-Quadrate-Schätzer $\hat{\beta}_i$, $i = 0, 1, \ldots, p$ an. Die Schätzwerte stehen unter `value` und die Quadratwurzeln aus den Varianzen der Schätzer, die auch Standardfehler genant werden, unter `Std. Error`. Die Werte des t-Tests auf $H_0 : \beta_i = 0$ kann man der Spalte mit der Überschrift `t value` entnehmen. Die Überschreitungswahrscheinlichkeit steht in der letzten Spalte.

Im nächsten Abschnitt des Outputs findet man den Schätzer $\hat{\sigma}$ unter `Residual standard error` und den Wert des Bestimmtheitsmaßes R^2 unter `Multiple R-Squared`. Es folgt der Test von $H_0 : \beta_1 = \beta_2 = 0$. Hier wird der Wert der F-Statistik und die Überschreitungswahrscheinlichkeit ausgegeben.

Im letzten Abschnitt ist die Korrelationsmatrix der Parameterschätzer zu finden.

Abbildung 8.3 erzeugt man mit folgender Befehlsfolge:

```
> plot(Kilometer,Preis)
> e<-lm(Preis~Kilometer,data=golf)[[1]]
> abline(e)
```

Dabei zeichnet die Funktion `abline` eine Gerade. Ihr Argument ist ein Vektor, dessen erste Komponente das Absolutglied und dessen zweite Komponente die Steigung der Geraden ist. Diese Information steckt in der ersten Komponente von `lm(Preis Kilometer,data=golf)`:

```
> e
 (Intercept) Kilometer
   21825.53 -88.05833
```

Um den Residuenplot in Abbildung 8.5 auf Seite 238 zu erstellen, benötigt man die Residuen e_i und die $\hat{y}_i$ für $i = 1, \ldots, n$. Diese erhält man mit den Funktionen `resid` und `fitted`. Die Abbildung liefert dann folgende Befehlsfolge:

```
> e<-lm(Preis~Kilometer+Alter,data=golf)
> plot(fitted(e),resid(e),ylab="e",xlab="")
> text(16000,-4000,"y")
> text(16000,-3900,"^")
```

Weitere Residuenplots erhält man mit `plot(e)`.

8.5 Ergänzungen und weiterführende Literatur

Wir haben hier nur einen kleinen Einblick in die lineare Regressionsanalyse gegeben. Nahezu alle theoretischen Aspekte sind bei Seber (1977) zu finden. Einen umfassenden Überblick aus der Sicht des Anwenders unter Berücksichtigung theoretischer Konzepte geben Draper & Smith (1998). Unterschiedliche Schätzverfahren werden von Birkes & Dodge (1993) behandelt, die die Algorithmen sehr detailliert beschreiben. Wie man einflussreiche Beobachtungen entdecken kann, findet man bei Cook & Weisberg (1982). Nichtparametrische Verfahren der Regressionsanalyse werden umfassend von Härdle (1990a) und Hastie & Tibshirani (1991) behandelt.

8.6 Übungen

Exercise 19. Stellen Sie sich vor, Sie haben einen Studienplatz in Bielefeld gefunden und suchen eine geeignete Wohnung. Sie schlagen die Neue Westfälische auf und suchen alle Einzimmerwohnungen heraus, die explizit in Uninähe liegen. Es sind acht. In Tabelle 8.1 sind die Flächen (in m^2) und Kaltmieten (in DM) der Wohnungen zu finden.

Sei x_i die Fläche und Y_i die Kaltmiete der i-ten Wohnung. Wir unterstellen folgendes Modell:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{8.57}$$

mit den üblichen Annahmen.

Table 8.1. Fläche und Kaltmiete von 8 Einzimmerwohnungen

Wohnung	Fläche	Kaltmiete
1	20	270
2	26	460
3	32	512
4	48	550
5	26	360
6	30	399
7	30	419
8	40	390

1. Erstellen Sie das Streudiagramm der Daten.
2. Bestimmen Sie die Kleinste-Quadrate-Schätzer von β_0 und β_1.
3. Zeichnen Sie die Regressionsgerade in das Streudiagramm.
4. Bestimmen Sie $\widehat{Var}(\beta_0)$ und $\widehat{Var}(\beta_1)$.
5. Testen Sie zum Niveau 0.05, ob x_i im Modell benötigt wird.
6. Bestimmen Sie R^2.
7. Erstellen Sie den Residuenplot. Deutet dieser auf Verletzungen der Annahmen des Modells hin?
8. Welche Miete erwarten Sie für eine Wohnung mit einer Fläche von 28 m^2?

Exercise 20. In `S-PLUS` gibt es einen Datensatz `air`. Mit `help(air)` können Sie Näheres über diesen Datensatz erfahren. Wir betrachten aus diesem Datensatz die Variablen `ozone`, `temperature` und `wind`.

Führen Sie in `S-PLUS` eine lineare Regression von `ozone` auf `temperature` und `wind` durch. Interpretieren Sie jede Zahl des Ergebnisses dieser Regression. Worauf deutet der Residuenplot hin?

9 Explorative Faktorenanalyse

9.1 Problemstellung und Grundlagen

Example 34. Bei einer Befragung von Erstsemestern wurde unter anderem nach den Merkmalen `Körpergröße` y_1, `Körpergewicht` y_2 und `Schuhgröße` y_3 gefragt. Die Antworten von 20 Studenten sind in Tabelle 9.1 zu finden.

Table 9.1. Körpergröße, Körpergewicht und Schuhgröße von 20 Studenten

Student	Körper-größe	Körper-gewicht	Schuh-größe	Student	Körper-größe	Körper-gewicht	Schuh-größe
1	171	58	40	11	201	93	48
2	180	80	44	12	180	67	42
3	178	80	42	13	183	73	42
4	171	60	41	14	176	65	42
5	182	73	44	15	170	65	41
6	180	70	41	16	182	85	40
7	180	77	43	17	180	80	41
8	170	55	42	18	190	83	44
9	163	50	37	19	180	67	39
10	169	51	38	20	183	75	45

Wir bestimmen die empirische Korrelationsmatrix

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.882 & 0.796 \\ 0.882 & 1.000 & 0.712 \\ 0.796 & 0.712 & 1.000 \end{pmatrix}. \tag{9.1}$$

□

Zwischen allen Merkmalen in Beispiel 34 besteht eine hohe positive Korrelation. Bei der Korrelation zwischen den Merkmalen `Körpergröße` und `Körpergewicht` wundert uns das nicht. Je größer eine Person ist, umso mehr wird sie auch wiegen. Die starke positive Korrelation zwischen den Merkmalen `Körpergröße` und `Schuhgröße` haben wir auch erwartet. Dass aber

die Merkmale `Körpergewicht` und `Schuhgröße` eine starke positive Korrelation aufweisen, ist verwunderlich. Warum sollten schwerere Personen größere Füße haben? Wir hätten hier eher einen Wert des empirischen Korrelationskoeffizienten in der Nähe von 0 erwartet. Woher kommt dieser hohe positive Wert? Der Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße` kann am Merkmal `Körpergröße` liegen, denn das Merkmal `Körpergröße` bedingt im Regelfall sowohl das Merkmal `Körpergewicht` als auch das Merkmal `Schuhgröße`. Um zu überprüfen, ob das Merkmal `Körpergröße` den Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße` bedingt, müssen wir es kontrollieren. Hierzu haben wir zwei Möglichkeiten:

- Wir betrachten nur Personen, die die gleiche Ausprägung des Merkmals `Körpergröße` besitzen, und bestimmen bei diesen den Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße`. Besteht bei Personen, die die gleiche Ausprägung des Merkmals `Körpergröße` besitzen, kein Zusammenhang zwischen den Merkmalen `Körpergewicht` und `Schuhgröße`, so sollte der Wert des empirischen Korrelationskoeffizienten gleich 0 sein.
- Wir können den Effekt des Merkmals `Körpergröße` auf die Merkmale `Körpergewicht` und `Schuhgröße` statistisch bereinigen und den Zusammenhang zwischen den bereinigten Merkmalen bestimmen.

Die erste Vorgehensweise ist auf Grund der Datenlage nicht möglich, also schauen wir uns die zweite an. Um die Merkmale `Körpergewicht` und `Schuhgröße` um den Effekt des Merkmals `Körpergröße` zu bereinigen, regressieren wir das Merkmal `Körpergewicht` y_2 auf das Merkmal `Körpergröße` y_1. Die Residuen e_{21t} dieser Regression können wir auffassen als das um den linearen Effekt des Merkmals `Körpergröße` bereinigte Merkmal `Körpergewicht`. Entsprechend regressieren wir das Merkmal `Schuhgröße` y_3 auf das Merkmal `Körpergröße` y_1. Die Residuen e_{31t} dieser Regression können wir auffassen als das um den linearen Effekt des Merkmals `Körpergröße` bereinigte Merkmal `Schuhgröße`. Der Wert des empirischen Korrelationskoeffizienten zwischen e_{21t} und e_{31t} ist ein Maß für die Stärke des Zusammenhangs zwischen den Merkmalen `Körpergewicht` und `Schuhgröße`, wenn man beide um den Effekt des Merkmals `Körpergröße` bereinigt hat. hmcounterend. (fortgesetzt)

Example 34. Abbildung 9.1 zeigt das Streudiagramm von e_{31t} gegen e_{21t}. Dieses deutet auf keinen linearen Zusammenhang zwischen e_{31t} und e_{21t} hin. Der Wert des empirischen Korrelationskoeffizienten zwischen e_{21t} und e_{31t} ist gleich 0.0348. Wir sehen, dass dieser Korrelationskoeffizient nahe 0 ist.

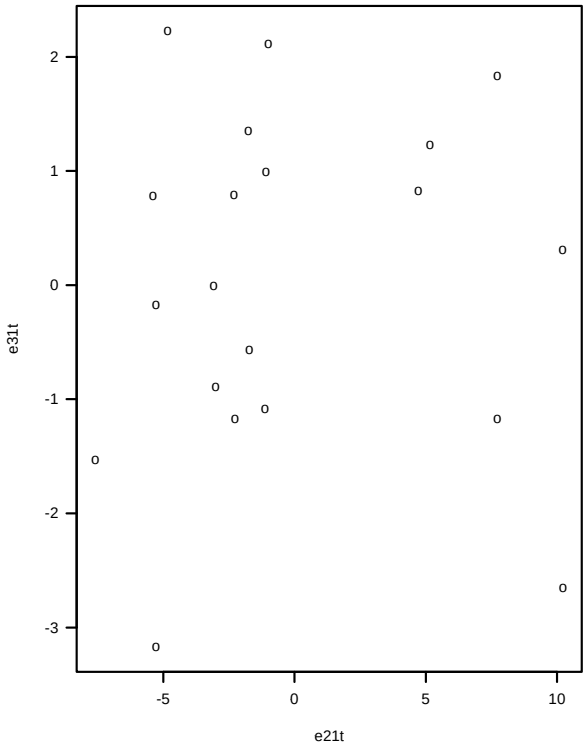


Fig. 9.1. Streudiagramm von e_{31t} gegen e_{21t}

□

Man nennt den Korrelationskoeffizienten zwischen e_{21t} und e_{31t} auch den *partiellen Korrelationskoeffizienten* $r_{23.1}$. Die Notation macht deutlich, dass man die Korrelation zwischen y_2 und y_3 betrachtet, wobei beide um y_1 bereinigt sind. Man kann den partiellen Korrelationskoeffizienten $r_{23.1}$ folgendermaßen aus den Korrelationen r_{12}, r_{13} und r_{23} bestimmen:

$$r_{23.1} = \frac{r_{23} - r_{12}\, r_{13}}{\sqrt{(1 - r_{12}^2)\,(1 - r_{13}^2)}}. \tag{9.2}$$

Eine Herleitung ist bei Krzanowski (2000) zu finden. hmcounterend. (fortgesetzt)

Example 34. Aus (9.1) entnehmen wir

$$r_{12} = 0.882, \qquad r_{13} = 0.796, \qquad r_{23} = 0.712.$$

Also gilt

$$r_{23.1} = \frac{0.712 - 0.882 \cdot 0.796}{\sqrt{(1 - 0.882^2)\,(1 - 0.796^2)}} = 0.0348.$$

□

Ist der partielle Korrelationskoeffizient $r_{23.1}$ nahe 0, während r_{23} betragsmäßig groß ist, so sagen wir, dass das Merkmal y_1 den Zusammenhang zwischen y_2 und y_3 erklärt. hmcounterend. (fortgesetzt)

Example 34. Wir bestimmen noch $r_{12.3}$ und $r_{13.2}$. Es gilt

$$r_{12.3} = \frac{0.882 - 0.712 \cdot 0.796}{\sqrt{(1 - 0.712^2)\,(1 - 0.796^2)}} = 0.742$$

und

$$r_{13.2} = \frac{0.796 - 0.882 \cdot 0.712}{\sqrt{(1 - 0.882^2)\,(1 - 0.712^2)}} = 0.508.$$

Wir sehen, dass die beiden anderen partiellen Korrelationskoeffizienten nicht verschwinden. □

Im Beispiel wird die Korrelation zwischen zwei Merkmalen durch ein drittes beobachtbares Merkmal erklärt. Das erklärende Merkmal muss aber nicht immer beobachtbar sein. Schauen wir uns auch hier ein Beispiel an.

Example 35. In der Übung 1 auf Seite 68 wurden die mittleren Punktezahlen von 31 Ländern in den drei Bereichen `Ermitteln von Informationen`, `Textbezogenes Interpretieren` und `Reflektieren und Bewerten` ermittelt. Die Korrelationsmatrix sieht folgendermaßen aus:

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.981 & 0.910 \\ 0.981 & 1.000 & 0.925 \\ 0.910 & 0.925 & 1.000 \end{pmatrix}.$$

Es fällt auf, dass alle Merkmale sehr stark positiv miteinander korreliert sind. Ist ein Land also überdurchschnittlich gut in einem Bereich, so ist es auch überdurchschnittlich gut in jedem der anderen Bereiche. □

Die hohen Korrelationen im Beispiel können an einer Variablen liegen, die mit allen drei Merkmalen positiv korreliert ist und die positive Korrelation zwischen diesen bewirkt. Dies ist eine Variable, die wir nicht messen können. Wir nennen diese unbeobachtbare Variable *Faktor* F. Wir wollen nun ein Modell entwickeln, bei dem die Korrelationen zwischen den Variablen $Y_1, \ldots, Y_p$ durch einen Faktor erklärt werden. Hierzu vollziehen wir noch einmal nach, wie wir vorgegangen sind, um zwei Variablen um den linearen Effekt einer dritten Variablen zu bereinigen. Schauen wir uns dazu zunächst die Variablen

Y_1 und Y_2 an. Wenn der Faktor F die Korrelation zwischen den Variablen Y_1 und Y_2 erklärt, dann sollten die Variablen Y_1 und Y_2 linear vom Faktor F abhängen. Wir unterstellen also

$$Y_1 = \mu_1 + l_1 F + \epsilon_1,$$
$$Y_2 = \mu_2 + l_2 F + \epsilon_2.$$

Dabei sind ϵ_1 und ϵ_2 Zufallsvariablen und μ_1, l_1, μ_2 und l_2 Parameter. Die Gleichungen allein reichen aber nicht aus, um den Tatbestand zu beschreiben, dass der Faktor F den Zusammenhang zwischen Y_1 und Y_2 erklärt. Der partielle Korrelationskoeffizient $\rho_{12.F}$ muss gleich 0 sein. Das heißt aber, dass die Korrelation und damit auch die Kovarianz zwischen ϵ_1 und ϵ_2 gleich 0 sein müssen:

$$Cov(\epsilon_1, \epsilon_2) = 0. \tag{9.3}$$

Allgemein unterstellen wir also folgendes Modell:

$$Y_i = \mu_i + l_i F + \epsilon_i \tag{9.4}$$

für $i = 1, \ldots, p$. Dabei heißt l_i *Faktorladung*. Den i-ten Zufallsfehler ϵ_i nennt man auch *spezifischen Faktor*. Für den Faktor F unterstellen wir

$$E(F) = 0 \tag{9.5}$$

und

$$Var(F) = 1. \tag{9.6}$$

Für die Zufallsfehler unterstellen wir in Anlehnung an (9.3)

$$Cov(\epsilon_i, \epsilon_j) = 0 \quad \text{für } i \neq j. \tag{9.7}$$

Außerdem nehmen wir an:

$$E(\epsilon_i) = 0 \tag{9.8}$$

und

$$Var(\epsilon_i) = \psi_i \tag{9.9}$$

für $i = 1, \ldots, p$. Außerdem unterstellen wir, dass die Zufallsfehler und der Faktor unkorreliert sind:

$$Cov(\epsilon_i, F) = 0 \tag{9.10}$$

für $i = 1, \ldots, p$. Aus den Annahmen folgt eine spezielle Gestalt der Varianz-Kovarianz-Matrix der Y_i, die wir jetzt herleiten wollen. Beginnen wir mit den Varianzen der Y_i. Es gilt

$$Var(Y_i) = l_i^2 + \psi_i \tag{9.11}$$

für $i = 1, \ldots, p$. Dies sieht man mit (3.14), (3.15), (9.6), (9.9) und (9.10) folgendermaßen:

$$\begin{aligned}
Var(Y_i) &= Cov(Y_i, Y_i) = Cov(\mu_i + l_i\, F + \epsilon_i, \mu_i + l_i\, F + \epsilon_i) \\
&= Cov(l_i\, F + \epsilon_i, l_i\, F + \epsilon_i) \\
&= Cov(l_i\, F, l_i\, F) + Cov(l_i\, F, \epsilon_i) + Cov(\epsilon_i, l_i\, F) + Cov(\epsilon_i, \epsilon_i) \\
&= l_i\, l_i\, Cov(F, F) + l_i\, Cov(F, \epsilon_i) + l_i\, Cov(\epsilon_i, F) + Cov(\epsilon_i, \epsilon_i) \\
&= l_i^2\, Var(F) + Var(\epsilon_i) = l_i^2 + \psi_i\,.
\end{aligned}$$

Die Varianz von Y_i setzt sich also aus zwei Summanden zusammen. Der Summand l_i^2 wird *Kommunalität* des Faktors F genannt. Dies ist der Anteil des Faktors an der Varianz von Y_i. Der Summand ψ_i heißt *spezifische Varianz*.

Schauen wir uns die Kovarianz zwischen Y_i und Y_j für $i \neq j$ an. Es gilt

$$Cov(Y_i, Y_j) = l_i\, l_j \tag{9.12}$$

Dies sieht man mit (3.14), (3.15), (9.7), (9.6) und (9.10) folgendermaßen:

$$\begin{aligned}
Cov(Y_i, Y_j) &= Cov(\mu_i + l_i\, F + \epsilon_i, \mu_j + l_j\, F + \epsilon_j) \\
&= Cov(l_i\, F + \epsilon_i, l_j\, F + \epsilon_j) \\
&= Cov(l_i\, F, l_j\, F) + Cov(l_i\, F, \epsilon_j) + Cov(\epsilon_i, l_j\, F) + Cov(\epsilon_i, \epsilon_j) \\
&= l_i\, l_j\, Cov(F, F) + l_i\, Cov(F, \epsilon_j) + l_j\, Cov(\epsilon_i, F) \\
&= l_i\, l_j\, Var(F) \\
&= l_i\, l_j\,.
\end{aligned}$$

Schauen wir uns nun die Struktur der Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$ der Y_i an, die aus den Annahmen des Modells folgt. Wegen (9.11) und (9.12) gilt

$$\boldsymbol{\Sigma} = \begin{pmatrix}
l_1^2 + \psi_1 & l_1\, l_2 & \ldots & l_1\, l_p \\
l_2\, l_1 & l_2^2 + \psi_2 & \ldots & l_2\, l_p \\
\vdots & \vdots & \ddots & \vdots \\
l_p\, l_1 & l_p\, l_2 & \ldots & l_p^2 + \psi_p
\end{pmatrix}. \tag{9.13}$$

Wir formen (9.13) um:

$$\boldsymbol{\Sigma} = \begin{pmatrix} l_1^2 & l_1 l_2 & \dots & l_1 l_p \\ l_2 l_1 & l_2^2 & \dots & l_2 l_p \\ \vdots & \vdots & \ddots & \vdots \\ l_p l_1 & l_p l_2 & \dots & l_p^2 \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{pmatrix}$$

$$= \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_p \end{pmatrix} \begin{pmatrix} l_1 & l_2 & \dots & l_p \end{pmatrix} + \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{pmatrix}.$$

Mit

$$\mathbf{l} = \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_p \end{pmatrix}$$

und

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_p \end{pmatrix}$$

können wir (9.13) folgendermaßen schreiben:

$$\boldsymbol{\Sigma} = \mathbf{l}\,\mathbf{l}' + \boldsymbol{\Psi}. \tag{9.14}$$

Ziel der Faktorenanalyse ist es, die l_i und ψ_i aus der empirischen Varianz-Kovarianz-Matrix $\mathbf{S}$ zu schätzen. Wir suchen also im Modell mit einem Faktor einen Vektor $\hat{\mathbf{l}}$ und eine Matrix $\hat{\boldsymbol{\Psi}}$, sodass gilt

$$\mathbf{S} = \hat{\mathbf{l}}\hat{\mathbf{l}}' + \hat{\boldsymbol{\Psi}}.$$

hmcounterend. (fortgesetzt)

Example 35. Es gilt

$$\mathbf{S} = \begin{pmatrix} 68.997 & 85.834 & 16.568 \\ 85.834 & 137.397 & 20.916 \\ 16.568 & 20.916 & 6.274 \end{pmatrix}.$$

Bei einem Faktor und 3 Variablen erhält man folgende 6 Gleichungen:

$$s_1^2 = \hat{l}_1^2 + \hat{\psi}_1, \tag{9.15}$$

$$s_2^2 = \hat{l}_2^2 + \hat{\psi}_2, \tag{9.16}$$

$$s_3^2 = \hat{l}_3^2 + \hat{\psi}_3, \tag{9.17}$$

$$s_{12} = \hat{l}_1 \hat{l}_2, \tag{9.18}$$

$$s_{13} = \hat{l}_1 \hat{l}_3, \tag{9.19}$$

$$s_{23} = \hat{l}_2 \hat{l}_3 \,. \tag{9.20}$$

Wir lösen Gleichung (9.18) nach $\hat{l}_2$ und Gleichung (9.19) nach $\hat{l}_3$ auf und setzen die neu gewonnenen Gleichungen in Gleichung (9.20) ein:

$$s_{23} = \frac{s_{12} s_{13}}{\hat{l}_1^2} \,.$$

Also gilt

$$\hat{l}_1 = \sqrt{\frac{s_{12}s_{13}}{s_{23}}} = \sqrt{\frac{85.834 \cdot 16.568}{20.916}} = 8.2457\,.$$

Entsprechend erhalten wir

$$\hat{l}_2 = \sqrt{\frac{s_{12}s_{23}}{s_{13}}} = \sqrt{\frac{85.834 \cdot 20.916}{16.568}} = 10.4096$$

und

$$\hat{l}_3 = \sqrt{\frac{s_{13}s_{23}}{s_{12}}} = \sqrt{\frac{16.568 \cdot 20.916}{85.834}} = 2.0093\,.$$

Aus den Gleichungen (9.15), (9.16) und (9.17) folgt

$$\begin{aligned}
\hat{\psi}_1 &= s_1^2 - \hat{l}_1^2 = 68.997 - 67.9909 = 1.006,\\
\hat{\psi}_2 &= s_2^2 - \hat{l}_2^2 = 137.397 - 108.3597 = 29.037,\\
\hat{\psi}_3 &= s_3^2 - \hat{l}_3^2 = 6.274 - 4.037285 = 2.237\,.
\end{aligned}$$

Es gilt

$$\mathbf{S} = \hat{\mathbf{l}}\hat{\mathbf{l}}' + \hat{\boldsymbol{\Psi}}$$

mit

$$\hat{\mathbf{l}} = \begin{pmatrix} 8.2457 \\ 10.4096 \\ 2.0093 \end{pmatrix}$$

und

$$\hat{\boldsymbol{\Psi}} = \begin{pmatrix} 1.006 & 0 & 0 \\ 0 & 29.037 & 0 \\ 0 & 0 & 2.237 \end{pmatrix}.$$

□

Wir haben bisher nur einen Faktor betrachtet. In diesem Fall kann man die Schätzer mit dem im Beispiel beschriebenen Verfahren bestimmen. Bei mehr als einem Faktor muss man andere Schätzverfahren anwenden. Mit diesen werden wir uns beschäftigen, nachdem wir das allgemeine Modell dargestellt haben.

9.2 Theorie

9.2.1 Das allgemeine Modell

Ausgangspunkt ist die p-dimensionale Zufallsvariable $\mathbf{y} = (Y_1, \ldots, Y_p)'$ mit der Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$. Wir wollen die Kovarianzen zwischen den Zufallsvariablen $Y_1, \ldots, Y_p$ durch k Faktoren $F_1, \ldots, F_k$ erklären. Wir stellen folgendes Modell für $i = 1, \ldots, p$ auf:

$$Y_i = \mu_i + l_{i1}\, F_1 + l_{i2}\, F_2 + \ldots + l_{ik}\, F_k + \epsilon_i\,. \tag{9.21}$$

Mit

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} F_1 \\ \vdots \\ F_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{pmatrix}$$

und

$$\mathbf{L} = \begin{pmatrix} l_{11} & l_{12} & \ldots & l_{1k} \\ l_{21} & l_{22} & \ldots & l_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \ldots & l_{pk} \end{pmatrix}$$

können wir (9.21) schreiben als

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{L}\,\mathbf{F} + \boldsymbol{\epsilon}. \tag{9.22}$$

Die Annahmen des Modells lauten:

1. $E(\mathbf{F}) = \mathbf{0}$,
2. $Var(\mathbf{F}) = \mathbf{I}_n$,
3. $E(\boldsymbol{\epsilon}) = \mathbf{0}$,
4. $Var(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$,
5. $Cov(\boldsymbol{\epsilon}, \mathbf{F}) = \mathbf{0}$

mit

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{pmatrix}.$$

Unter diesen Annahmen gilt folgende Beziehung für die Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$ von $\mathbf{Y}$:

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}. \tag{9.23}$$

Dies sieht man unter Berücksichtigung von (3.27) und (3.28) folgendermaßen:

$$\begin{aligned}
\boldsymbol{\Sigma} &= Var(\mathbf{Y}) = Cov(\mathbf{Y},\mathbf{Y}) = Cov(\boldsymbol{\mu} + \mathbf{LF} + \boldsymbol{\epsilon}, \boldsymbol{\mu} + \mathbf{LF} + \boldsymbol{\epsilon}) \\
&= Cov(\mathbf{LF} + \boldsymbol{\epsilon}, \mathbf{LF} + \boldsymbol{\epsilon}) \\
&= Cov(\mathbf{LF},\mathbf{LF}) + Cov(\mathbf{LF},\boldsymbol{\epsilon}) + Cov(\boldsymbol{\epsilon},\mathbf{LF}) + Cov(\boldsymbol{\epsilon},\boldsymbol{\epsilon}) \\
&= \mathbf{L}\, Cov(\mathbf{F},\mathbf{F})\, \mathbf{L}' + \mathbf{L}\, Cov(\mathbf{F},\boldsymbol{\epsilon}) + Cov(\boldsymbol{\epsilon},\mathbf{F})\, \mathbf{L}' + \boldsymbol{\Psi} \\
&= \mathbf{L}\, Var(\mathbf{F})\, \mathbf{L}' + \boldsymbol{\Psi} \\
&= \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}.
\end{aligned}$$

Die Beziehung (9.23) wird das *Fundamentaltheorem der Faktorenanalyse* genannt. Schauen wir uns diese Beziehung genauer an. Es gilt

$$\sigma_i^2 = \sum_{j=1}^{k} l_{ij}^2 + \psi_i\,, \tag{9.24}$$

wobei σ_i^2, $i = 1, \ldots, p$ die Elemente auf der Hauptdiagonalen von $\boldsymbol{\Sigma}$ sind.

Wie das einfaktorielle Modell postuliert das allgemeine Modell eine Zerlegung der Varianz der i-ten Variablen in zwei Summanden. Der erste Summand heißt *Kommunalität*

$$h_i^2 = \sum_{j=1}^{k} l_{ij}^2. \tag{9.25}$$

Dies ist der Teil der Varianz von Y_i, der über die Faktoren mit den anderen Variablen geteilt wird. Der zweite Summand ψ_i ist der Teil der Varianz, der spezifisch für die Variable Y_i ist. Er wird deshalb spezifische Varianz genannt. Dies ist der Teil der Varianz von Y_i, der nicht mit den anderen Variablen geteilt wird.

Bisher haben wir nur die Varianz-Kovarianz-Matrix betrachtet. Es stellt sich die Frage, welche Konsequenzen es hat, wenn man statt der Varianz-Kovarianz-Matrix die Korrelationsmatrix betrachtet. Um diese Frage zu beantworten, schauen wir uns Gleichung (9.22) an, wobei wir $\boldsymbol{\mu}$ auf die linke Seite bringen:

$$\mathbf{Y} - \boldsymbol{\mu} = \mathbf{L}\,\mathbf{F} + \boldsymbol{\epsilon}. \tag{9.26}$$

Sei

$$\mathbf{D} = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_p \end{pmatrix}.$$

Multiplizieren wir $\mathbf{y}-\boldsymbol{\mu}$ von links mit $\mathbf{D}^{-1}$, so erhalten wir die standardisierte Zufallsvariable $\mathbf{y}^*$. Wir haben in (3.36) gezeigt, dass die Varianz-Kovarianz-Matrix von $\mathbf{y}^*$ gleich der Korrelationsmatrix $\mathbf{P}$ von $\mathbf{y}$ ist. Also gilt

$$\begin{aligned}\mathbf{P} &= Var(\mathbf{D}^{-1}(\mathbf{Y}-\boldsymbol{\mu})) = \mathbf{D}^{-1}Var(\mathbf{Y}-\boldsymbol{\mu})\big(\mathbf{D}^{-1}\big)' \\ &= \mathbf{D}^{-1}Var(\mathbf{Y})\big(\mathbf{D}^{-1}\big)' = \mathbf{D}^{-1}(\mathbf{L}\mathbf{L}'+\boldsymbol{\Psi})\big(\mathbf{D}^{-1}\big)' \\ &= \mathbf{D}^{-1}\mathbf{L}\mathbf{L}'\big(\mathbf{D}^{-1}\big)' + \mathbf{D}^{-1}\boldsymbol{\Psi}\big(\mathbf{D}^{-1}\big)' \\ &= \mathbf{D}^{-1}\mathbf{L}\big(\mathbf{D}^{-1}\mathbf{L}\big)' + \mathbf{D}^{-1}\boldsymbol{\Psi}\big(\mathbf{D}^{-1}\big)' = \tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}}\end{aligned}$$

mit

$$\tilde{\mathbf{L}} = \mathbf{D}^{-1}\mathbf{L}$$

und

$$\tilde{\boldsymbol{\Psi}} = \mathbf{D}^{-1}\boldsymbol{\Psi}\big(\mathbf{D}^{-1}\big)'.$$

Das faktoranalytische Modell gilt also auch für die standardisierten Variablen, wobei man die Ladungsmatrix $\tilde{\mathbf{L}}$ der standardisierten Variablen durch Skalierung der Ladungsmatrix $\mathbf{L}$ der ursprünglichen Variablen erhält. Das gleiche gilt für die spezifischen Varianzen. hmcounterend. (fortgesetzt)

Example 35. Wir wollen exemplarisch zeigen, dass die obigen Aussagen gelten. Die Analyse auf Basis der empirischen Varianz-Kovarianz-Matrix haben wir bereits durchgeführt. Wir betrachten nun das Modell

$$\mathbf{P} = \tilde{\mathbf{l}}\tilde{\mathbf{l}}' + \tilde{\boldsymbol{\Psi}}\,.$$

Wir suchen einen Vektor $\hat{\tilde{\mathbf{l}}}$ und eine Matrix $\hat{\tilde{\boldsymbol{\Psi}}}$, sodass gilt

$$\mathbf{R} = \hat{\tilde{\mathbf{l}}}\hat{\tilde{\mathbf{l}}}' + \hat{\tilde{\boldsymbol{\Psi}}}\,.$$

Dabei ist $\mathbf{R}$ die empirische Korrelationsmatrix. Es müssen folgende Gleichungen erfüllt sein:

$$1 = \hat{\tilde{l}}_1^2 + \hat{\tilde{\psi}}_1, \quad 1 = \hat{\tilde{l}}_2^2 + \hat{\tilde{\psi}}_2, \quad 1 = \hat{\tilde{l}}_3^2 + \hat{\tilde{\psi}}_3,$$

$$r_{12} = \hat{\tilde{l}}_1\hat{\tilde{l}}_2, \quad r_{13} = \hat{\tilde{l}}_1\hat{\tilde{l}}_3, \quad r_{23} = \hat{\tilde{l}}_2\hat{\tilde{l}}_3\,.$$

Setzen wir die Werte von $\mathbf{R}$ aus (9.1) ein und lösen die Gleichungen, so ergibt sich

$$\hat{\tilde{l}}_1 = 0.993, \quad \hat{\tilde{l}}_2 = 0.888, \quad \hat{\tilde{l}}_3 = 0.802$$

und

$$\hat{\tilde{\psi}}_1 = 0.015, \quad \hat{\tilde{\psi}}_2 = 0.211, \quad \hat{\tilde{\psi}}_3 = 0.357\,.$$

Bei der Faktorenanalyse auf Basis der empirischen Varianz-Kovarianz-Matrix $\mathbf{S}$ ist $\hat{l}_1 = 8.2457$.

Mit $s_1 = 8.306$ folgt

$$\frac{\hat{l}_1}{s_1} = 0.993 = \hat{\tilde{l}}_1 \,.$$

□

Wir können also auch die Korrelationsmatrix als Ausgangspunkt einer Faktorenanalyse nehmen. Dies werden wir auch machen. Wir bezeichnen die Ladungsmatrix im Folgenden mit $\mathbf{L}$ und die Matrix der spezifischen Varianzen mit $\boldsymbol{\Psi}$. Wir gehen also aus vom Modell

$$\mathbf{P} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}. \tag{9.27}$$

9.2.2 Nichteindeutigkeit der Lösung

Gegeben sei folgende Korrelationsmatrix:

$$\mathbf{P} = \begin{pmatrix} 1.00 & 0.58 & 0.66 & 0.22 & 0.16 \\ 0.58 & 1.00 & 0.78 & 0.32 & 0.26 \\ 0.66 & 0.78 & 1.00 & 0.42 & 0.36 \\ 0.22 & 0.32 & 0.42 & 1.00 & 0.74 \\ 0.16 & 0.26 & 0.36 & 0.74 & 1.00 \end{pmatrix}.$$

Es gilt

$$\mathbf{P} = \mathbf{L}_1\mathbf{L}_1' + \boldsymbol{\Psi} \tag{9.28}$$

mit

$$\mathbf{L}_1 = \begin{pmatrix} 0.7 & -0.1 \\ 0.8 & -0.2 \\ 0.9 & -0.3 \\ 0.2 & -0.8 \\ 0.1 & -0.9 \end{pmatrix}$$

und

$$\boldsymbol{\Psi} = \begin{pmatrix} 0.50 & 0 & 0 & 0 & 0 \\ 0 & 0.32 & 0 & 0 & 0 \\ 0 & 0 & 0.10 & 0 & 0 \\ 0 & 0 & 0 & 0.32 & 0 \\ 0 & 0 & 0 & 0 & 0.18 \end{pmatrix}.$$

Mit

$$\mathbf{L}_2 = \begin{pmatrix} 0.1 & 0.7 \\ 0.2 & 0.8 \\ 0.3 & 0.9 \\ 0.8 & 0.2 \\ 0.9 & 0.1 \end{pmatrix}$$

gilt aber auch

$$\mathbf{P} = \mathbf{L}_2\mathbf{L}_2' + \boldsymbol{\Psi}.$$

Bei festem k und $\boldsymbol{\Psi}$ gibt es also mindestens zwei unterschiedliche Ladungsmatrizen $\mathbf{L}_1$ und $\mathbf{L}_2$, die die Gleichung (9.27) erfüllen. Wie hängen die beiden Lösungen zusammen? Abbildung 9.2 zeigt die Punkte der beiden Lösungen, wobei die Punkte der ersten Lösung durch einen Stern und die Punkte der zweiten Lösung durch einen Kreis gekennzeichnet sind.

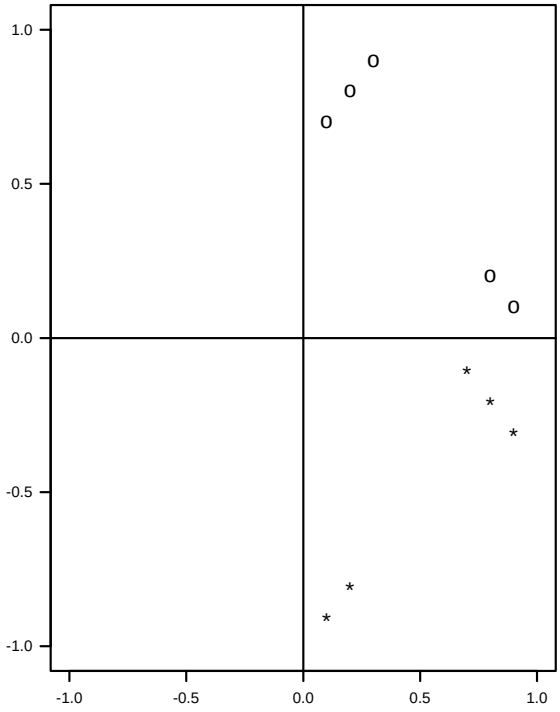


Fig. 9.2. Streudiagramm von zwei Lösungen der Fundamentalgleichung der Faktorenanalyse

Wir sehen, dass die gesamte erste Konfiguration durch eine Drehung um 90 Grad im Uhrzeigersinn in die zweite Konfiguration übergeht. Drehungen im $\mathbb{R}^2$ um den Winkel α werden bewirkt durch Multiplikation von rechts mit einer Matrix $\mathbf{T}$, für die gilt

$$\mathbf{T} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}.$$

Im Beispiel ist die Matrix $\mathbf{T}$ gegeben durch

$$\mathbf{T} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Es gilt also

$$\mathbf{L}_2 = \mathbf{L}_1\mathbf{T}\,.$$

Die Matrix $\mathbf{T}$ ist orthogonal. Es gilt also

$$\mathbf{T}\mathbf{T}' = \mathbf{I}_n. \tag{9.29}$$

Wegen der Orthogonalität von $\mathbf{T}$ gilt

$$\mathbf{P} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}' + \boldsymbol{\Psi} = \mathbf{L}\mathbf{T}(\mathbf{L}\mathbf{T})' + \boldsymbol{\Psi}\,.$$

Die Fundamentalgleichung ändert sich nicht, wenn wir die Faktorladungsmatrix $\mathbf{L}$ mit einer orthogonalen Matrix multiplizieren. In diesem Fall ändert sich das Modell (9.22) zu

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \epsilon. \tag{9.30}$$

Die transformierten Faktoren $\mathbf{T}'\mathbf{F}$ erfüllen die Bedingungen des faktoranalytischen Modells. Es gilt

$$E(\mathbf{T}'\mathbf{F}) = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$$

und

$$Var(\mathbf{T}'\mathbf{F}) = \mathbf{T}'Var(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}_n.$$

Die Faktorladungen sind also bis auf orthogonale Transformationen eindeutig. Durch eine Forderung an die Faktorladungsmatrix wird die Lösung eindeutig. Die übliche Forderung ist

$$\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L} = \boldsymbol{\Delta}, \tag{9.31}$$

wobei $\boldsymbol{\Delta}$ eine Diagonalmatrix ist. Eine Begründung für diese Forderung ist bei Fahrmeir et al. (1996), S. 646 zu finden.

9.2.3 Schätzung von L und $\boldsymbol{\Psi}$

Wir wollen uns in diesem Abschnitt damit beschäftigen, wie man im Modell (9.22) auf Seite 256 die unbekannten $\boldsymbol{\mu}$, $\mathbf{L}$, $\boldsymbol{\Psi}$ und k schätzen kann. Wir werden zwei Schätzverfahren betrachten. Beim *Maximum-Likelihood-Verfahren* unterstellt man, dass $\mathbf{y}$ multivariat normalverteilt ist, und bestimmt für festes k die Maximum-Likelihood-Schätzer der Parameter. Bei der *Hauptfaktorenanalyse* geht man auch von einem festen Wert von k aus

und schätzt $\boldsymbol{\mu}$ durch $\bar{\mathbf{x}}$. Die Schätzung von $\mathbf{L}$ und $\boldsymbol{\Psi}$ geht aus von der Gleichung (9.23). Es liegt nahe, in dieser Gleichung $\boldsymbol{\Sigma}$ durch die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ beziehungsweise durch die empirische Korrelationsmatrix $\mathbf{R}$ zu ersetzen. Dann erhält man folgende Schätzgleichung:

$$\mathbf{R} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}}. \tag{9.32}$$

Aus dieser bestimmt man dann die Schätzer $\hat{\mathbf{L}}$ und $\hat{\boldsymbol{\Psi}}$. Wie man dabei vorgeht, wollen wir uns im nächsten Abschnitt anschauen. Vorher wollen wir uns überlegen, wie viele Faktoren man höchstens verwenden darf, wenn p Variablen vorliegen. Wir schauen uns dies für die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ an. Für die empirische Korrelationsmatrix ergibt sich das gleiche Ergebnis, wie Mardia et al. (1979), S.260 zeigen.

Da die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ symmetrisch ist, enthält sie $0.5\,p\,(p+1)$ Größen, die zur Schätzung verwendet werden. In der Matrix $\hat{\mathbf{L}}$ gibt es pk und in der Matrix $\hat{\boldsymbol{\Psi}}$ p unbekannte Parameter. Die $pk+p$ Parameter unterliegen aber den Restriktionen in (9.31), sodass nur $pk + p - 0.5k(k-1)$ frei gewählt werden können. Da die Anzahl der Größen, die zur Schätzung verwendet werden, mindestens genau so groß sein sollte wie die Anzahl der geschätzten Parameter, muss also gelten:

$$0.5p(p+1) \geq pk + p - 0.5k(k-1). \tag{9.33}$$

Dies können wir vereinfachen zu

$$(p-k)^2 \geq p + k. \tag{9.34}$$

Tabelle 9.2 gibt die maximale Anzahl k von Faktoren in Abhängigkeit von der Anzahl p der Variablen an.

Table 9.2. Maximale Anzahl von Faktoren in Abhängigkeit von der Anzahl p der Variablen

Anzahl Variablen	Maximalzahl Faktoren
3	1
4	1
5	2
6	3
7	3
8	4
9	5

Schauen wir uns nun die beiden Schätzverfahren auf Basis des folgenden Beispiels genauer an.

Example 36. Im Beispiel 8 auf Seite 8 wurden Unternehmen gebeten, den Nutzen anzugeben, den sie von einem Virtual-Reality-System erwarten. Auf einer Skala von 1 bis 5 sollte dabei angegeben werden, wie wichtig die Merkmale `Veranschaulichung von Fehlfunktionen`, `Qualitätsverbesserung`, `Entwicklungszeitverkürzung`, `Ermittlung von Kundenanforderungen`, `Angebotserstellung` und `Kostenreduktion` sind. Es ergab sich folgende empirische Korrelationsmatrix:

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.223 & 0.133 & 0.625 & 0.506 & 0.500 \\ 0.223 & 1.000 & 0.544 & 0.365 & 0.320 & 0.361 \\ 0.133 & 0.544 & 1.000 & 0.248 & 0.179 & 0.288 \\ 0.625 & 0.365 & 0.248 & 1.000 & 0.624 & 0.630 \\ 0.506 & 0.320 & 0.179 & 0.624 & 1.000 & 0.625 \\ 0.500 & 0.361 & 0.288 & 0.630 & 0.625 & 1.000 \end{pmatrix}.$$

Wir wollen die Korrelationen durch zwei Faktoren erklären. Wir suchen also eine $(5, 2)$-Matrix $\hat{\mathbf{L}}$ und eine $(5, 5)$-Matrix $\hat{\boldsymbol{\Psi}}$, die

$$\mathbf{R} = \hat{\mathbf{L}}\,\hat{\mathbf{L}}' + \hat{\boldsymbol{\Psi}} \tag{9.35}$$

erfüllen und suchen für $k = 2$ die Schätzer $\hat{\mathbf{L}}$ und $\hat{\boldsymbol{\Psi}}$. □

Hauptfaktorenanalyse Die Hauptfaktorenanalyse beruht auf der Spektralzerlegung (A.51) einer symmetrischen Matrix. Wir gehen aus von (9.35) und bilden

$$\mathbf{R} - \hat{\boldsymbol{\Psi}} = \hat{\mathbf{L}}\hat{\mathbf{L}}'. \tag{9.36}$$

Da wir von der Matrix $\mathbf{R}$ die Diagonalmatrix $\hat{\boldsymbol{\Psi}}$ subtrahieren, unterscheiden sich nur die Hauptdiagonalelemente von $\mathbf{R}$ und $\mathbf{R} - \hat{\boldsymbol{\Psi}}$. Auf der Hauptdiagonale von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ stehen die Kommunalitäten h_i^2, $i = 1, \ldots, p$. Die i-te Kommunalität ist der Teil der Varianz von Y_i, den Y_i mit den anderen Variablen über die gemeinsamen Faktoren teilt. Einen Schätzer der Kommunalität von Y_i erhält man, indem man Y_i auf alle anderen Variablen regressiert. Das Bestimmtheitsmaß R_i^2 dieser Regression ist der Anteil der Varianz von Y_i, der durch die Regression von Y_i auf die anderen Variablen erklärt wird. Dies ist ein Schätzer von h_i^2. Das Bestimmtheitsmaß R_i^2 kann folgendermaßen auf Basis der empirischen Korrelationsmatrix ermittelt werden:

$$R_i^2 = 1 - \frac{1}{r^{ii}}. \tag{9.37}$$

Dabei ist r^{ii} das i-te Hauptdiagonalelement von $\mathbf{R}^{-1}$. Ein Beweis dieser Tatsache ist bei Seber (1977) auf Seite 333 zu finden. Ist die empirische Varianz-Kovarianz-Matrix $\mathbf{S}$ Ausgangspunkt der Schätzung, so gilt

$$R_i^2 = 1 - \frac{1}{s_{ii}\, s^{ii}}. \tag{9.38}$$

Dabei ist s^{ii} das i-te Hauptdiagonalelement von $\mathbf{S}^{-1}$ und s_{ii} das i-te Hauptdiagonalelement von $\mathbf{S}$. hmcounterend. (fortgesetzt)

Example 36. Wir bestimmen zunächst die Schätzer $\hat{h}_i^2$ der Kommunalitäten h_i^2. Es gilt

$$\mathbf{R}^{-1} = \begin{pmatrix} 1.733 & 0.040 & 0.061 & -0.811 & -0.250 & -0.231 \\ 0.040 & 1.576 & -0.738 & -0.235 & -0.169 & -0.122 \\ 0.061 & -0.738 & 1.450 & -0.070 & 0.123 & -0.214 \\ -0.811 & -0.235 & -0.07 & 2.366 & -0.599 & -0.606 \\ -0.250 & -0.169 & 0.123 & -0.599 & 1.972 & -0.703 \\ -0.231 & -0.122 & -0.214 & -0.606 & -0.703 & 2.043 \end{pmatrix}.$$

Also gilt

$$\hat{h}_1^2 = 0.422,\ \hat{h}_2^2 = 0.366,\ \hat{h}_3^2 = 0.311,\ \hat{h}_4^2 = 0.577,\ \hat{h}_5^2 = 0.493,\ \hat{h}_6^2 = 0.511.$$

Mit diesen Schätzern gilt

$$\mathbf{R} - \hat{\boldsymbol{\Psi}} = \begin{pmatrix} 0.422 & 0.223 & 0.133 & 0.625 & 0.506 & 0.500 \\ 0.223 & 0.366 & 0.544 & 0.365 & 0.320 & 0.361 \\ 0.133 & 0.544 & 0.311 & 0.248 & 0.179 & 0.288 \\ 0.625 & 0.365 & 0.248 & 0.577 & 0.624 & 0.630 \\ 0.506 & 0.320 & 0.179 & 0.624 & 0.493 & 0.625 \\ 0.500 & 0.361 & 0.288 & 0.630 & 0.625 & 0.511 \end{pmatrix}. \tag{9.39}$$

□

Kehren wir wieder zur Gleichung (9.36) zurück, deren linke Seite jetzt bekannt ist. Da die Matrix $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ symmetrisch ist, gilt wegen (A.51)

$$\mathbf{R} - \hat{\boldsymbol{\Psi}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'. \tag{9.40}$$

Wir betrachten zunächst den Fall, dass alle Eigenwerte λ_i von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ nichtnegativ sind. In diesem Fall können wir eine Diagonalmatrix $\boldsymbol{\Lambda}^{0.5}$ bilden, auf deren Hauptdiagonale die $\sqrt{\lambda_i}$ stehen. Es gilt

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{0.5}\boldsymbol{\Lambda}^{0.5}.$$

Wir können die Gleichung (9.40) umformen zu

$$\mathbf{R} - \hat{\boldsymbol{\Psi}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}' = \mathbf{U}\boldsymbol{\Lambda}^{0.5}\boldsymbol{\Lambda}^{0.5}\mathbf{U}' = \mathbf{U}\boldsymbol{\Lambda}^{0.5}(\mathbf{U}\boldsymbol{\Lambda}^{0.5})'.$$

Mit

$$\hat{\mathbf{L}} = \mathbf{U}\boldsymbol{\Lambda}^{0.5}$$

ist also Gleichung (9.36) erfüllt.

Sind aber Eigenwerte von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ negativ, so können wir $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ nur approximativ in der Form (9.36) darstellen. Hierzu bilden wir aus den Spalten von $\mathbf{U}$, die zu positiven Eigenwerten gehören, die Matrix $\mathbf{U}_1$. Entsprechend bilden wir die Diagonalmatrix $\boldsymbol{\Lambda}_1$ mit den positiven Eigenwerten von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ auf der Hauptdiagonalen. Wir erhalten hierdurch eine Approximation von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$:

$$\mathbf{R} - \hat{\boldsymbol{\Psi}} \doteq \mathbf{U}_1\boldsymbol{\Lambda}_1^{0.5}(\mathbf{U}_1\boldsymbol{\Lambda}_1^{0.5})'. \tag{9.41}$$

Mit

$$\hat{\mathbf{L}}_1 = \mathbf{U}_1\boldsymbol{\Lambda}_1^{0.5}$$

können wir dies auch schreiben als

$$\mathbf{R} - \hat{\boldsymbol{\Psi}} \doteq \hat{\mathbf{L}}_1\hat{\mathbf{L}}_1'.$$

Meist wollen wir die Korrelationen in $\mathbf{R}$ durch k Faktoren erklären. In diesem Fall wählen wir für die Spalten von $\mathbf{U}_1$ in (9.41) die normierten Eigenvektoren, die zu den k größten Eigenwerten gehören, und für die Hauptdiagonalelemente von $\boldsymbol{\Lambda}_1$ die k größten Eigenwerte. Dabei kann die Anzahl der Faktoren natürlich nicht größer werden als die Anzahl der positiven Eigenwerte. hmcounterend. (fortgesetzt)

Example 36. Wir führen eine Spektralzerlegung der Matrix in Gleichung (9.39) durch. Nur die beiden größten Eigenwerte sind nichtnegativ. Sie sind $\lambda_1 = 2.61$ und $\lambda_2 = 0.57$. Die zugehörigen Eigenvektoren sind

$$\mathbf{u}_1 = \begin{pmatrix} 0.403 \\ 0.322 \\ 0.248 \\ 0.497 \\ 0.452 \\ 0.470 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0.315 \\ -0.602 \\ -0.669 \\ 0.191 \\ 0.219 \\ 0.081 \end{pmatrix}.$$

Multiplizieren wir $\mathbf{u}_1$ mit $\sqrt{\lambda_1} = 1.62$ und $\mathbf{u}_2$ mit $\sqrt{\lambda_2} = 0.76$, so erhalten wir die Spalten der Matrix $\hat{\mathbf{L}}_1$:

$$\hat{\mathbf{L}}_1 = \begin{pmatrix} 0.653 & 0.239 \\ 0.523 & -0.458 \\ 0.401 & -0.508 \\ 0.805 & 0.145 \\ 0.732 & 0.166 \\ 0.761 & 0.062 \end{pmatrix}.$$

Es liegt nahe, diese Faktorladungen wie bei der Hauptkomponentenanalyse zu interpretieren. Wir werden aber im Kapitel 9.3.2 durch Rotation eine einfache Interpretation erhalten. □

Nachdem wir nun den Schätzer $\mathbf{L}$ gewonnen haben, können wir diesen in die Gleichung (9.35) einsetzen und so einen neuen Schätzer für $\boldsymbol{\Psi}$ bestimmen. hmcounterend. (fortgesetzt)

Example 36. Es gilt

$$\hat{\mathbf{L}}_1\hat{\mathbf{L}}_1' = \begin{pmatrix} 0.484 & 0.231 & 0.141 & 0.560 & 0.518 & 0.512 \\ 0.231 & 0.482 & 0.443 & 0.354 & 0.306 & 0.369 \\ 0.141 & 0.443 & 0.420 & 0.250 & 0.210 & 0.274 \\ 0.560 & 0.354 & 0.250 & 0.669 & 0.613 & 0.622 \\ 0.518 & 0.306 & 0.210 & 0.613 & 0.563 & 0.567 \\ 0.512 & 0.369 & 0.274 & 0.622 & 0.567 & 0.583 \end{pmatrix}.$$

Also gilt

$$\hat{\psi}_1 = 0.516, \;\; \hat{\psi}_2 = 0.518,$$
$$\hat{\psi}_3 = 0.580, \;\; \hat{\psi}_4 = 0.331,$$
$$\hat{\psi}_5 = 0.437, \;\; \hat{\psi}_6 = 0.417.$$

□

Nun können wir die neue Matrix $\hat{\boldsymbol{\Psi}}$ benutzen, um einen neuen Schätzer $\hat{\mathbf{L}}$ zu bestimmen.

Der folgende Algorithmus bestimmt die Schätzer einer Hauptfaktorenanalyse:

1. Bestimme für $i = 1, \ldots, p$ den Schätzer $\hat{\psi}_i$ der i-ten spezifischen Varianz durch $1 - R_i^2$. Dabei ist R_i^2 das Bestimmtheitsmaß einer Regression von Y_i auf die restlichen Variablen.
2. Stelle die Diagonalmatrix $\hat{\boldsymbol{\Psi}}$ mit den $\hat{\psi}_i$ auf der Hauptdiagonalen auf.
3. Berechne $\mathbf{R} - \hat{\boldsymbol{\Psi}}$.
4. Bestimme den Schätzer $\hat{\mathbf{L}}$ durch eine Spektralzerlegung von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$.
5. Stelle die Diagonalmatrix $\hat{\boldsymbol{\Psi}}$ auf mit den Hauptdiagonalelementen von $\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}'$ auf der Hauptdiagonalen.
6. Wiederhole die Schritte 3., 4. und 5. so lange, bis aufeinander folgende Paare von $\hat{\boldsymbol{\Psi}}$ und $\hat{\mathbf{L}}$ in einer vorgegebenen Genauigkeit identisch sind.

Maximum-Likelihood-Faktorenanalyse Es wird unterstellt, dass $\mathbf{y}$ multivariat normalverteilt ist mit Erwartungswert $\boldsymbol{\mu}$ und Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$. Die Herleitung der Log-Likelihood-Funktion ist bei Mardia et al. (1979), S. 97 zu finden. Die Log-Likelihood-Funktion lautet

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2} \ln |2\pi\boldsymbol{\Sigma}| - \frac{n}{2}\, tr\boldsymbol{\Sigma}^{-1}\mathbf{S} - \frac{n}{2}(\bar{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}).$$

Offensichtlich ist $\bar{\mathbf{y}}$ der Maximum-Likelihood-Schätzer von $\boldsymbol{\mu}$. Ersetzen wir $\boldsymbol{\mu}$ in der Log-Likelihood-Funktion durch $\bar{\mathbf{y}}$, so gilt

$$l(\boldsymbol{\Sigma}) = -\frac{n}{2} \ln |2\pi\boldsymbol{\Sigma}| - \frac{n}{2} tr\boldsymbol{\Sigma}^{-1}\mathbf{S}\,. \tag{9.42}$$

Ersetzen wir $\boldsymbol{\Sigma}$ in (9.42) durch $\mathbf{L}\mathbf{L}'+\boldsymbol{\Psi}$, so hängt die Log-Likelihood-Funktion von $\mathbf{L}$ und $\boldsymbol{\Psi}$ ab. Ein Algorithmus zur Maximierung der Log-Likelihood-Funktion unter der Nebenbedingung (9.31) ist bei Mardia et al. (1979), S. 263-266 zu finden. hmcounterend. (fortgesetzt)

Example 36. Wir werden in Kapitel 9.4 sehen, wie man den Maximum-Likelihood-Schätzer mit `S-PLUS` bestimmt. Schauen wir uns aber hier schon die Ergebnisse für $k = 2$ an. Es gilt

$$\hat{\mathbf{L}} = \begin{pmatrix} 0.659 & -0.255 \\ 0.537 & 0.446 \\ 0.450 & 0.678 \\ 0.822 & -0.182 \\ 0.734 & -0.201 \\ 0.764 & -0.089 \end{pmatrix}.$$

Die Schätzwerte der spezifischen Varianzen sind

$$\hat{\psi}_1 = 0.501, \;\; \hat{\psi}_2 = 0.513,$$
$$\hat{\psi}_3 = 0.338, \;\; \hat{\psi}_4 = 0.290,$$
$$\hat{\psi}_5 = 0.421, \;\; \hat{\psi}_6 = 0.408.$$

□

9.3 Praktische Aspekte

9.3.1 Bestimmung der Anzahl der Faktoren

Wie auch bei der Hauptkomponentenanalyse basiert die Bestimmung der Anzahl der Faktoren auf Eigenwerten von speziellen Matrizen. Diese Verfahren beruhen auf Guttman (1954), der eine untere Schranke für die Anzahl der Faktoren angegeben hat. Diese untere Schranke wird in der Praxis als Wert für die Anzahl der Faktoren gewählt. Wir wollen uns drei Kriterien anschauen.

Das erste Kriterium beruht auf der empirischen Korrelationsmatrix $\mathbf{R}$. Hier ist die untere Schranke für die Anzahl der Faktoren die Anzahl der Eigenwerte von $\mathbf{R}$, die größer als 1 sind. Nimmt man diesen Wert, so erhält man das im Kapitel 5.4.1 beschriebene Kriterium von Kaiser.

Das zweite Kriterium basiert auf der Matrix $\mathbf{R} - \hat{\boldsymbol{\Psi}}$. Hierbei wird $\hat{\boldsymbol{\Psi}}$ so gewählt, dass das i-te Hauptdiagonalelement von $\mathbf{R}$ durch das Bestimmtheitsmaß einer Regression von Y_i auf die restlichen Variablen ersetzt wird. Eine untere Schranke für die Anzahl der Faktoren ist hier die Anzahl der positiven Eigenwerte von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$.

Auf den Eigenwerten von $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ beruht ein weiteres Kriterium, das bei Krzanowski (2000), S.494 zu finden ist. Hier wird für die Anzahl der Faktoren der Wert k gewählt, bei dem die Summe der k größten Eigenwerte zum ersten Mal die Summe aller Eigenwerte übertrifft.

hmcounterend. (fortgesetzt)

Example 36. Die Eigenwerte der Matrix $\mathbf{R}$ sind:

$$\lambda_1 = 3.13, \quad \lambda_2 = 1.21, \quad \lambda_3 = 0.53, \quad \lambda_4 = 0.45, \quad \lambda_5 = 0.36, \quad \lambda_6 = 0.32.$$

Da zwei der Eigenwerte größer als 1 sind, entscheiden wir uns für zwei Faktoren. Zur gleichen Entscheidung gelangen wir bei den beiden anderen Kriterien. Die Eigenwerte der Matrix $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ lauten:

$$\lambda_1 = 2.61, \lambda_2 = 0.57, \lambda_3 = -0.006, \lambda_4 = -0.12, \lambda_5 = -0.15, \lambda_6 = -0.23.$$

Zwei der Eigenwerte sind größer als 0. Die Summe aller Eigenwerte beträgt 2.68. Somit würden wir uns auch nach dem dritten Kriterium für zwei Faktoren entscheiden, da die Summe der ersten beiden Eigenwerte 3.1786 und der erste Eigenwert 2.6056 beträgt. □

9.3.2 Rotation

Wir haben gesehen, dass die Faktorladungsmatrix bis auf eine Multiplikation mit einer orthogonalen Matrix eindeutig ist. Multipliziert man die Faktorladungsmatrix von rechts mit einer orthogonalen Matrix $\mathbf{T}$, so wird die Konfiguration der Punkte um den Nullpunkt gedreht. Man nennt die Matrix $\mathbf{T}$ auch *Rotationsmatrix*. Ein Ziel der Rotation ist es, Faktorladungen zu finden, die eine sogenannte *Einfachstruktur* besitzen. Bei dieser haben die Faktoren bei einigen Variablen eine sehr hohe Ladung, bei den anderen Variablen hingegen eine sehr niedrige. hmcounterend. (fortgesetzt)

Example 36. Schauen wir uns noch einmal die Ladungsmatrix $\hat{\mathbf{L}}$ an, die wir durch die Maximum-Likelihood-Faktorenanalyse gewonnen haben:

$$\hat{\mathbf{L}} = \begin{pmatrix} 0.659 & -0.255 \\ 0.537 & 0.446 \\ 0.450 & 0.678 \\ 0.822 & -0.182 \\ 0.734 & -0.201 \\ 0.764 & -0.089 \end{pmatrix}.$$

Diese weist offensichtlich keine Einfachstruktur auf. Die folgende durch Rotation aus $\hat{\mathbf{L}}$ gewonnene Ladungsmatrix $\check{\mathbf{L}}$ hingegen weist Einfachstruktur auf:

$$\check{\mathbf{L}} = \begin{pmatrix} 0.702 & 0.079 \\ 0.270 & 0.643 \\ 0.085 & 0.810 \\ 0.813 & 0.219 \\ 0.744 & 0.161 \\ 0.718 & 0.275 \end{pmatrix}. \tag{9.43}$$

Der erste Faktor hat hohe Ladungen bei den Merkmalen `Veranschaulichung von Fehlfunktionen`, `Qualitätsverbesserung`, `Entwicklungszeitverkürzung` und `Kostenreduktion`, während die Ladungen bei den beiden anderen Merkmalen niedrig sind. Aus Sicht von Bödeker & Franke (2001) bezieht sich dieser Faktor auf den Produktionsbereich. Der zweite Faktor hingegen hat hohe Ladungen bei den Merkmalen `Ermittlung von Kundenanforderungen` und `Angebotserstellung`, während die Ladungen der anderen Merkmale niedrig sind. Bödeker & Franke (2001) ordnen diesen Faktor dem Verkauf zu. □

Das Beispiel zeigt, wie leicht eine Einfachstruktur interpretiert werden kann. Wie kann man eine Einfachstruktur finden? Im Folgenden sei $\hat{\mathbf{L}}$ die Ladungsmatrix und $\mathbf{T}$ die Rotationsmatrix. Für die rotierte Ladungsmatrix gilt

$$\check{\mathbf{L}} = \hat{\mathbf{L}}\mathbf{T}.$$

Die Einfachstruktur sollte sich in der Matrix $\check{\mathbf{L}}$ zeigen. Seien $\check{l}_{ij}$ die Elemente von $\check{\mathbf{L}}$. Kaiser (1958) hat vorgeschlagen, die $\check{l}_{ij}$ so zu bestimmen, dass

$$\sum_{j=1}^{k} \sum_{i=1}^{p} \left(\check{l}_{ij}^2 - \overline{\check{l}_{\cdot j}^2} \right)^2$$

mit

$$\overline{\check{l}_{\cdot j}^2} = \frac{1}{p} \sum_{i=1}^{p} \check{l}_{ij}^2$$

maximal wird. Man spricht auch vom *Varimax-Kriterium*, da die Varianz der Ladungsquadrate eines Faktors maximiert wird. Bei einem Faktor wird die Varianz der quadrierten Ladungen maximal, wenn ein Teil der quadrierten Ladungen groß und der andere Teil klein wird. Dies wird dann über alle Faktoren ausbalanciert. Neuhaus & Wrigley (1954) haben das *Quartimax-Kriterium* vorgeschlagen. Bei diesem wird die Varianz der Ladungsquadrate einer Variablen maximiert. Es wird also folgender Ausdruck maximiert:

$$\sum_{i=1}^{p} \sum_{j=1}^{k} \left(\check{l}_{ij}^2 - \overline{\check{l}_{i \cdot}^2} \right)^2$$

mit

$$\overline{\check{l}_{i \cdot}^2} = \frac{1}{k} \sum_{j=1}^{k} \check{l}_{ij}^2.$$

Hierdurch soll erreicht werden, dass die Ladungen bei einigen Faktoren groß und bei den anderen klein sind. hmcounterend. (fortgesetzt)

Example 36. Die sich nach dem Varimax-Kriterium ergebende Faktorladungsmatrix ist in (9.43) zu finden. Die nach dem Quartimax-Kriterium gewonnene Faktorladungsmatrix sieht folgendermaßen aus:

$$\check{\mathbf{L}} = \begin{pmatrix} 0.706 & -0.008 \\ 0.347 & 0.605 \\ 0.184 & 0.793 \\ 0.834 & 0.117 \\ 0.758 & 0.069 \\ 0.747 & 0.185 \end{pmatrix}. \tag{9.44}$$

Wir sehen, dass beide Verfahren die gleiche Interpretation liefern. □

9.4 Faktorenanalyse in S-PLUS

Wir wollen die Daten in Tabelle 1.9 auf Seite 9 mit der Faktorenanalyse in S-PLUS analysieren. Die Korrelationsmatrix möge in der Variablen `rnutzen` stehen:

```
> rnutzen
           Fehler Kunden Angebot Qualitaet  Zeit Kosten
    Fehler  1.000  0.223   0.133     0.625 0.506  0.500
    Kunden  0.223  1.000   0.544     0.365 0.320  0.361
   Angebot  0.133  0.544   1.000     0.248 0.179  0.288
 Qualitaet  0.625  0.365   0.248     1.000 0.624  0.630
      Zeit  0.506  0.320   0.179     0.624 1.000  0.625
    Kosten  0.500  0.361   0.288     0.630 0.625  1.000
```

In S-PLUS gibt es eine Funktion `factanal`, mit der man eine Faktorenanalyse durchführen kann. Bevor wir uns diese aber näher anschauen, wollen wir die Anzahl der Faktoren mit den drei Verfahren aus Kapitel 9.3.1 ermitteln. Wir bestimmen zuerst mit der Funktion `eigen` die Eigenwerte der empirischen Korrelationsmatrix:

```
> eigen(rnutzen)[[1]]
[1] 3.12635 1.21132 0.52972 0.45244 0.35870 0.32145
```

Zwei der Eigenwerte sind größer als 1. Für die beiden anderen Kriterien benötigen wir $\mathbf{R} - \hat{\boldsymbol{\Psi}}$. Wir wollen diese Matrix `rh` nennen. Wir initialisieren `rh` mit `rnutzen`:

```
> rh<-rnutzen
```

Dann bestimmen wir die Bestimmtheitsmaße der Regressionen der einzelnen Variablen auf alle anderen Variablen:

```
> r2<-1-1/diag(solve(rnutzen))
> r2
[1] 0.42240 0.36613 0.31095 0.57713 0.49330 0.51063
```

Wir ersetzen die Hauptdiagonalelemente von $\mathbf{R}$ durch die Komponenten von `r2`:

```
> diag(rh)<-r2
```

Dann schauen wir uns die Eigenwerte dieser Matrix an:

```
> ew<-eigen(rh)[[1]]
> ew
[1]  2.605660078  0.573116835 -0.005630934 -0.118488447
    -0.147746771 -0.226342075
```

Zwei Eigenwerte sind größer als 0.

Nun bestimmen wir noch den Index von `ew`, bei dem die kumulierte Summe der Eigenwerte größer als die Summe aller Eigenwerte ist:

```
> min((1:length(ew))[cumsum(ew)>sum(ew)])
[1] 2
```

Dabei bildet die Funktion `cumsum` die kumulierte Summe der Komponenten eines Vektors.

Schauen wir uns nun die Funktion `factanal` an. Der Aufruf von `factanal` ist

```
factanal(x, factors=1, method="principal", data=NULL,
         covlist=NULL, scores=T, type="regression",
         rotation="varimax", na.action, subset,
         start=<<see below>>, control=NULL, ...)
```

Wir betrachten die Argumente, die sich auf Charakteristika beziehen, mit denen wir uns befasst haben. Liegen die Daten in Form einer Datenmatrix vor, so weist man diese beim Aufruf dem Argument `x` zu. In diesem Fall ist es wie auch bei der Hauptkomponentenanalyse möglich, Scores zu bestimmen. Da wir uns hiermit nicht beschäftigt haben, gehen wir auch nicht auf die Argumente `scores` und `type` ein, die die Berechnung der Scores ermöglichen. Liegen die Daten in Form einer empirischen Varianz-Kovarianz-Matrix oder empirischen Korrelationsmatrix vor, so verwendet man das Argument `covlist`. Hier geht man genauso wie bei der Hauptkomponentenanalyse auf Seite 149 vor. Wir schauen uns dies gleich am Beispiel an. Durch das Argument `factors` wird die Anzahl der Faktoren festgelegt. Die Schätzmethode wird durch das Argument `method` mit den Optionen `"principal"` und `"mle"` festgelegt. Man kann auch schon beim Schätzen das Verfahren der Rotation mit dem Argument `"rotation"` wählen. Standardmäßig wird Varimax angewendet. Soll nicht rotiert werden, setzt man `rotation` auf `"none"`. Varimax erhält man durch `"varimax"`. Es sind noch eine Reihe anderer Verfahren der Rotation möglich. Das Ergebnis der Funktion `factanal` ist eine Liste. Schauen wir uns die relevanten Komponenten am Beispiel an. Dabei führen wir eine Maximum-Likelihood-Faktorenanalyse mit zwei Faktoren und keiner Rotation auf Basis der empirischen Korrelationsmatrix durch. Wir müssen zuerst das Argument `covlist` bilden:

```
> cov.obj<-list(cov=rnutzen,center=rep(0,6))
```

Nun rufen wir die Funktion `factanal` auf:

```
> e<-factanal(covlist=cov.obj, factors=2,
              method="mle", rotation="none")
```

Die Faktorladungsmatrix erhalten wir durch

```
> e$loadings
          Factor1 Factor2
   Fehler  0.658  -0.255
   Kunden  0.537   0.445
  Angebot  0.451   0.679
Qualitaet  0.822  -0.183
     Zeit  0.734  -0.203
   Kosten  0.764
```

Die spezifischen Varianzen finden wir durch

```
> e$uniquenesses
    Fehler    Kunden  Angebot Qualitaet      Zeit    Kosten
 0.5014281 0.5133394 0.336325 0.2906119 0.4205383 0.4079054
```

Nun wollen wir noch die Faktorladungsmatrix rotieren. Dies geschieht mit der Funktion `rotate`. Der folgende Aufruf liefert die mit Varimax rotierte Faktorladungsmatrix:

```
> rotate(e$loadings,rotation="varimax")$rmat
          Factor1 Factor2
   Fehler 0.702
   Kunden 0.270   0.643
  Angebot         0.810
Qualitaet 0.813   0.219
     Zeit 0.744   0.161
   Kosten 0.718   0.275
```

Wir sehen, dass die Faktoren nun einfach interpretiert werden können.

9.5 Ergänzungen und weiterführende Literatur

Viele weitere Aspekte der explorativen Faktorenanalyse sind bei Basilevsky (1994) und Jackson (1991) zu finden. Im Gegensatz zur explorativen Faktorenanalyse geht man bei der konfirmatorischen Faktorenanalyse von einem Modell aus, das den Zusammenhang zwischen den Variablen beschreibt. In diesem wird auch die Anzahl der latenten Variablen vorgegeben. Die Parameter des Modells werden geschätzt und die Angemessenheit des Modells wird überprüft. Eine hervorragende Einführung in die konfirmatorische Faktorenanalyse gibt Bollen (1989).

9.6 Übungen

Exercise 21. Die Korrelationen zwischen den Variablen Y_1, Y_2 ,Y_3, Y_4 und Y_5 sollen durch einen Faktor F erklärt werden. Es gilt

$$\begin{aligned} Y_1 &= 0.65\,F + \epsilon_1, \\ Y_2 &= 0.84\,F + \epsilon_2, \\ Y_3 &= 0.70\,F + \epsilon_3, \\ Y_4 &= 0.32\,F + \epsilon_4, \\ Y_5 &= 0.28\,F + \epsilon_5. \end{aligned}$$

Es sollen die üblichen Annahmen der Faktorenanalyse gelten.

1. Geben Sie diese Annahmen an und interpretieren Sie sie.
2. Bestimmen Sie die Kommunalitäten des Faktors F mit den einzelnen Variablen.
3. Bestimmen Sie die spezifischen Varianzen der einzelnen Variablen.
4. Bestimmen Sie die Korrelationen zwischen den Variablen.

Exercise 22. Die Korrelationen zwischen den Variablen Y_1, Y_2 und Y_3 sind in der folgenden Korrelationsmatrix zu finden:

$$\mathbf{P} = \begin{pmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{pmatrix}$$

Zeigen Sie, dass die Korrelationsmatrix durch das folgende Modell erzeugt werden kann:

$$\begin{aligned} Y_1 &= 0.9\,F + \epsilon_1 \\ Y_2 &= 0.7\,F + \epsilon_2 \\ Y_3 &= 0.5\,F + \epsilon_3 \end{aligned}$$

mit

$$Var(F) = 1$$

und

$$Cov(\epsilon_i, F) = 0 \quad \text{für} \quad i = 1, 2, 3\,.$$

Wie groß sind die Kommunalitäten und die spezifischen Varianzen?

Exercise 23. Die Korrelationen zwischen den Variablen Y_1, Y_2 und Y_3 sind in der folgenden Korrelationsmatrix zu finden:

$$\mathbf{P} = \begin{pmatrix} 1 & 0.4 & 0.9 \\ 0.4 & 1 & 0.7 \\ 0.9 & 0.7 & 1 \end{pmatrix}.$$

Zeigen Sie, dass es eine eindeutige Lösung von

$$\mathbf{R} = \begin{pmatrix} \lambda_1^2 + \psi_1 & \lambda_1 \lambda_2 & \lambda_1 \lambda_3 \\ \lambda_2 \lambda_1 & \lambda_2^2 + \psi_2 & \lambda_2 \lambda_3 \\ \lambda_3 \lambda_1 & \lambda_3 \lambda_2 & \lambda_3^2 + \psi_3 \end{pmatrix}$$

gibt, die aber nicht zulässig ist, da $\psi_3 < 0$ gilt. Man nennt dies den Heywood-Fall.

Exercise 24. Bödeker & Franke (2001) beschäftigen sich in Ihrer Diplomarbeit mit den Möglichkeiten und Grenzen von Virtual-Reality-Technologien auf industriellen Anwendermärkten. Hierbei führten Sie eine Befragung bei Unternehmen durch, in der Sie unter anderem die Anforderungen ermittelten, die Unternehmen an ein Virtual-Reality-System stellen. Auf einer Skala von 1 bis 5 sollte dabei angegeben werden, wie wichtig die Merkmale `Simulation`, `Audiounterstützung`, `Internetfähigkeit`, `Detailtreue` und `Realitätsnähe` sind. In Tabelle 9.3 sind die Korrelationen zwischen den Merkmalen zu finden.

Table 9.3. Korrelationen zwischen Merkmalen

	Simulation	Audio	Internet	Detail	Real
Simulation	1	0.354	0.314	0.231	0.333
Audio	0.354	1	0.437	0.156	0.271
Internet	0.314	0.437	1	0.139	0.303
Detail	0.231	0.156	0.139	1	0.622
Real	0.333	0.271	0.303	0.622	1

Dabei kürzen wir `Audiounterstützung` durch `Audio`, `Internetfähigkeit` durch `Internet`, `Detailtreue` durch `Detail` und `Realitätsnähe` durch `Real` ab. Es soll eine Faktorenanalyse durchgeführt werden.

1. Betrachten Sie zunächst die Variablen Y_1, Y_2 und Y_3. Die Korrelationen zwischen diesen Variablen sollen durch einen Faktor F erklärt werden. Es sollen die üblichen Annahmen der Faktorenanalyse gelten.
   a) Geben Sie das Modell und die Annahmen an.
   b) Interpretieren Sie die Annahmen.
   c) Bestimmen Sie die Kommunalitäten des Faktors F mit den einzelnen Variablen.
   d) Bestimmen Sie die spezifischen Varianzen der einzelnen Variablen.

2. Nun soll die gesamte Korrelationsmatrix durch zwei Faktoren erklärt werden.
   a) Die Hauptfaktorenanalyse liefert folgende Ladungsmatrix:

$$\hat{\mathbf{L}} = \begin{pmatrix} 0.50 & 0.19 \\ 0.53 & 0.42 \\ 0.51 & 0.38 \\ 0.65 & -0.47 \\ 0.76 & -0.27 \end{pmatrix}.$$

   Wie groß sind die spezifischen Varianzen der Variablen?
   b) Nach der Rotation der Ladungsmatrix mit Varimax erhält man folgende Ladungsmatrix:

$$\check{\mathbf{L}} = \begin{pmatrix} 0.24 & 0.47 \\ 0.10 & 0.67 \\ 0.12 & 0.63 \\ 0.79 & 0.09 \\ 0.75 & 0.31 \end{pmatrix}.$$

   Interpretieren Sie die beiden Faktoren.
3. Führen Sie mit `S-PLUS` eine Maximum-Likelihood-Faktorenanalyse für die Korrelationsmatrix in Tabelle 9.3 durch. Testen Sie, ob die Korrelationen durch einen Faktor erklärt werden können.

10 Hierarchische loglineare Modelle

10.1 Problemstellung und Grundlagen

Im Kapitel 2.2.2 haben wir uns mit der Darstellung von Datensätzen befasst, die qualitative Merkmale enthalten. Wir haben gelernt, die Häufigkeitsverteilung von mehreren qualitativen Merkmalen in einer Kontingenztabelle zusammenzustellen. Wir wollen uns nun mit Modellen beschäftigen, die die Abhängigkeitsstruktur zwischen den Merkmalen beschreiben, und zeigen, wie man ein geeignetes Modell auswählen kann. Wir betrachten zunächst eine Grundgesamtheit, in der bei jedem Objekt zwei qualitative Merkmale A und B mit den Merkmalsausprägungen $A_1, \ldots, A_I$ und $B_1, \ldots, B_J$ von Interesse sind. Sei $P(A_i, B_j)$ die Wahrscheinlichkeit, dass ein zufällig aus der Grundgesamtheit ausgewähltes Objekt die Merkmalsausprägung A_i beim Merkmal A und die Merkmalsausprägung B_j beim Merkmal B aufweist.

Für $i = 1, \ldots, I$ gilt

$$P(A_i) = \sum_{j=1}^{J} P(A_i, B_j)\,.$$

Für $j = 1, \ldots, J$ gilt

$$P(B_j) = \sum_{i=1}^{I} P(A_i, B_j)\,.$$

Example 37. Wir betrachten die Merkmale `Geschlecht` A und `Interesse an Fußball` B in einer Population von 100 Personen. Von diesen sind 45 weiblich. 15 Frauen und 45 Männer sind interessiert am Fußball. Bezeichnen wir `weiblich` mit A_1, `männlich` mit A_2, `interessiert an Fußball` mit B_1 und `nicht interessiert an Fußball` mit B_2, so gilt

$$P(A_1, B_1) = 0.15,\ \ P(A_1, B_2) = 0.30,$$
$$P(A_2, B_1) = 0.45,\ \ P(A_2, B_2) = 0.10\,.$$

Somit gilt

$$P(A_1) = 0.45, \quad P(A_2) = 0.55, \quad P(B_1) = 0.60, \quad P(B_2) = 0.40\,.$$

□

Wir sind interessiert, die Abhängigkeitsstruktur zwischen A und B zu modellieren. Hierzu schauen wir uns die Verteilung des Merkmals B für die Ausprägungen $A_1, \ldots, A_I$ des Merkmals A an.

Definition 17. *Seien A und B Merkmale mit den Merkmalsausprägungen $A_1, \ldots, A_I$ und $B_1, \ldots, B_J$. Die bedingte Wahrscheinlichkeit von B_j gegeben A_i ist für $i = 1, \ldots, I$, $j = 1, \ldots, J$ definiert durch*

$$P(B_j|A_i) = \frac{P(A_i, B_j)}{P(A_i)}, \tag{10.1}$$

falls $P(A_i) > 0$ gilt. Ansonsten ist $P(B_j|A_i)$ gleich 0.

hmcounterend. (fortgesetzt)

Example 37. Es gilt

$$P(B_1|A_1) = \frac{1}{3}, \;\; P(B_2|A_1) = \frac{2}{3},$$
$$P(B_1|A_2) = \frac{9}{11}, \;\; P(B_2|A_2) = \frac{2}{11}\,.$$

□

Definition 18. *Die Merkmale A und B mit den Merkmalsausprägungen $A_1, \ldots, A_I$ und $B_1, \ldots, B_J$ heißen unabhängig, wenn für $i = 1, \ldots, I$, $j = 1, \ldots, J$ gilt*

$$P(B_j|A_i) = P(B_j)\,. \tag{10.2}$$

Aus (10.1) und (10.2) folgt, dass die Merkmale A und B genau dann unabhängig sind, wenn für $i = 1, \ldots, I$, $j = 1, \ldots, J$ gilt

$$P(A_i, B_j) = P(A_i)P(B_j)\,. \tag{10.3}$$

Wir wollen nun die Abhängigkeitsstruktur zwischen A und B durch eine Maßzahl beschreiben, die die Interpretation bestimmter loglinearer Modelle erleichtert. Hierzu betrachten wir den Fall $I = 2$ und $J = 2$. Schauen wir uns zunächst nur ein Merkmal an.

Definition 19. *Sei A ein Merkmal mit den Merkmalsausprägungen A_1 und A_2. Das Verhältnis*

$$\frac{P(A_1)}{P(A_2)} \tag{10.4}$$

nennt man Wettchance 1.Ordnung (vgl. Fahrmeir et al. (1996), S. 548).

Über Wettchancen 1.Ordnung werden bei Sportwetten die Quoten festgelegt.
hmcounterend. (fortgesetzt)

Example 37. Es gilt

$$\frac{P(A_1)}{P(A_2)} = \frac{0.45}{0.55} = \frac{9}{11}$$

und

$$\frac{P(B_1)}{P(B_2)} = \frac{0.6}{0.4} = 1.5\,.$$

□

Mit (10.4) können wir eine Maßzahl gewinnen, die den Zusammenhang zwischen zwei Merkmalen beschreibt. Man bestimmt die Wettchance 1.Ordnung von B, wenn die Merkmalsausprägung A_1 von A vorliegt:

$$\frac{P(B_1|A_1)}{P(B_2|A_1)} \tag{10.5}$$

und die Wettchance 1.Ordnung von B, wenn die Merkmalsausprägung A_2 von A vorliegt:

$$\frac{P(B_1|A_2)}{P(B_2|A_2)}\,. \tag{10.6}$$

hmcounterend. (fortgesetzt)

Example 37. Es gilt

$$\frac{P(B_1|A_1)}{P(B_2|A_1)} = \frac{\frac{1}{3}}{\frac{2}{3}} = 0.5$$

und

$$\frac{P(B_1|A_2)}{P(B_2|A_2)} = \frac{\frac{9}{11}}{\frac{2}{11}} = 4.5\,.$$

Wir sehen, dass die Wettchancen des Merkmals `Interesse an Fußball` sich bei den Frauen und Männern beträchtlich unterscheiden. □

Unterscheiden sich (10.5) und (10.6), so unterscheidet sich die Verteilung des Merkmals B für die beiden Kategorien des Merkmals A.

Definition 20. *Sei A ein Merkmal mit den Merkmalsausprägungen A_1 und A_2 und B ein Merkmal mit den Merkmalsausprägungen B_1 und B_2. Das Verhältnis*

$$\theta = \frac{P(B_1|A_1)/P(B_2|A_1)}{P(B_1|A_2)/P(B_2|A_2)} \tag{10.7}$$

heißt das Kreuzproduktverhältnis θ.

hmcounterend. (fortgesetzt)

Example 37. Es gilt

$$\theta = \frac{0.5}{4.5} = \frac{1}{9}\,.$$

□

Es gilt

$$\theta = \frac{P(A_1, B_1)P(A_2, B_2)}{P(A_1, B_2)P(A_2, B_1)}\,. \tag{10.8}$$

Dies sieht man folgendermaßen:

$$\theta = \frac{P(B_1|A_1)/P(B_2|A_1)}{P(B_1|A_2)/P(B_2|A_2)} = \frac{P(B_1|A_1)P(B_2|A_2)}{P(B_1|A_2)P(B_2|A_1)}$$

$$= \frac{\frac{P(A_1,B_1)}{P(A_1)}\frac{P(A_2,B_2)}{P(A_2)}}{\frac{P(A_1,B_2)}{P(A_1)}\frac{P(A_2,B_1)}{P(A_2)}} = \frac{P(A_1, B_1)P(A_2, B_2)}{P(A_1, B_2)P(A_2, B_1)}\,.$$

Das folgende Theorem zeigt, dass man am Kreuzproduktverhältnis erkennen kann, ob zwei Merkmale unabhängig sind.

Theorem 11. *Seien A und B zwei Merkmale mit Merkmalsausprägungen A_1 und A_2 beziehungsweise B_1 und B_2 und zugehörigen Wahrscheinlichkeiten $P(A_i, B_j)$ für $i = 1, 2$ und $j = 1, 2$. Das Kreuzproduktverhältnis θ ist genau dann gleich 1, wenn A und B unabhängig sind.*

Beweis:

Aus Gründen der Übersichtlichkeit setzen wir für $i = 1, 2$ und $j = 1, 2$:

$$p_{ij} = P(A_i, B_j), \quad p_{i.} = P(A_i), \quad p_{.j} = P(B_j)\,.$$

Sind A und B unabhängig, so gilt (10.3), also

$$p_{ij} = p_{i.}p_{.j}\,.$$

Somit gilt

$$\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{p_{1.}p_{.1}p_{2.}p_{.2}}{p_{1.}p_{.2}p_{2.}p_{.1}} = 1\,.$$

Sei $\theta = 1$. Es gilt also

$$p_{11}p_{22} = p_{12}p_{21}\,. \tag{10.9}$$

Wir addieren auf beiden Seiten von Gleichung (10.9) den Ausdruck

$$p_{11}(p_{11} + p_{12} + p_{21}) = p_{11}^2 + p_{11}p_{12} + p_{11}p_{21}$$

und erhalten folgende Gleichung

$$p_{11}^2 + p_{11}p_{12} + p_{11}p_{21} + p_{11}p_{22} = p_{11}^2 + p_{11}p_{12} + p_{11}p_{21} + p_{12}p_{21}\,.$$

Diesen können wir umformen zu

$$p_{11}(p_{11} + p_{12} + p_{21} + p_{22}) = p_{11}(p_{11} + p_{21}) + p_{12}(p_{11} + p_{21})\,.$$

Mit

$$p_{11} + p_{12} + p_{12} + p_{22} = 1$$

gilt also

$$p_{11} = (p_{11} + p_{12})(p_{11} + p_{21}) = p_{1\cdot}p_{\cdot 1}\,.$$

Entsprechend können wir zeigen

$$\begin{aligned} p_{12} &= p_{1\cdot}p_{\cdot 2}, \\ p_{21} &= p_{2\cdot}p_{\cdot 1}, \\ p_{22} &= p_{2\cdot}p_{\cdot 2}\,. \end{aligned}$$

Also sind A und B unabhängig.

Ist das Kreuzproduktverhältnis also 1, so sind die Merkmale A und B unabhängig. Ist es aber ungleich 1, so sind sie abhängig. Wir haben das Kreuzproduktverhältnis nur für $I = 2$ und $J = 2$ betrachtet. Agresti (1990), S. 18-19 beschreibt, wie man es für Merkmale mit mehr als zwei Merkmalsausprägungen erweitern kann.

Bisher haben wir die Abhängigkeitsstruktur unter der Annahme betrachtet, dass alle Informationen über die Grundgesamtheit vorliegen. Normalerweise ist dies nicht der Fall, und man wird eine Zufallsstichprobe vom Umfang n aus der Grundgesamtheit ziehen und die Daten in einer Kontingenztabelle zusammenstellen. Wir bezeichnen die absolute Häufigkeit für das gleichzeitige Auftreten der Merkmalsausprägungen A_i und B_j mit n_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Außerdem ist

$$n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$$

für $i = 1, \ldots, I$, und

$$n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$$

für $j = 1, \ldots, J$.

Example 38. Im Beispiel 9 auf Seite 9 ist das Ergebnis einer Befragung von Studenten zu finden. Diese wurden nach ihrem Geschlecht, ihrem Studienfach und ihrem Wahlverhalten gefragt. Schauen wir uns zunächst die Merkmale `Studienfach` A und `Wahlverhalten` B an. Im Folgenden entspricht `BWL` A_1, `VWL` A_2, `CDU` B_1 und `SPD` B_2. Tabelle 10.1 zeigt die Kontingenztabelle.

Table 10.1. Studienfach und Wahlverhalten bei Studenten

Studienfach \ Wahlverhalten	CDU	SPD
BWL	50	36
VWL	6	8

Es gilt

$$n_{11} = 50, \qquad n_{12} = 36, \qquad n_{21} = 6, \qquad n_{22} = 8$$

und

$$n_{1.} = 86, \qquad n_{2.} = 14, \qquad n_{.1} = 56, \qquad n_{.2} = 44.$$

□

Die unbekannten Wahrscheinlichkeiten $P(A_i, B_j)$ schätzen wir durch die relativen Häufigkeiten:

$$\hat{P}(A_i, B_j) = \frac{n_{ij}}{n}. \tag{10.10}$$

hmcounterend. (fortgesetzt)

Example 38. Es gilt

$$\hat{P}(A_1, B_1) = 0.50, \;\; \hat{P}(A_1, B_2) = 0.36,$$
$$\hat{P}(A_2, B_1) = 0.06, \;\; \hat{P}(A_2, B_2) = 0.08.$$

□

Wir schätzen das Kreuzproduktverhältnis durch

$$\hat{\theta} = \frac{\hat{P}(A_1, B_1)\hat{P}(A_2, B_2)}{\hat{P}(A_1, B_2)\hat{P}(A_2, B_1)}. \tag{10.11}$$

hmcounterend. (fortgesetzt)

Example 38. Es gilt

$$\hat{\theta} = \frac{0.5 \cdot 0.08}{0.36 \cdot 0.06} = 1.85.$$

□

Es stellt sich die Frage, ob $\hat{\theta}$ signifikant von 1 verschieden ist, die Merkmale also abhängig sind. Das Testproblem lautet

H_0: Die Merkmale A und B sind unabhängig,

H_1: Die Merkmale A und B sind nicht unabhängig.

Tests, die auf dem Kreuzproduktverhältnis beruhen, sind bei Agresti (1990), S. 54 ff. zu finden. Wir verwenden das Kreuzproduktverhältnis nur zur Beschreibung der Abhängigkeitsstruktur und betrachten den χ^2-Unabhängigkeitstest, um H_0 zu überprüfen. Bei diesem werden die beobachteten Häufigkeiten n_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ mit den Häufigkeiten verglichen, die man für das gleichzeitige Auftreten von A_i und B_j erwartet, wenn H_0 zutrifft. Trifft H_0 zu, so gilt

$$P(A_i, B_j) = P(A_i)\, P(B_j)$$

für $i = 1, \ldots, I$ und $j = 1, \ldots, J$. Multiplizieren wir diesen Ausdruck mit n, so erhalten wir die *erwarteten absoluten Häufigkeiten*

$$n\, P(A_i, B_j) = n\, P(A_i)\, P(B_j). \tag{10.12}$$

Die Wahrscheinlichkeiten $P(A_i)$ und $P(B_j)$ sind unbekannt. Wir schätzen sie durch die entsprechenden relativen Häufigkeiten. Wir schätzen $P(A_i)$ durch $n_{i.}/n$ und $P(B_j)$ durch $n_{.j}/n$. Setzen wir diese Schätzer in (10.12) ein, so erhalten wir die folgenden *geschätzten erwarteten Häufigkeiten*, die wir mit $\hat{n}_{ij}$ bezeichnen:

$$\hat{n}_{ij} = n \cdot \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}.$$

Dies können wir vereinfachen zu

$$\hat{n}_{ij} = \frac{n_{i.}\, n_{.j}}{n}. \tag{10.13}$$

hmcounterend. (fortgesetzt)

Example 38. Die geschätzten erwarteten Häufigkeiten sind

$$\begin{aligned}
\hat{n}_{11} &= \frac{86 \cdot 56}{100} = 48.16, \\
\hat{n}_{12} &= \frac{86 \cdot 44}{100} = 37.84, \\
\hat{n}_{21} &= \frac{14 \cdot 56}{100} = 7.84, \\
\hat{n}_{22} &= \frac{14 \cdot 44}{100} = 6.16\,.
\end{aligned}$$

□

Die Teststatistik des χ^2-Unabhängigkeitstests lautet

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}. \tag{10.14}$$

hmcounterend. (fortgesetzt)

Example 38. Es gilt

$$X^2 = \frac{(50 - 48.16)^2}{48.16} + \frac{(36 - 37.84)^2}{37.84} + \frac{(6 - 7.84)^2}{7.84} + \frac{(8 - 6.16)^2}{6.16} = 1.1412.$$

□

Trifft H_0 zu, so ist X^2 approximativ χ^2-verteilt mit $(I-1)(J-1)$ Freiheitsgraden. Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt $X^2 \geq \chi^2_{(I-1)(J-1);1-\alpha}$, wobei $\chi^2_{(I-1)(J-1);1-\alpha}$ das $1-\alpha$-Quantil der χ^2-Verteilung mit $(I-1)(J_1)$ ist. hmcounterend. (fortgesetzt)

Example 38. Sei $\alpha = 0.05$. Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{1;0.95} = 3.84$. Wir lehnen H_0 also nicht ab. □

Zwei Merkmale sind entweder unabhängig oder abhängig. Bei drei Merkmalen wird es komplizierter. Wir betrachten eine Grundgesamtheit, in der bei jedem Objekt drei Merkmale A, B und C mit den Merkmalsausprägungen $A_1, \ldots, A_I$, $B_1, \ldots, B_J$ und $C_1, \ldots, C_K$ von Interesse sind. Sei $P(A_i, B_j, C_k)$ die Wahrscheinlichkeit, dass ein zufällig aus der Grundgesamtheit ausgewähltes Objekt die Merkmalsausprägung A_i beim Merkmal A, die Merkmalsausprägung B_j beim Merkmal B und die Merkmalsausprägung C_k beim Merkmal C aufweist.

Example 39. Wir schauen uns wieder das Beispiel 9 auf Seite 9 an und berücksichtigen jetzt alle drei Merkmale. Das Merkmal A sei das `Studienfach`, das Merkmal B das `Wahlverhalten` und das Merkmal C das `Geschlecht`. □

Eine Möglichkeit, die Abhängigkeitsstruktur zwischen drei Merkmalen A, B und C herauszufinden, besteht darin, die Abhängigkeitsstruktur zwischen jeweils zwei Merkmalen zu untersuchen. Man überprüft also, ob die Merkmale paarweise unabhängig sind. Liegt eine Zufallsstichprobe aus der Grundgesamtheit vor, so kann man die Hypothesen mit dem χ^2-Unabhängigkeitstest überprüfen. Bei drei Merkmalen gibt es drei Paare von Merkmalen, sodass man drei Tests durchführen muss. hmcounterend. (fortgesetzt)

Example 39. Wir fassen die Bobachtungen als Zufallsstichprobe auf. Die Merkmale `Studienfach` und `Wahlverhalten` haben wir bereits untersucht. Bei den Merkmalen `Geschlecht` und `Studienfach` gilt $X^2 = 0.7475$. Bei den Merkmalen `Geschlecht` und `Wahlverhalten` gilt $X^2 = 6.86$. □

Alle Paare von Merkmalen zu untersuchen ist aus einer Reihe von Gründen nicht unproblematisch. Im Beispiel führen wir drei Tests am gleichen Datensatz durch. Man spricht von einem *multiplen Testproblem*. Bei einem multiplen Testproblem begeht man einen Fehler 1. Art, wenn mindestens eine der wahren Nullhypothesen abgelehnt wird. Führen wir jeden Test zum Niveau α durch, so wird die Wahrscheinlichkeit für den Fehler 1. Art im multiplen Testproblem größer als α sein. Bei k Tests wird das vorgegebene multiple Niveau

α nicht überschritten, wenn man jeden der Tests zum Niveau α/k durchführt. Ein Beweis dieser Tatsache ist bei Schlittgen (1996), S.383 zu finden. Man spricht vom *Bonferroni-Test*. Die Verkleinerung des Signifikanzniveaus der Einzeltests vermindert die Güte der Tests. hmcounterend. (fortgesetzt)

Example 39. Es gilt $0.05/3 = 0.0167$. Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{1;0.983} = 5.73$. Wir lehnen also die Hypothesen, dass `Studienfach` und `Wahlverhalten` und dass `Geschlecht` und `Studienfach` unabhängig sind, nicht ab. Die Hypothese, dass `Geschlecht` und `Wahlverhalten` unabhängig sind, lehnen wir ab. □

Kann man das multiple Testproblem noch in den Griff bekommen, so wird die ausschließliche Betrachtung der paarweisen Zusammenhänge in vielen Fällen der Abhängigkeitsstruktur nicht gerecht, da viele Abhängigkeitsstrukturen durch sie nicht erfasst werden. So folgt aus der paarweisen Unabhängigkeit der Ereignisse A, B und C nicht die vollständige Unabhängigkeit. Bei dieser gilt

$$P(A_i, B_j, C_k) = P(A_i)\, P(B_j)\, P(C_k)$$

für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$. Ein Beispiel hierfür ist bei Schlittgen (2000) auf Seite 82 zu finden. Es gibt noch weitere Abhängigkeitsstrukturen, die man in Betracht ziehen muss. Schauen wir uns auch hierfür ein Beispiel an.

Example 40. In einer Grundstudiumsveranstaltung wurden die Studenten unter anderem gefragt, ob sie den Film Titanic gesehen haben. Wir bezeichnen dieses Merkmal mit `Titanic`. Außerdem wurden Sie gebeten, den nachfolgenden Satz richtig zu vervollständigen:

> Zu Risiken und Nebenwirkungen ...

Dieses Merkmal bezeichnen wir mit `Satz`. In Tabelle 10.2 ist die Kontingenztabelle der drei Merkmale zu finden.

Table 10.2. Kontingenztabelle der Merkmale Geschlecht, Titanic und Satz

		Satz	
Geschlecht	Titanic	richtig	falsch
w	ja	64	16
	nein	14	6
m	ja	28	32
	nein	14	26

Führen wir für alle Paare von Merkmalen einen χ^2-Unabhängigkeitstest durch, so erhalten wir die Werte der Teststatistik in Tabelle 10.3.

Table 10.3. Werte des χ^2-Unabhängigkeitstests

Merkmale	X^2
`Geschlecht - Satz`	27.00
`Geschlecht - Titanic`	9.52
`Titanic - Satz`	6.35

Berücksichtigt man, dass es sich um ein multiples Testproblem handelt, so ist der kritische Wert gleich 5.73. Somit sind alle Paare von Merkmalen voneinander abhängig. Bei den Paaren `Geschlecht` und `Satz` und `Geschlecht` und `Titanic` ist dies nicht verwunderlich, aber woher kommt die Abhängigkeit zwischen den Merkmalen `Titanic` und `Satz`? Haben die Personen, die den Film Titanic gesehen haben, ein besseres Gedächtnis? Eine Antwort auf diese Frage erhalten wir, wenn wir die gemeinsame Verteilung der drei Merkmale aus einem anderen Blickwinkel anschauen. Wir betrachten die Merkmale `Satz` und `Titanic` zum einen bei den Studentinnen und zum anderen bei den Studenten. Der Wert von X^2 bei den Studentinnen ist gleich 0.93. Bei den Studenten beträgt er 1.34. Bei den Studentinnen und bei den Studenten besteht also kein Zusammenhang zwischen den Merkmalen `Titanic` und `Satz`. Aggregiert man über alle Personen, so sind die beiden Merkmale abhängig. Woran liegt die Abhängigkeit in der aggregierten Tabelle? Wir haben gesehen, dass die Merkmale `Geschlecht` und `Satz` und die Merkmale `Geschlecht` und `Titanic` abhängig sind. Schaut man sich die bedingten relativen Häufigkeiten an, so stellt man fest, dass die Chance, sich den Film Titanic anzusehen, bei den Studentinnen größer ist als bei den Studenten. Die Chance, den Satz richtig zu vollenden, ist ebenfalls bei den Studentinnen größer als bei den Studenten. Dies führt bei der Betrachtung der Merkmale `Titanic` und `Satz` dazu, dass die Personen, die Titanic gesehen haben, auch häufiger den Satz richtig vollenden können. □

Das Beispiel zeigt, dass die Merkmale A und B unter der Bedingung, dass die Merkmalsausprägungen von C festgehalten werden, unabhängig sein können, aber aggregiert abhängig sind. Man sagt, dass das Merkmal C den Zusammenhang zwischen den Merkmalen A und B erklärt. Sind die Merkmale A und B für die Ausprägungen des Merkmals C unabhängig, so liegt das Modell der bedingten Unabhängigkeit vor. In diesem gilt für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$:

$$P(A_i, B_j|C_k) = P(A_i|C_k)\, P(B_j|C_k) \tag{10.15}$$

für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$.

Wir haben eine Reihe von Modellen zur Beschreibung der Abhängigkeitsstruktur in einer dreidimensionalen Kontingenztabelle kennengelernt. Loglineare Modelle bieten die Möglichkeit, systematisch ein geeignetes Modell zu finden. Mit diesen werden wir uns in den nächsten Abschnitten beschäftigen. Da die

Theorie loglinearer Modelle an einer zweidimensionalen Kontingenztabelle am einfachsten veranschaulicht werden kann, beginnen wir mit diesem Fall.

10.2 Zweidimensionale Kontingenztabellen

Wir gehen davon aus, dass eine Zufallsstichprobe vom Umfang n aus einer Grundgesamtheit vorliegt, in der die Merkmale A und B von Interesse sind. Das Merkmal A besitze die Ausprägungen $A_1, \ldots, A_I$ und das Merkmal B die Ausprägungen $B_1, \ldots, B_J$. Die Wahrscheinlichkeit, dass ein zufällig aus der Grundgesamtheit ausgewähltes Objekt die Merkmalsausprägung A_i beim Merkmal A und die Merkmalsausprägung B_j beim Merkmal B aufweist, bezeichnen wir mit $P(A_i, B_j)$. Die Anzahl der Objekte mit Merkmalsausprägungen A_i und B_j in der Stichprobe bezeichnen wir mit n_{ij}. Wir stellen die Häufigkeiten in einer Kontingenztabelle zusammen.

Example 41. Wir betrachten wieder die Merkmale `Studienfach` und `Wahlverhalten` im Rahmen des Beispiels 9 auf Seite 9. Tabelle 10.4 zeigt die Kontingenztabelle.

Table 10.4. Studienfach und Wahlverhalten bei Studenten

Studienfach \ Wahlverhalten	CDU	SPD
BWL	50	36
VWL	6	8

□

Ziel ist es, ein Modell zu finden, das die Abhängigkeitsstruktur zwischen den beiden Merkmalen gut beschreibt. Wir betrachten eine Reihe von Modellen.

10.2.1 Modell 0

Das Modell 0 beruht auf folgenden Annahmen:

1. Das Merkmal A ist gleichverteilt:
$$P(A_i) = \frac{1}{I} \quad \text{für } i = 1, \ldots, I.$$
2. Das Merkmal B ist gleichverteilt:
$$P(B_j) = \frac{1}{J} \quad \text{für } j = 1, \ldots, J.$$
3. Die Merkmale A und B sind unabhängig:
$$P(A_i, B_j) = P(A_i)\, P(B_j) \quad \text{für } i = 1, \ldots, I, j = 1, \ldots, J.$$

Unter diesen Annahmen gilt

$$P(A_i, B_j) = \frac{1}{IJ}. \tag{10.16}$$

Dies sieht man folgendermaßen:

$$P(A_i, B_j) = P(A_i)\, P(B_j) = \frac{1}{IJ}.$$

Wenn das Modell 0 zutrifft, erwarten wir für das gleichzeitige Auftreten von A_i und B_j für $i = 1, \ldots, I$, $j = 1, \ldots, J$:

$$n\, P(A_i, B_j) = \frac{n}{IJ}.$$

Im Modell 0 müssen wir die erwarteten Häufigkeiten nicht schätzen. Wir bezeichnen sie aber trotzdem mit $\hat{n}_{ij}$. hmcounterend. (fortgesetzt)

Example 41. Tabelle 10.5 enthält die $\hat{n}_{ij}$.

Table 10.5. Geschätzte absolute Häufigkeiten im Modell 0

	CDU	SPD
BWL	25	25
VWL	25	25

□

Vergleichen wir die beobachteten Häufigkeiten mit den geschätzten Häufigkeiten, so sehen wir, dass die beobachteten Häufigkeiten n_{ij} nicht gut mit den geschätzten Häufigkeiten $\hat{n}_{ij}$ übereinstimmen. Zur Messung der Übereinstimmung kann man die Teststatistik des χ^2-Unabhängigkeitstests in Gleichung (10.14) verwenden. hmcounterend. (fortgesetzt)

Example 41. Es gilt

$$X^2 = 55.84.$$

□

Wir werden im Folgenden eine andere Teststatistik verwenden, da diese, wie wir später sehen werden, bessere Eigenschaften besitzt. Die *Likelihood-Quotienten-Teststatistik* zur Überprüfung eines Modells M ist definiert durch

$$G(M) = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \ln \frac{n_{ij}}{\hat{n}_{ij}}. \tag{10.17}$$

Dabei sind $\hat{n}_{ij}$ die erwarteten beziehungsweise geschätzten erwarteten Häufigkeiten des Auftretens von A_i und B_j unter der Annahme, dass das Modell M zutrifft. hmcounterend. (fortgesetzt)

Example 41. Es gilt

$$G(0) = 2\left[50 \ln \frac{50}{25} + 36 \ln \frac{36}{25} + 6 \ln \frac{6}{25} + 8 \ln \frac{8}{25}\right] = 60.21.$$

□

Man kann zeigen, dass G approximativ χ^2-verteilt ist mit $IJ - 1$ Freiheitsgraden, wenn das Modell 0 zutrifft. Wir können also testen

H_0: Das Modell 0 trifft zu,

H_1: Das Modell 0 trifft nicht zu.

hmcounterend. (fortgesetzt)

Example 41. Im Beispiel gilt $IJ - 1 = 3$. Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{3;0.95} = 7.82$. Wir lehnen das Modell 0 zum Signifikanzniveau 0.05 ab. □

Das Modell 0 liefert keine adäquate Beschreibung des Zusammenhangs zwischen den beiden Merkmalen. Schauen wir uns andere Modelle an, die wir erhalten, indem wir eine oder mehrere Forderungen fallenlassen, die das Modell 0 an den datengenerierenden Prozess stellt.

10.2.2 Modell *A*

Als erstes lassen wir die Annahme der Gleichverteilung von A fallen. Wir unterstellen also:

1. Das Merkmal B ist gleichverteilt:

   $P(B_j) = \dfrac{1}{J}$ für $j = 1, \ldots, J$.

2. Die Merkmale A und B sind unabhängig:

   $P(A_i, B_j) = P(A_i)\, P(B_j)$ für $i = 1, \ldots, I, j = 1, \ldots, J$.

Unter diesen Annahmen gilt

$$P(A_i, B_j) = \frac{1}{J} P(A_i). \tag{10.18}$$

Dies sieht man folgendermaßen:

$$P(A_i, B_j) = P(A_i)\, P(B_j) = \frac{1}{J} P(A_i).$$

Wir schätzen $P(A_i)$ durch die relative Häufigkeit $n_{i.}/n$ der i-ten Kategorie von A. Ersetzen wir $P(A_i)$ in

$$P(A_i, B_j) = \frac{1}{J} P(A_i)$$

durch $n_{i.}/n$, so erhalten wir die geschätzten Wahrscheinlichkeiten

$$\hat{P}(A_i, B_j) = \frac{n_{i.}}{n\,J}.$$

Wir erhalten somit als Schätzer für die erwartete Häufigkeit des gleichzeitigen Auftretens von A_i und B_j:

$$\hat{n}_{ij} = \frac{n_{i.}}{J}. \tag{10.19}$$

hmcounterend. (fortgesetzt)

Example 41. Tabelle 10.6 enthält die geschätzten Häufigkeiten.

Table 10.6. Geschätzte absolute Häufigkeiten im Modell A

	CDU	SPD
BWL	43	43
VWL	7	7

Der Wert der Likelihood-Quotienten-Teststatistik ist

$$G(A) = 2\left[50\,\ln\frac{50}{43} + 36\,\ln\frac{36}{43} + 6\,\ln\frac{6}{7} + 8\,\ln\frac{8}{7}\right] = 2.58.$$

□

Im Modell A ist $G(A)$ approximativ χ^2-verteilt mit $(J-1)I$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 41. Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{2;0.95} = 5.99$. Wir lehnen das Modell A zum Niveau 0.05 nicht ab. □

10.2.3 Der IPF-Algorithmus

hmcounterend. (fortgesetzt)

Example 41. Die Tabellen 10.7 und 10.8 zeigen die Tabellen 10.4 und 10.6, wobei in beiden Fällen die Randverteilungen angegeben sind.
Wir sehen, dass die geschätzte Randverteilung von A mit der beobachteten Randverteilung von A übereinstimmt. □

Der im Beispiel beobachtete Sachverhalt gilt generell im Modell A. Aus (10.19) folgt nämlich

$$\hat{n}_{i.} = \sum_{j=1}^{J} \hat{n}_{i.} = \sum_{j=1}^{J} \frac{n_{i.}}{J} = n_{i.} \sum_{j=1}^{J} \frac{1}{J} = n_{i.} J \frac{1}{J} = n_{i.}\,.$$

Table 10.7. Studienfach und Wahlverhalten bei Studenten mit Randverteilung

	CDU	SPD	
BWL	50	36	86
VWL	6	8	14
	56	44	100

Table 10.8. Geschätzte absolute Häufigkeiten im Modell A mit Randverteilung

	CDU	SPD	
BWL	43	43	86
VWL	7	7	14
	50	50	100

Man spricht deshalb vom Modell A. Ein Modell erhält immer den Namen der Randverteilungen, die festgehalten werden. Man sagt auch, dass die Verteilung von A angepasst wird. Im Modell A muss also gelten

$$\hat{n}_{i.} = n_{i.} \, .$$

Diese Forderung nimmt man als Ausgangspunkt für die Anpassung mit dem IPF-Algorithmus (Iteratively Proportional Fitting-Algorithmus). Schauen wir uns diesen für das Modell A an.
Wir gehen aus von

$$\hat{n}_{i.} = n_{i.} \, ,$$

multiplizieren diese Gleichung mit $\hat{n}_{ij}$ und erhalten

$$\hat{n}_{ij} \, \hat{n}_{i.} = \hat{n}_{ij} \, n_{i.} \, .$$

Hieraus folgt die Identität

$$\hat{n}_{ij} = \frac{n_{i.}}{\hat{n}_{i.}} \, \hat{n}_{ij} \, ,$$

auf der der Algorithmus beruht. Man geht aus von den Startwerten $\hat{n}_{ij}^{(0)}$. Dann werden die geschätzten Häufigkeiten folgendermaßen iterativ bestimmt:

$$\hat{n}_{ij}^{(1)} = \frac{n_{i.}}{\hat{n}_{i.}^{(0)}} \, \hat{n}_{ij}^{(0)} \, .$$

In der Regel setzt man

$$\hat{n}_{ij}^{(0)} = 1 \qquad \text{für } i = 1, \ldots, I, j = 1, \ldots, J.$$

Schauen wir uns dies für die Anpassung von Modell A an. Wir setzen

$$\hat{n}_{ij}^{(0)} = 1.$$

Hieraus folgt

$$\hat{n}_{i.}^{(0)} = \sum_{j=1}^{J} \hat{n}_{ij}^{(0)} = \sum_{j=1}^{J} 1 = J.$$

Also gilt

$$\hat{n}_{ij}^{(1)} = \frac{n_{i.}}{\hat{n}_{i.}^{(0)}} \hat{n}_{ij}^{(0)} = \frac{n_{i.}}{J}.$$

Dies sind die Bedingungen, die wir bereits kennen. Wir lassen den Index (1) weg und erhalten

$$\hat{n}_{ij} = \frac{n_{i.}}{J}.$$

Das Ergebnis stimmt mit dem überein, das wir weiter oben bereits entwickelt haben. Jedes loglineare Modell ist charakterisiert durch die Randverteilungen, die angepasst werden. Die Anpassung erfolgt mit dem IPF-Algorithmus, wobei dieser gegebenenfalls iteriert werden muss. Wir werden bei den einzelnen Modellen den IPF-Algorithmus anwenden.

10.2.4 Modell *B*

Anstatt der Randverteilung von A können wir auch die Randverteilung von B festhalten. Man spricht dann vom Modell B. Wir unterstellen also:

1. Das Merkmal A ist gleichverteilt:

   $P(A_i) = \dfrac{1}{I}$ für $i = 1, \ldots, I$.

2. Die Merkmale A und B sind unabhängig:

   $P(A_i, B_j) = P(A_i)\, P(B_j)$ für $i = 1, \ldots, I, j = 1, \ldots, J$.

Unter diesen Annahmen gilt

$$P(A_i, B_j) = \frac{1}{I}\, P(B_j)\,.$$

Dies sieht man folgendermaßen:

$$P(A_i, B_j) = P(A_i)\, P(B_j) = \frac{1}{I}\, P(B_j)\,.$$

Wir schätzen $P(B_j)$ durch die relative Häufigkeit der $n_{.j}/n$ der j-ten Kategorie von B.

Ersetzen wir $P(B_j)$ in

$$P(A_i, B_j) = \frac{1}{I} P(B_j)$$

durch $n_{\cdot j}/n$, so erhalten wir folgende geschätzte Zellwahrscheinlichkeiten:

$$\hat{P}(A_i, B_j) = \frac{n_{\cdot j}}{n\, I}\,.$$

Wir erhalten somit folgende Schätzer:

$$\hat{n}_{ij} = \frac{n_{\cdot j}}{I}\,.$$

hmcounterend. (fortgesetzt)

Example 41. Tabelle 10.9 enthält die geschätzten Häufigkeiten.

Table 10.9. Geschätzte absolute Häufigkeiten im Modell B

	CDU	SPD
BWL	28	22
VWL	28	22

Der Wert der Likelihood-Quotienten-Teststatistik $G(B)$ ist

$$G(B) = 2\left[50 \ln \frac{50}{28} + 36 \ln \frac{36}{22} + 6 \ln \frac{6}{28} + 8 \ln \frac{8}{22}\right] = 58.77\,.$$

□

Im Modell B ist $G(B)$ approximativ χ^2-verteilt mit $(I-1)J$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 41. Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{2;0.95} = 5.99$. Wir lehnen das Modell B zum Niveau 0.05 ab. □

Schauen wir uns den IPF-Algorithmus an. Es muss gelten

$$\hat{n}_{\cdot j} = n_{\cdot j}\,.$$

Wir setzen

$$\hat{n}_{ij}^{(0)} = 1 \qquad \text{für } i = 1, \ldots, I, j = 1, \ldots, J$$

und passen die Randverteilung von B an. Es gilt

$$\hat{n}_{ij}^{(1)} = \frac{n_{\cdot j}}{\hat{n}_{\cdot j}^{(0)}}\, \hat{n}_{ij}^{(0)} = \frac{n_{\cdot j}}{I}$$

wegen

$$\hat{n}_{\cdot j}^{(0)} = \sum_{i=1}^{I} \hat{n}_{ij}^{(0)} = \sum_{i=1}^{I} 1 = I.$$

10.2.5 Modell A, B

Bevor wir das nächste Modell betrachten, wollen wir uns kurz überlegen, wodurch sich die bisher betrachteten Modelle unterscheiden. Modell 0 fordert Gleichverteilung von A, Gleichverteilung von B und Unabhängigkeit zwischen A und B. Modell A verzichtet im Vergleich zu Modell 0 auf die Gleichverteilung von A, während Modell B im Vergleich zu Modell 0 auf die Gleichverteilung von B verzichtet. Es liegt nun nahe, auch die Gleichverteilung des jeweils anderen Merkmals fallenzulassen. Es wird also nur die Unabhängigkeit zwischen A und B gefordert. Es muss also gelten

$$P(A_i, B_j) = P(A_i)P(B_j) \tag{10.20}$$

für $i = 1, \ldots, I$, $j = 1, \ldots, J$.

Wir schätzen die Wahrscheinlichkeit $P(A_i, B_j)$, indem wir $P(A_i)$ durch $n_{i.}/n$ und $P(B_j)$ durch $n_{.j}/n$ schätzen und dann in (10.20) einsetzen. Wir erhalten also als Schätzer für die erwarteten Häufigkeiten

$$\hat{n}_{ij} = n \frac{n_{i.}}{n} \frac{n_{.j}}{n} = \frac{n_{i.}\, n_{.j}}{n}\,.$$

hmcounterend. (fortgesetzt)

Example 41. Tabelle 10.10 enthält die geschätzten erwarteten Häufigkeiten. Wir sehen, dass die beobachteten Häufigkeiten n_{ij} sehr gut mit den geschätzten Häufigkeiten $\hat{n}_{ij}$ übereinstimmen. Der Wert der Likelihood-Quotienten-Teststatistik ist

$$G(A, B) = 2 \left[50 \ln \frac{50}{48.16} + 36 \ln \frac{36}{37.84} + 6 \ln \frac{6}{7.84} + 8 \ln \frac{8}{6.16} \right] = 1.13\,.$$

Table 10.10. Geschätzte erwartete Häufigkeiten im Modell A, B

	CDU	SPD
BWL	48.16	37.84
VWL	7.84	6.16

□

Im Modell A, B ist $G(A, B)$ approximativ χ^2-verteilt mit $(I-1)(J-1)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 41. Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{1;0.95} = 3.84$. Wir lehnen das Modell A, B zum Niveau 0.05 also nicht ab. □

Warum haben wir das Modell eigentlich mit A, B bezeichnet? Die Notation deutet darauf hin, dass sowohl die Randverteilung von A als auch die Randverteilung von B angepasst wird.

Es muss also gelten

$$\hat{n}_{i.} = n_{i.} \tag{10.21}$$

und

$$\hat{n}_{.j} = n_{.j}\,. \tag{10.22}$$

Für die Schätzer der absoluten Häufigkeiten gilt

$$\hat{n}_{ij} = \frac{n_{.j}\, n_{i.}}{n}\,.$$

Hieraus folgt

$$\hat{n}_{i.} = \sum_{j=1}^{J} \frac{n_{i.}\, n_{.j}}{n} = \frac{n_{i.}}{n} \sum_{j=1}^{J} n_{.j} = \frac{n_{i.}}{n}\, n = n_{i.}$$

und

$$\hat{n}_{.j} = \sum_{i=1}^{I} \frac{n_{i.}\, n_{.j}}{n} = \frac{n_{.j}}{n} \sum_{i=1}^{I} n_{i.} = \frac{n_{.j}}{n}\, n = n_{.j}\,.$$

Wir sehen, dass die Bezeichnung des Modells gerechtfertigt ist. Wir können aber auch diese Bedingungen als Ausgangspunkt nehmen und den IPF-Algorithmus anwenden. Die Gleichungen (10.21) und (10.22) müssen erfüllt sein. Wir passen zuerst die Randverteilung von A an und erhalten

$$\hat{n}_{ij}^{(1)} = \frac{n_{i.}}{J}\,.$$

Nun müssen wir noch die Randverteilung von B anpassen. Wir erhalten

$$\hat{n}_{ij}^{(2)} = \frac{n_{.j}}{\hat{n}_{.j}^{(1)}}\, \hat{n}_{ij}^{(1)} = \frac{n_{.j}}{\frac{n}{J}}\, \frac{n_{i.}}{J} = \frac{n_{i.} n_{.j}}{n},$$

da gilt

$$\hat{n}_{.j}^{(1)} = \sum_{i=1}^{I} \hat{n}_{ij}^{(1)} = \sum_{i=1}^{I} \frac{n_{i.}}{J} = \frac{n}{J}\,.$$

Wir sehen also, dass beim Unabhängigkeitsmodell A, B die Randverteilung von A und die Randverteilung von B angepasst wird. Wir sehen aber auch, dass wir durch die Anpassung der Randverteilung von B nicht die Anpassung der Randverteilung von A zerstört haben. Wäre durch die Anpassung der Randverteilung von B die Anpassung der Randverteilung von A zerstört worden, so hätten wir wieder die Randverteilung von A anpassen müssen.

10.2.6 Modell AB

Das Unabhängigkeitsmodell war das bisher schwächste Modell. Ein noch schwächeres Modell ist das Modell AB. Bei diesem wird die gemeinsame Verteilung von A und B angepasst. Dies liefert die perfekte Anpassung an die Daten.

Die geschätzten absoluten Häufigkeiten sind

$$\hat{n}_{ij} = n_{ij} \, .$$

hmcounterend. (fortgesetzt)

Example 41. Tabelle 10.11 enthält die geschätzten Häufigkeiten.

Table 10.11. Geschätzte absolute Häufigkeiten im Modell A, B

	CDU	SPD
BWL	50	36
VWL	6	8

Der Wert der Likelihood-Quotienten-Teststatistik ist

$$G(AB) = 0.$$

Die Anzahl der Freiheitsgrade ist 0. Die Angemessenheit des Modells müssen wir nicht testen. □

10.2.7 Modellselektion

Wir haben in den letzten Abschnitten eine Reihe von Modellen zur Beschreibung einer zweidimensionalen Kontingenztabelle kennengelernt. Wir wollen nun das Modell wählen, bei dem die geschätzten erwarteten Häufigkeiten mit den beobachteten Häufigkeiten übereinstimmen. Wäre dies das einzige Kriterium der Modellwahl, so würden wir immer das Modell AB wählen. Das Modell sollte aber nicht nur gut angepasst sein, es sollte auch einfach zu interpretieren sein. Bei der Beschreibung der Modelle haben wir mit dem Modell 0 begonnen. Bei diesem werden die meisten Annahmen getroffen. Es ist einfach zu interpretieren, denn jede Merkmalskombination besitzt in diesem Modell die gleiche Wahrscheinlichkeit. Beim Übergang zu den Modellen A und B haben wir jeweils eine der Annahmen des Modells 0 fallengelassen. Dies hat zur Konsequenz, dass diese Modelle nicht mehr so einfach zu interpretieren sind. Außerdem sind die Annahmen der Modelle A beziehungsweise B erfüllt, wenn die Annahmen des Modells 0 erfüllt sind. In diesem Sinne bilden die betrachteten Modelle eine Hierarchie, bei der das Modell 0 auf der höchsten Stufe steht und die Modelle A und B gemeinsam die nächste Stufe bilden.

Läßt man bei den Modellen A beziehungsweise B jeweils eine der Annahmen fallen, so landet man beim Modell A, B. Im letzten Schritt gelangt man zum Modell AB. Abbildung 10.1 zeigt die Hierarchie der Modelle.

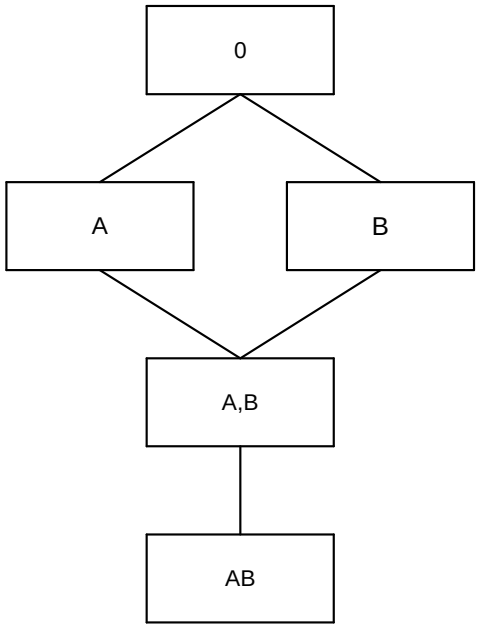


Fig. 10.1. Modellhierarchie eines loglinearen Modells mit zwei Merkmalen

Es gibt eine Reihe von Suchverfahren in loglinearen Modellen. Wir schauen uns hier ein Verfahren von Goodman (1971) an. Dieses beruht auf einer wichtigen Eigenschaft der Likelihood-Quotienten-Teststatistik. Von zwei Modellen, die auf unterschiedlichen Hierarchiestufen stehen, bezeichnen wir das Modell, das auf der höheren Hierarchiestufe steht, als stärkeres Modell S und das Modell, das auf der niedrigeren Hierarchiestufe steht, als schwächeres Modell W. $G(S)$ ist der Wert der Likelihood-Quotienten-Teststatistik des stärkeren Modells und $G(W)$ ist der Wert der Likelihood-Quotienten-Teststatistik des schwächeren Modells. Ist das stärkere Modell das wahre Modell, so ist die Differenz $G(S) - G(W)$ approximativ χ^2-verteilt mit der Differenz aus der Anzahl der Freiheitsgrade des Modells S und der Anzahl der Frei-

heitsgrade des Modells W. Der Beweis ist bei Andersen (1991), S. 147-148 zu finden. Testen wir also

$$H_0: \text{ Modell S trifft zu}$$

gegen

$$H_1: \text{ Modell W trifft zu,}$$

so lehnen wir H_0 ab, wenn gilt

$$G(S) - G(W) > \chi^2_{df;1-\alpha},$$

wobei $\chi^2_{df;1-\alpha}$ das $1-\alpha$-Quantil der χ^2-Verteilung ist, wobei df die Differenz der Freiheitsgrade des Modells S und der Freiheitsgrade des Modells W ist.

Goodman (1971) schlägt vor, mit dem stärksten Modell zu beginnen. In unserem Fall ist dies das Modell 0. Lehnen wir das Modell 0 nicht ab, so beenden wir die Suche und wählen Modell 0 zur Beschreibung der Abhängigkeitsstruktur zwischen den Merkmalen.

hmcounterend. (fortgesetzt)

Example 41. Tabelle 10.12 zeigt die Werte von $G(M)$ und die Freiheitsgrade der einzelnen Modelle.

Table 10.12. Werte von $G(M)$ und Freiheitsgrade df loglinearer Modelle

Modell M	$G(M)$	df
0	60.21	3
A	2.58	2
B	58.77	2
A, B	1.13	1
AB	0	0

Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{3,0.95} = 7.82$. Wir lehnen das Modell 0 zum Niveau 0.05 also ab. □

Lehnen wir das Modell 0 jedoch ab, suchen wir nach einem besseren Modell. Dabei betrachten wir alle Modelle, die in der Hierarchie auf der Stufe unterhalb des Modells 0 stehen. Im Beispiel sind das die Modelle A und B. Wir fragen uns, ob die Anpassung bedeutend verbessert wird, wenn wir eines dieser Modelle betrachten. Hierbei benutzen wir die oben beschriebene Eigenschaft der Likelihood-Quotienten-Teststatistik. Wir bestimmen $G(0) - G(A)$ und $G(0) - G(B)$ und wählen unter den signifikanten Übergängen den mit der größten Verbesserung. Dann testen wir, ob das so gefundene Modell abgelehnt wird. Wird es nicht abgelehnt, so wird es zur Beschreibung des Zusammenhangs gewählt. Wird es abgelehnt, gehen wir zur nächsten Stufe in der Hierarchie. Der Prozess wird so lange fortgesetzt, bis ein geeignetes Modell gefunden wurde. hmcounterend. (fortgesetzt)

Example 41. Beim Übergang vom Modell 0 zum Modell A gilt

$$G(0) - G(A) = 60.21 - 2.58 = 57.63\,.$$

Die Differenz der Freiheitsgrade ist 1. Der Übergang ist signifikant, da $\chi^2_{1,0.95} = 3.84$ gilt. Beim Übergang vom Modell 0 zum Modell B gilt:

$$G(0) - G(B) = 60.21 - 58.77 = 1.44.$$

Die Differenz der Freiheitsgrade ist 1. Der Übergang ist nicht signifikant, da $\chi^2_{1,0.95} = 3.84$ gilt.
Wir gehen also vom Modell 0 zum Modell A. Im Modell A gilt $G(A) = 2.58$. Da im Modell A die Anzahl der Freiheitsgrade gleich 2 ist, lehnen wir wegen $\chi^2_{2,0.95} = 5.99$ das Modell A nicht ab. Wir haben mit dem Modell A ein Modell zur Beschreibung der Abhängigkeitsstruktur gefunden. Die Merkmale `Studienfach` und `Wahlverhalten` sind unabhängig. Außerdem ist das Merkmal `Wahlverhalten` gleichverteilt. □

10.3 Dreidimensionale Kontingenztabellen

Wir haben im Kapitel 10.1 gesehen, dass es zwischen drei qualitativen Merkmalen eine Reihe von Abhängigkeitsstrukturen geben kann. Wir wollen diese mit Hilfe von loglinearen Modellen strukturieren. Dabei gehen wir davon aus, dass eine Zufallsstichprobe vom Umfang n aus einer Grundgesamtheit vorliegt, in der die Merkmale A, B und C von Interesse sind. Das Merkmal A besitze die Ausprägungen $A_1, \ldots, A_I$, das Merkmal B die Ausprägungen $B_1, \ldots, B_J$ und das Merkmal C die Ausprägungen $C_1, \ldots, C_K$. Die Wahrscheinlichkeit, aus der Grundgesamtheit ein Objekt mit den Merkmalsausprägungen A_i, B_j und C_k zufällig auszuwählen, bezeichnen wir mit $P(A_i, B_j, C_k)$. Die Anzahl der Objekte mit den Merkmalsausprägungen A_i, B_j und C_k in der Stichprobe bezeichnen wir mit n_{ijk}. Wir stellen die Häufigkeiten in einer Kontingenztabelle zusammen.

Example 42. Wir greifen das Beispiel 40 auf. Dabei bezeichnen wir das Merkmal `Titanic` mit A, das Merkmal `Satz` mit B und das Merkmal `Geschlecht` mit C. Die Daten sind in Tabelle 10.2 auf Seite 285 zu finden. □

Damit unsere Ausführungen nicht ausufern, betrachten wir nur Modelle, bei denen in der Definition des Modells alle drei Merkmale auftauchen. Wir werden also nicht eingehen auf Modelle wie das Modell 0 oder das Modell AC. Eine Beschreibung dieser Modelle ist bei Fahrmeir et al. (1996) zu finden. Wir starten mit dem Modell der totalen Unabhängigkeit.

10.3.1 Das Modell der totalen Unabhängigkeit

Im Modell A, B, C der totalen Unabhängigkeit unterstellen wir, dass alle Merkmale unabhängig sind. Es gilt somit

$$P(A_i, B_j, C_k) = P(A_i)P(B_j)P(C_k)$$

für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$. Aus dem Modell der totalen Unabhängigkeit folgt die paarweise Unabhängigkeit:

1. Die Merkmale A und B sind unabhängig:
   $P(A_i, B_j) = P(A_i)\,P(B_j) \quad$ für $i = 1, \ldots, I, j = 1, \ldots, J$.
2. Die Merkmale A und C sind unabhängig:
   $P(A_i, C_k) = P(A_i)\,P(C_k) \quad$ für $i = 1, \ldots, I, k = 1, \ldots, K$.
3. Die Merkmale B und C sind unabhängig:
   $P(B_j, C_k) = P(B_j)\,P(C_k) \quad$ für $j = 1, \ldots, J, k = 1, \ldots, K$.

Wir zeigen die erste Behauptung. Die anderen ergeben sich analog.

$$P(A_i, B_j) = \sum_{k=1}^{K} P(A_i, B_j, C_k) = \sum_{k=1}^{K} P(A_i)P(B_j)P(C_k)$$

$$= P(A_i)P(B_j) \sum_{k=1}^{K} P(C_k) = P(A_i)P(B_j).$$

Die erwartete Häufigkeit des gemeinsamen Auftretens von A_i, B_j und C_k im Modell A, B, C ist

$$n\,P(A_i, B_j, C_k) = n\,P(A_i)P(B_j)P(C_k).$$

Wir schätzen $P(A_i)$ durch $n_{i..}/n$, $P(B_j)$ durch $n_{.j.}/n$ und $P(C_k)$ durch $n_{..k}/n$ mit

$$n_{i..} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk},$$

$$n_{.j.} = \sum_{i=1}^{I} \sum_{k=1}^{K} n_{ijk},$$

$$n_{..k} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ijk}$$

für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$. Wir erhalten als geschätzte erwartete Häufigkeiten

$$\hat{n}_{ijk} = n\,\frac{n_{i..}}{n}\,\frac{n_{.j.}}{n}\,\frac{n_{..k}}{n}\,.$$

Dies kann man vereinfachen zu

$$\hat{n}_{ijk} = \frac{n_{i..}n_{.j.}n_{..k}}{n^2}\,.$$

hmcounterend. (fortgesetzt)

Example 42. Es gilt

$$n_{1..} = 140, \;\; n_{2..} = 60,$$
$$n_{.1.} = 120, \;\; n_{.2.} = 80,$$
$$n_{..1} = 100, \;\; n_{..2} = 100\,.$$

Also erhalten wir folgende geschätzte erwartete Häufigkeiten:

$$\hat{n}_{111} = \frac{n_{1..}n_{.1.}n_{..1}}{n^2} = \frac{140 \cdot 120 \cdot 100}{200^2} = 42,$$

$$\hat{n}_{121} = \frac{n_{1..}n_{.2.}n_{..1}}{n^2} = \frac{140 \cdot 80 \cdot 100}{200^2} = 28,$$

$$\hat{n}_{211} = \frac{n_{2..}n_{.1.}n_{..1}}{n^2} = \frac{60 \cdot 120 \cdot 100}{200^2} = 18,$$

$$\hat{n}_{221} = \frac{n_{2..}n_{.2.}n_{..1}}{n^2} = \frac{60 \cdot 80 \cdot 100}{200^2} = 12,$$

$$\hat{n}_{112} = \frac{n_{1..}n_{.1.}n_{..2}}{n^2} = \frac{140 \cdot 120 \cdot 100}{200^2} = 42,$$

$$\hat{n}_{122} = \frac{n_{1..}n_{.2.}n_{..2}}{n^2} = \frac{140 \cdot 80 \cdot 100}{200^2} = 28,$$

$$\hat{n}_{212} = \frac{n_{2..}n_{.1.}n_{..2}}{n^2} = \frac{60 \cdot 120 \cdot 100}{200^2} = 18,$$

$$\hat{n}_{222} = \frac{n_{2..}n_{.2.}n_{..2}}{n^2} = \frac{60 \cdot 80 \cdot 100}{200^2} = 12.$$

Tabelle 10.13 zeigt die dreidimensionale Kontingenztabelle mit den geschätzten erwarteten Häufigkeiten.

Table 10.13. Geschätzte erwartete Häufigkeiten des Modells A, B, C

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	42	28
	nein	18	12
m	ja	42	28
	nein	18	12

□

Zur Überprüfung der Güte eines Modells M bestimmen wir die Likelihood-Quotienten-Teststatistik

$$G(M) = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk} \ln \frac{n_{ijk}}{\hat{n}_{ijk}}.$$

Im Modell A, B, C ist $G(A, B, C)$ approximativ χ^2-verteilt mit $IJK - I - J - K + 2$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 42. Es gilt $G(A, B, C) = 39.6621$. Die Anzahl der Freiheitsgrade ist 4. Wir sehen, dass das Modell nicht gut passt. □

Schauen wir uns den IPF-Algorithmus für das Modell A, B, C an. Es muss gelten

$$\begin{aligned} \hat{n}_{i..} &= n_{i..}, \\ \hat{n}_{.j.} &= n_{.j.}, \\ \hat{n}_{..k} &= n_{..k}\,. \end{aligned}$$

Ausgehend von Startwerten $\hat{n}_{ijk}^{(0)} = 1$ passen wir zunächst die Randverteilung von A an. Es gilt

$$\hat{n}_{ijk}^{(1)} = \frac{n_{i..}}{\hat{n}_{i..}^{(0)}} \hat{n}_{ijk}^{(0)} = \frac{n_{i..}}{JK},$$

da gilt

$$\hat{n}_{i..}^{(0)} = \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{n}_{ijk}^{(0)} = \sum_{j=1}^{J} \sum_{k=1}^{K} 1 = JK.$$

Anschließend passen wir die Randverteilung von B an:

$$\hat{n}_{ijk}^{(2)} = \frac{n_{.j.}}{\hat{n}_{.j.}^{(1)}} \hat{n}_{ijk}^{(1)} = \frac{n_{.j.}}{\frac{n}{J}} \frac{n_{i..}}{JK} = \frac{n_{i..} n_{.j.}}{nK}.$$

Dabei haben wir benutzt, dass gilt:

$$\hat{n}_{.j.}^{(1)} = \sum_{i=1}^{I} \sum_{k=1}^{K} \frac{n_{i..}}{JK} = \sum_{i=1}^{I} n_{i..} \sum_{k=1}^{K} \frac{1}{JK} = nK \frac{1}{JK} = \frac{n}{J}\,.$$

Nun passen wir noch die Randverteilung von C an:

$$\hat{n}_{ijk}^{(3)} = \frac{n_{..k}}{\hat{n}_{..k}^{(2)}} \hat{n}_{ijk}^{(2)} = \frac{n_{..k}}{\frac{n}{K}} \frac{n_{i..} n_{.j.}}{nK} = \frac{n_{i..} n_{.j.} n_{..k}}{n^2}.$$

Dies gilt wegen

$$\hat{n}_{..k}^{(2)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{i..} n_{.j.}}{nK} = \frac{1}{nK} \sum_{i=1}^{I} n_{i..} \sum_{j=1}^{J} n_{.j.} = \frac{1}{nK} n\, n = \frac{n}{K}\,.$$

Wie man leicht erkennt, wurde auch hier durch die Anpassung der Randverteilung von C nicht die Anpassung der Randverteilungen von A und B zerstört.

10.3.2 Das Modell der Unabhängigkeit einer Variablen

Im Modell A, B, C sind alle drei Merkmale vollständig voneinander unabhängig. Hieraus folgt die Unabhängigkeit aller Paare. Gegenüber dem Modell A, B, C lassen wir im Modell AB, C die Annahme der Unabhängigkeit von A und B fallen. Im Modell AB, C gilt

$$P(A_i, B_j, C_k) = P(A_i, B_j)P(C_k) \tag{10.23}$$

für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$. Aus (10.23) folgt:

1. Die Merkmale A und C sind unabhängig:
   $P(A_i, C_k) = P(A_i)\,P(C_k) \quad$ für $i = 1, \ldots, I, k = 1, \ldots, K$.
2. Die Merkmale B und C sind unabhängig:
   $P(B_j, C_k) = P(B_j)\,P(C_k) \quad$ für $j = 1, \ldots, J, k = 1, \ldots, K$.

Wir zeigen die erste Behauptung. Die andere folgt analog.

$$P(A_i, C_k) = \sum_{j=1}^{J} P(A_i, B_j, C_k) = \sum_{j=1}^{J} P(A_i, B_j)P(C_k)$$

$$= P(C_k) \sum_{j=1}^{J} P(A_i, B_j) = P(A_i)P(C_k).$$

Die erwartete Häufigkeit des gemeinsamen Auftretens von A_i, B_j und C_k im Modell AB, C ist

$$n\,P(A_i, B_j, C_k) = n\,P(A_i, B_j)P(C_k).$$

Wir schätzen $P(A_i, B_j)$ durch $n_{ij.}/n$ und $P(C_k)$ durch $n_{..k}/n$ und erhalten folgende Schätzer der erwarteten Häufigkeiten:

$$\hat{n}_{ijk} = n\,\frac{n_{ij.}}{n}\,\frac{n_{..k}}{n}\,. \tag{10.24}$$

Dabei gilt für $i = 1, \ldots, I$, $j = 1, \ldots, J$:

$$n_{ij.} = \sum_{k=1}^{K} n_{ijk}\,.$$

(10.24) kann man vereinfachen zu

$$\hat{n}_{ijk} = \frac{n_{ij.}n_{..k}}{n}\,. \tag{10.25}$$

hmcounterend. (fortgesetzt)

Example 42. Es gilt

$$n_{11.} = 92, \;\; n_{12.} = 48,$$
$$n_{21.} = 28, \;\; n_{22.} = 32,$$
$$n_{..1} = 100, \;\; n_{..2} = 100\,.$$

Also erhalten wir folgende geschätzte erwartete Häufigkeiten:

$$\hat{n}_{111} = \frac{n_{11.}n_{..1}}{n} = \frac{92 \cdot 100}{200} = 46,$$
$$\hat{n}_{121} = \frac{n_{12.}n_{..1}}{n} = \frac{48 \cdot 100}{200} = 24,$$
$$\hat{n}_{211} = \frac{n_{21.}n_{..1}}{n} = \frac{28 \cdot 100}{200} = 14,$$
$$\hat{n}_{221} = \frac{n_{22.}n_{..1}}{n} = \frac{32 \cdot 100}{200} = 16,$$
$$\hat{n}_{112} = \frac{n_{11.}n_{..2}}{n} = \frac{92 \cdot 100}{200} = 46,$$
$$\hat{n}_{122} = \frac{n_{12.}n_{..2}}{n} = \frac{48 \cdot 100}{200} = 24,$$
$$\hat{n}_{212} = \frac{n_{21.}n_{..2}}{n} = \frac{28 \cdot 100}{200} = 14,$$
$$\hat{n}_{222} = \frac{n_{22.}n_{..2}}{n} = \frac{32 \cdot 100}{200} = 16\,.$$

Tabelle 10.14 zeigt die dreidimensionale Kontingenztabelle mit den geschätzten Häufigkeiten.

Table 10.14. Geschätzte erwartete Häufigkeiten des Modells AB, C

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	46	24
	nein	14	16
m	ja	46	24
	nein	14	16

□

Im Modell AB, C ist $G(AB, C)$ approximativ χ^2-verteilt mit $(K-1)(IJ-1)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 42. Es gilt $G(AB, C) = 33.3837$. Die Anzahl der Freiheitsgrade ist 3. Wir sehen, dass auch dieses Modell nicht gut passt. □

Schauen wir uns den IPF-Algorithmus für das Modell AB, C an. Es muss gelten

$$\hat{n}_{ij.} = n_{ij.},$$
$$\hat{n}_{..k} = n_{..k}\,.$$

Ausgehend von den Startwerten $\hat{n}_{ijk}^{(0)} = 1$ passen wir zunächst die Randverteilung von AB an. Es gilt

$$\hat{n}_{ijk}^{(1)} = \frac{n_{ij.}}{\hat{n}_{ij.}^{(0)}}\,\hat{n}_{ijk}^{(0)} = \frac{n_{ij.}}{K}\,,$$

da gilt

$$\hat{n}_{ij.}^{(0)} = \sum_{k=1}^{K} \hat{n}_{ijk}^{(0)} = \sum_{k=1}^{K} 1 = K\,.$$

Dann passen wir die Randverteilung von C an. Es gilt

$$\hat{n}_{ijk}^{(2)} = \frac{n_{..k}}{\hat{n}_{..k}^{(1)}}\,\hat{n}_{ijk}^{(1)} = \frac{n_{..k}}{\frac{n}{K}}\,\frac{n_{ij.}}{K} = \frac{n_{ij.}n_{..k}}{n}\,.$$

Dabei haben wir benutzt:

$$\hat{n}_{..k}^{(1)} = \sum_{i=1}^{I}\sum_{j=1}^{J} \hat{n}_{ijk}^{(1)} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij.}}{K} = \frac{n}{K}\,.$$

Beim Modell AB, C fordern wir gegenüber dem Modell A, B, C nicht mehr, dass A und B unabhängig sind. Wir können aber ausgehend vom Modell A, B, C auch auf die Forderung nach der Unabhängigkeit zwischen A und C beziehungsweise zwischen B und C verzichten. Dann erhalten wir die Modelle AC, B beziehungsweise BC, A. Schauen wir uns diese kurz an. Im Modell AC, B bestimmen wir die geschätzten Häufigkeiten durch

$$\hat{n}_{ijk} = \frac{n_{i.k}n_{.j.}}{n} \tag{10.26}$$

und im Modell BC, A durch

$$\hat{n}_{ijk} = \frac{n_{.jk}n_{i..}}{n}\,. \tag{10.27}$$

Dabei ist für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$:

$$n_{i.k} = \sum_{j=1}^{J} n_{ijk} \,,$$

$$n_{.jk} = \sum_{i=1}^{I} n_{ijk} \,.$$

hmcounterend. (fortgesetzt)

Example 42. Tabelle 10.15 zeigt die dreidimensionale Tabelle mit den geschätzten Häufigkeiten des Modells AC, B und Tabelle 10.16 die dreidimensionale Tabelle mit den geschätzten Häufigkeiten des Modells BC, A.

Table 10.15. Geschätzte erwartete Häufigkeiten des Modells AC, B

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	48	32
	nein	12	8
m	ja	36	24
	nein	24	16

Table 10.16. Geschätzte erwartete Häufigkeiten des Modells BC, A

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	54.6	15.4
	nein	23.4	6.6
m	ja	29.4	40.6
	nein	12.6	17.4

□

Im Modell AC, B ist $G(AC, B)$ approximativ χ^2-verteilt mit $(J-1)(IK-1)$ Freiheitsgraden und im Modell BC, A ist $G(BC, A)$ approximativ χ^2-verteilt mit $(I-1)(JK-1)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 42. Es gilt $G(AC, B) = 29.9992$ und $G(BC, A) = 11.8974$. □

10.3.3 Das Modell der bedingten Unabhängigkeit

Wir haben im Beispiel 39 auf Seite 284 das Modell der bedingten Unabhängigkeit kennengelernt. Die Merkmale A und B sind für jede Ausprägung des Merkmals C unabhängig. Es gilt also für $i = 1, \ldots, I$, $j = 1, \ldots, J$ und $k = 1, \ldots, K$:

$$P(A_i, B_j|C_k) = P(A_i|C_k)\, P(B_j|C_k). \tag{10.28}$$

Aus dieser Gleichung folgt

$$P(A_i, B_j, C_k) = \frac{P(A_i, C_k)P(B_j, C_k)}{P(C_k)}. \tag{10.29}$$

Dies sieht man folgendermaßen:

$$\begin{aligned} P(A_i, B_j, C_k) &= P(A_i, B_j|C_k)\, P(C_k) = P(A_i|C_k)\, P(B_j|C_k)\, P(C_k) \\ &= \frac{P(A_i, C_k)}{P(C_k)} \frac{P(B_j, C_k)}{P(C_k)} P(C_k) = \frac{P(A_i, C_k)P(B_j, C_k)}{P(C_k)}\,. \end{aligned}$$

Wir bezeichnen das Modell mit AC, BC.

Die erwartete Häufigkeit des gemeinsamen Auftretens von A_i, B_j und C_k im Modell AC, BC ist

$$n\, P(A_i, B_j, C_k) = n\, \frac{P(A_i, C_k)P(B_j, C_k)}{P(C_k)}.$$

Wir schätzen $P(A_i, C_k)$ durch $n_{i.k}/n$, $P(B_j, C_k)$ durch $n_{.jk}/n$ und $P(C_k)$ durch $n_{..k}/n$ und erhalten folgende Schätzungen:

$$\hat{n}_{ijk} = n\, \frac{\frac{n_{i.k}}{n} \frac{n_{.jk}}{n}}{\frac{n_{..k}}{n}}.$$

Dies kann man vereinfachen zu:

$$\hat{n}_{ijk} = \frac{n_{i.k}\, n_{.jk}}{n_{..k}}. \tag{10.30}$$

hmcounterend. (fortgesetzt)

Example 42. Es gilt

$$\begin{aligned} n_{1.1} &= 80, \;\; n_{1.2} = 60, \\ n_{2.1} &= 20, \;\; n_{2.2} = 40, \\ n_{.11} &= 78, \;\; n_{.12} = 42, \\ n_{.21} &= 22, \;\; n_{.22} = 58, \\ n_{..1} &= 100, \;\; n_{..2} = 100\,. \end{aligned}$$

Also erhalten wir folgende geschätzte erwartete Häufigkeiten:

$$\hat{n}_{111} = \frac{n_{1.1}n_{.11}}{n_{..1}} = \frac{80 \cdot 78}{100} = 62.4,$$

$$\hat{n}_{121} = \frac{n_{1.1}n_{.21}}{n_{..1}} = \frac{80 \cdot 22}{100} = 17.6,$$

$$\hat{n}_{211} = \frac{n_{2.1}n_{.11}}{n_{..1}} = \frac{20 \cdot 78}{100} = 15.6,$$

$$\hat{n}_{221} = \frac{n_{2.1}n_{.21}}{n_{..1}} = \frac{20 \cdot 22}{100} = 4.4,$$

$$\hat{n}_{112} = \frac{n_{1.2}n_{.12}}{n_{..2}} = \frac{60 \cdot 42}{100} = 25.2,$$

$$\hat{n}_{122} = \frac{n_{1.2}n_{.22}}{n_{..2}} = \frac{60 \cdot 58}{100} = 34.8,$$

$$\hat{n}_{212} = \frac{n_{2.2}n_{.12}}{n_{..2}} = \frac{40 \cdot 42}{100} = 16.8,$$

$$\hat{n}_{222} = \frac{n_{2.2}n_{.22}}{n_{..2}} = \frac{40 \cdot 58}{100} = 23.2\,.$$

Tabelle 10.17 zeigt die dreidimensionale Tabelle mit den geschätzten Häufigkeiten.

Table 10.17. Geschätzte erwartete Häufigkeiten des Modells AC, BC

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	62.4	17.6
	nein	15.6	4.4
m	ja	25.2	34.8
	nein	16.8	23.2

□

Im Modell AC, BC ist $G(AC, BC)$ approximativ χ^2-verteilt mit $K(I-1)(J-1)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 42. Es gilt $G(AC, BC) = 2.2345$. Die Anzahl der Freiheitsgrade ist 2. Wir sehen, dass dieses Modell sehr gut passt. □

Schauen wir uns den IPF-Algorithmus für das Modell AC, BC an. Es muss gelten

$$\hat{n}_{i.k} = n_{i.k},$$
$$\hat{n}_{.jk} = n_{.jk}\,.$$

Ausgehend von den Startwerten $\hat{n}_{ijk}^{(0)} = 1$ passen wir zunächst die Randverteilung von AC an:

$$\hat{n}_{ijk}^{(1)} = \frac{n_{i.k}}{\hat{n}_{i.k}^{(0)}}\,\hat{n}_{ijk}^{(0)} = \frac{n_{i.k}}{J},$$

da gilt

$$\hat{n}_{i.k}^{(0)} = \sum_{j=1}^{J} \hat{n}_{ijk}^{(0)} = \sum_{j=1}^{J} 1 = J.$$

Dann passen wir die Randverteilung von BC an:

$$\hat{n}_{ijk}^{(2)} = \frac{n_{.jk}}{\hat{n}_{.jk}^{(1)}}\,\hat{n}_{ijk}^{(1)} = \frac{n_{.jk}}{\frac{n_{..k}}{J}}\,\frac{n_{i.k}}{J} = \frac{n_{.jk}\,n_{i.k}}{n_{..k}},$$

da gilt

$$\hat{n}_{.jk}^{(1)} = \sum_{i=1}^{I} \hat{n}_{ijk}^{(1)} = \sum_{i=1}^{I} \frac{n_{i.k}}{J} = \frac{n_{..k}}{J}.$$

Wie beim Modell der Unabhängigkeit einer Variablen gibt es auch beim Modell der bedingten Unabhängigkeit drei Fälle. Wir betrachten hier kurz die Modelle AB, AC und AB, BC. Im Modell AB, AC bestimmen wir die geschätzten Häufigkeiten durch

$$\hat{n}_{ijk} = \frac{n_{ij.}\,n_{i.k}}{n_{i..}} \tag{10.31}$$

und im Modell AB, BC durch

$$\hat{n}_{ijk} = \frac{n_{ij.}\,n_{.jk}}{n_{.j.}}. \tag{10.32}$$

hmcounterend. (fortgesetzt)

Example 42. Tabelle 10.18 zeigt die dreidimensionale Tabelle mit den geschätzten Häufigkeiten des Modells AB, AC und Tabelle 10.19 die dreidimensionale Tabelle mit den geschätzten Häufigkeiten des Modells AB, BC.

□

Im Modell AB, AC ist $G(AB, AC)$ approximativ χ^2-verteilt mit $I(J-1)(K-1)$ Freiheitsgraden und im Modell AB, BC ist $G(AB, BC)$ approximativ χ^2-verteilt mit $J(I-1)(K-1)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 42. Es gilt $G(AB, AC) = 23.7208$ und $G(AB, BC) = 5.619$. Die Anpassung des Modells AB, AC ist schlecht, während das Modell AB, BC gut geeignet ist. □

Table 10.18. Geschätzte erwartete Häufigkeiten des Modells AB, AC

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	52.57	27.43
	nein	9.33	10.67
m	ja	39.43	20.57
	nein	18.67	21.33

Table 10.19. Geschätzte erwartete Häufigkeiten des Modells AB, BC

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	59.8	13.2
	nein	18.2	8.8
m	ja	32.2	34.8
	nein	9.8	23.2

10.3.4 Das Modell ohne Drei-Faktor-Interaktion

Im Modell AC, BC der bedingten Unabhängigkeit sind die Merkmale A und B unabhängig, wenn man die einzelnen Ausprägungen des Merkmals C betrachtet. Wir können dieses Modell auch über das Kreuzproduktverhältnis charakterisieren. Das Kreuzproduktverhältnis von A und B muss gleich 1 sein, falls C die Merkmalsausprägung C_k, $k = 1, \ldots, K$ aufweist. Für $k = 1, \ldots, K$ muss also gelten

$$\frac{P(A_1, B_1, C_k)P(A_2, B_2, C_k)}{P(A_1, B_2, C_k)P(A_2, B_1, C_k)} = 1.$$

Wir betrachten nun den Fall $K = 2$. Für diesen gilt im Modell AC, BC:

$$\frac{P(A_1, B_1, C_1)P(A_2, B_2, C_1)}{P(A_1, B_2, C_1)P(A_2, B_1, C_1)} = \frac{P(A_1, B_1, C_2)P(A_2, B_2, C_2)}{P(A_1, B_2, C_2)P(A_2, B_1, C_2)} = 1.$$

Fordert man, dass das Kreuzproduktverhältnis von A und B für die einzelnen Merkmalsausprägungen von C gleich ist, so erhält man das Modell AB, AC, BC. Es muss also gelten

$$\frac{P(A_1, B_1, C_1)P(A_2, B_2, C_1)}{P(A_1, B_2, C_1)P(A_2, B_1, C_1)} = \frac{P(A_1, B_1, C_2)P(A_2, B_2, C_2)}{P(A_1, B_2, C_2)P(A_2, B_1, C_2)}. \quad (10.33)$$

Fahrmeir et al. (1996), S.554 zeigen, wie man diese Forderung auf Zusammenhänge zwischen drei Merkmalen mit mehr als zwei Merkmalsausprägungen übertragen kann. Die Gleichung (10.33) beinhaltet aber nicht nur,

dass das Kreuzproduktverhältnis von A und B für die einzelnen Merkmalsausprägungen von C gleich ist. Wir können die Gleichung (10.33) umformen zu

$$\frac{P(A_1, B_1, C_1)P(A_1, B_2, C_2)}{P(A_1, B_1, C_2)P(A_1, B_2, C_1)} = \frac{P(A_2, B_1, C_1)P(A_2, B_2, C_2)}{P(A_2, B_1, C_2)P(A_2, B_2, C_1)}. \quad (10.34)$$

Auf der linken Seite von Gleichung (10.34) ist A_1 und auf der rechten Seite A_2 konstant. Außerdem steht auf der linken Seite das Kreuzproduktverhältnis von B und C für festes A_1 und auf der rechten das Kreuzproduktverhältnis von B und C für festes A_2. Eine analoge Beziehung erhält man für das Kreuzproduktverhältnis von A und C für die einzelnen Merkmalsausprägungen von B. Im Modell AB, AC, BC ist der Zusammenhang zwischen zwei Merkmalen für jede Ausprägung des dritten Merkmals gleich. Man spricht deshalb auch vom Modell ohne Drei-Faktor-Interaktion. In diesem Modell können die $\hat{n}_{ijk}$ nicht explizit angegeben werden. Man muss in diesem Fall den IPF-Algorithmus anwenden. Bei diesem passt man zunächst das Modell AB, dann das Modell AC und dann das Modell BC an. Diesen Zyklus wiederholt man so lange, bis sich die $\hat{n}_{ijk}$ stabilisieren. hmcounterend. (fortgesetzt)

Example 42. In Kapitel 10.3.3 haben wir bereits das Modell AC, BC angepasst. Die geschätzten Häufigkeiten $\hat{n}^{(2)}_{ijk}$ dieses Modells sind in Tabelle 10.17 auf Seite 308 zu finden. Wir passen mit dem IPF-Algorithmus noch AB an. Die geschätzten Häufigkeiten $\hat{n}^{(3)}_{ijk}$ erhalten wir durch

$$\hat{n}^{(3)}_{ijk} = \frac{n_{ij.}}{\hat{n}^{(2)}_{ij.}} \hat{n}^{(2)}_{ijk}.$$

Es gilt

$$n_{11.} = 92, \quad n_{12.} = 48, \quad n_{21.} = 28, \quad n_{22.} = 32$$

und

$$\hat{n}_{11.} = 87.6, \quad \hat{n}_{12.} = 52.4, \quad \hat{n}_{21.} = 32.4, \quad \hat{n}_{22.} = 27.6.$$

Mit den Werten in Tabelle 10.17 folgt

$$\hat{n}^{(3)}_{111} = \frac{92}{87.6} 62.4 = 65.53,$$

$$\hat{n}^{(3)}_{121} = \frac{48}{52.4} 17.6 = 16.12,$$

$$\hat{n}^{(3)}_{211} = \frac{28}{32.4} 15.6 = 13.48,$$

$$\hat{n}^{(3)}_{221} = \frac{32}{27.6} 4.4 = 5.10,$$

$$\hat{n}^{(3)}_{112} = \frac{92}{87.6}\,25.2 = 26.47,$$

$$\hat{n}^{(3)}_{122} = \frac{48}{52.4}\,34.8 = 31.88,$$

$$\hat{n}^{(3)}_{212} = \frac{28}{32.4}\,16.8 = 14.52,$$

$$\hat{n}^{(3)}_{222} = \frac{32}{27.6}\,23.2 = 26.90.$$

Durch die Anpassung von AB haben wir uns aber die Anpassung von AC und BC zerstört. So gilt zum Beispiel

$$n_{1.1} = n_{111} + n_{121} = 80$$

und

$$\hat{n}^{(3)}_{1.1} = \hat{n}^{(3)}_{111} + \hat{n}^{(3)}_{121} = 65.53 + 16.12 = 81.65.$$

Wir müssen also AC wieder anpassen. Tabelle 10.20 zeigt die geschätzten Häufigkeiten, die sich ergeben, wenn man die Anpassung so lange iteriert, bis sich die relativen Häufigkeiten stabilisiert haben. □

Table 10.20. Geschätzte Häufigkeiten des Modells AB, AC, BC

Geschlecht	Titanic	Satz richtig	Satz falsch
w	ja	63.89	16.11
	nein	14.11	5.89
m	ja	28.11	31.89
	nein	13.89	26.11

Im Modell AB, AC, BC ist die Likelihood-Ratio-Statistik $G(AB, AC, BC)$ approximativ χ^2-verteilt mit $(I-1)(J-1)(K-1)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 42. Es gilt $G(AB, AC, BC) = 0.0058$. Die Anpassung des Modells AB, AC, BC ist hervorragend. □

10.3.5 Das saturierte Modell

Wir betrachten wie bei einer zweidimensionalen Kontingenztabelle das Modell, bei dem die Kontingenztabelle perfekt angepasst ist. Wir bezeichnen dies als Modell ABC oder saturiertes Modell. Der Wert der Likelihood-Quotienten-Teststatistik ist beim saturierten Modell gleich 0.

10.3.6 Modellselektion

hmcounterend. (fortgesetzt)

Example 42. Tabelle 10.21 zeigt die Werte von $G(M)$ und die Freiheitsgrade der einzelnen Modelle.

Table 10.21. Werte von $G(M)$ und Freiheitsgrade *df* loglinearer Modelle

Modell M	$G(M)$	*df*
A, B, C	39.6621	4
BC, A	11.8974	3
AC, B	29.9992	3
AB, C	33.3837	3
AB, AC	23.7208	2
AB, BC	5.6190	2
AC, BC	2.2345	2
AB, AC, BC	0.0058	1
ABC	0	0

□

Wir werden wieder das von Goodman (1971) vorgeschlagene Modellselektionsverfahren verwenden. Abbildung 10.2 zeigt die Modellhierarchie des loglinearen Modells mit drei Merkmalen.

hmcounterend. (fortgesetzt)

Example 42. Wir starten mit dem Modell A, B, C. In diesem Modell gilt $G(A, B, C) = 39.6621$. Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{4,0.95} = 9.49$. Wir verwerfen das Modell. Wir suchen unter den Modellen AB, C, AC, B und BC, A das beste. Es gilt

$$G(A, B, C) - G(AB, C) = 39.6621 - 33.3837 = 6.2784,$$

$$G(A, B, C) - G(AC, B) = 39.6621 - 29.9992 = 9.6629$$

und

$$G(A, B, C) - G(BC, A) = 39.6621 - 11.8974 = 27.7647\,.$$

Die größte Verbesserung tritt beim Übergang zum Modell BC, A auf. Es gilt $G(BC, A) = 11.8974$. Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{3,0.95} = 7.81$. Wir verwerfen dieses Modell und gehen weiter. Es gilt

$$G(BC, A) - G(AC, BC) = 11.8974 - 2.2345 = 9.6629$$

und

$$G(BC, A) - G(AB, BC) = 11.8974 - 5.6190 = 6.2784.$$

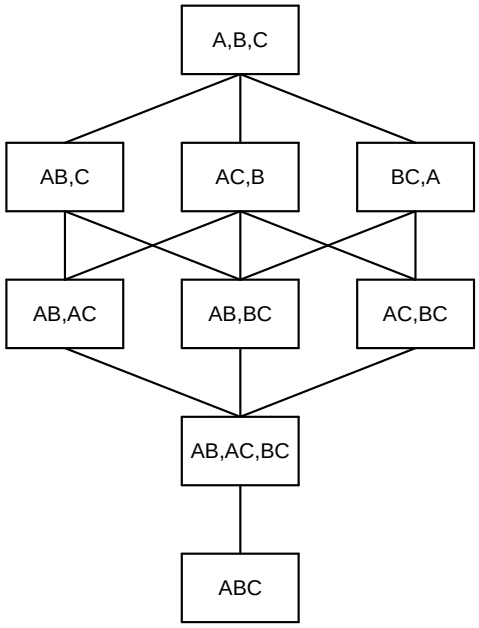


Fig. 10.2. Modellhierarchie eines loglinearen Modells mit drei Merkmalen

Die größte Verbesserung tritt beim Übergang zum Modell AC, BC auf. Es gilt $G(AC, BC) = 2.2345$. Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{2,0.95} = 5.99$. Wir verwerfen dieses Modell nicht. Das Modell AC, BC beschreibt den Zusammenhang zwischen den drei Merkmalen am besten. □

10.4 Loglineare Modelle in S-PLUS

Bevor wir uns anschauen, wie man loglineare Modelle in `S-PLUS` schätzt, wollen wir zeigen, wie man den χ^2-Unabhängigkeitstest durchführt. Hierzu betrachten wir die Daten in Tabelle 10.1 auf Seite 282. Wir geben diese in `S-PLUS` als Matrix `wahl` ein:

```
> wahl<-matrix(c(50,6,36,8),2,2)
> wahl
     [,1] [,2]
```

```
[1,]   50   36
[2,]    6    8
```

Mit der Funktion `chisq.test` kann man in `S-PLUS` einen χ^2-Unabhängigkeitstest durchführen. Wir geben ein

```
> chisq.test(wahl,correct=F)
```

und erhalten folgendes Ergebnis:

```
Pearson's chi-square test without Yates'
continuity correction

data:  wahl
X-square = 1.1412, df = 1, p-value = 0.2854
```

Wir haben das Argument `correct` auf `F` gesetzt, da in der Teststatistik keine Stetigkeitskorrektur berücksichtigt werden soll. Wir sehen, dass der Wert der Teststatistik gleich 1.1412 ist. Die Anzahl der Freiheitsgrade ist gleich 1. Die Überschreitungswahrscheinlichkeit beträgt 0.2854. Wir lehnen zum Niveau $\alpha = 0.05$ die Nullhypothese der Unabhängigkeit also nicht ab.

In `S-PLUS` gibt es eine Funktion `loglin`, mit der man loglineare Modelle an eine Tabelle anpassen kann. Die Funktion `loglin` wird aufgerufen durch

```
loglin(table, margin, start=<<see below>>, fit=F,
       eps=0.1, iter=20, param=F, print=T)
```

Die Kontingenztabelle wird dem Argument `table` übergeben. Mit dem Argument `margin` legt man die Randverteilungen fest, die angepasst werden sollen, wobei `margin` eine Liste ist. Jede Komponente enthält eine Randverteilung als Vektor, wobei die Komponenten des Vektors die Dimensionen der Randverteilung sind. Will man also das Modell AC, BC anpassen, so wählt man für das Argument `margin` die Liste `list(c(1,3),c(2,3))`. Setzt man das Argument `fit` auf `T`, so wird die angepasste Tabelle als Ergebnis zurückgegeben. Durch die Argumente `eps` und `iter` steuert man das Ende des IPF-Algorithmus. Setzt man das Argument `print` auf `F`, so ist die Anzahl der Iterationen am Ende nicht ausgegeben. Mit dem Argument `param` kann man sich die Parameterschätzer des loglinearen Modells ausgeben. Da wir uns mit diesen nicht beschäftigt haben, lassen wir dieses Argument auf dem Wert `F`. Schauen wir uns das Ergebnis der Funktion `loglin` am Beispiel 39 an. Wir erzeugen die dreidimensionale Tabelle mit der Funktion `array` , wie es auf Seite 67 beschrieben wird:

```
> loglinbsp<-array(c(64,14,16,6,28,14,32,26),c(2,2,2))
> dimnames(loglinbsp)<-list(c("Titanic.j","Titanic.n"),
          c("Satz.j","Satz.n"),c("w","m"))
> loglinbsp

 , , w
```

```
           Satz.j Satz.n
Titanic.j     64     16
Titanic.n     14      6

, , m

           Satz.j Satz.n
Titanic.j     28     32
Titanic.n     14     26
```

Wir passen das Modell AC, BC an:

```
> e<-loglin(loglinbsp,list(c(1,3),c(2,3)),print=F,fit=T)
```

und betrachten das Ergebnis:

```
> e
 $lrt:
 [1] 2.234499

 $pearson:
 [1] 2.273396

 $df:
 [1] 2

 $margin:
 $margin[[1]]:
 [1] 1 3

 $margin[[2]]:
 [1] 2 3


 $fit:

 , , w
           Satz.j Satz.n
 Titanic.j   62.4   17.6
 Titanic.n   15.6    4.4

 , , m
           Satz.j Satz.n
 Titanic.j   25.2   34.8
 Titanic.n   16.8   23.2
```

Das Ergebnis ist eine Liste. Sehen wir uns jede Komponente dieser Liste an. Der Wert der Likelihood-Quotienten-Teststatistik steht in `e$lrt`, während die Komponente `e$pearson` den Wert der Teststatistik des χ^2-Unabhängigkeitstests enthält. Die Anzahl der Freiheitsgrade des Modells steht in `e$df`. Die geschätzten Häufigkeiten finden wir in `e$fit`. Wenn wir nur die Güte dieses Modells testen wollen, können wir die Argumente geeignet benutzen. Den kritischen Wert des Tests der Hypothesen

H_0: Das Modell AC, BC ist das wahre Modell,

H_1: Das Modell AC, BC ist nicht das wahre Modell

zum Signifikanzniveau $\alpha = 0.05$ liefert der Aufruf

```
> qchisq(0.95,e$df)
[1] 5.991465
```

Dabei bestimmt die Funktion `qchisq` das 0.95-Quantil der χ^2-Verteilung mit `e$df` Freiheitsgraden. Der Aufruf

```
> e$lrt > qchisq(0.95,e$df)
[1] F
```

liefert die Entscheidung. Ist das Ergebnis `T`, so wird die Nullhypothese abgelehnt. Im Beispiel wird sie also nicht abgelehnt. Um die Überschreitungswahrscheinlichkeit zu bestimmen, benötigen wir den Wert der Verteilungsfunktion der χ^2-Verteilung mit `df` Freiheitsgraden an der Stelle `e$lrt`. Diesen erhalten wir durch

```
> pchisq(e$lrt,e$df)
[1] 0.6728215
```

Die Überschreitungswahrscheinlichkeit liefert dann folgender Aufruf:

```
> 1-pchisq(e$lrt,e$df)
[1] 0.3271785
```

Schauen wir uns die Modellselektion in `S-PLUS` an. Wir beginnen mit dem Modell A, B, C:

```
> e<-loglin(loglinbsp,list(1,2,3),print=F)
```

Wir weisen der Variablen `e.a.b.c` den Wert der Likelihood-Quotienten-Teststatistik und die Anzahl der Freiheitsgrade zu:

```
> e.a.b.c<-c(e$lrt,e$df)
> e.a.b.c
[1] 39.66208  4.00000
```

Wenden wir uns den Modellen BC, A, AC, B und AB, C zu. Wir wollen auch bei diesen den Wert der Likelihood-Quotienten-Teststatistik und die Anzahl der Freiheitsgrade speichern. Hierzu erzeugen wir eine Matrix `modelle1`. Jeder Zeile dieser Matrix weisen wir die Charakteristika eines Modells zu:

```
> modelle1<-matrix(0,3,2)
```

Die folgende Befehlsfolge erzeugt die Charakteristika der Modelle und weist sie der Matrix `modelle` zu:

```
> ind<-1:3
> for (i in 1:3)
  {m1<-ind[-i]
   m2<-i
  e<-loglin(loglinbsp,list(m1,m2),print=F)
   modelle1[i,]<-c(e$lrt,e$df)
  }
```

Wir schauen uns `modelle1` an:

```
> modelle1
          [,1] [,2]
 [1,] 11.89742    3
 [2,] 29.99918    3
 [3,] 33.38369    3
```

Wir vergleichen zunächst das Modell A, B, C mit den Modellen BC, A, AC, B und AB, C. Hierzu bestimmen wir zunächst die Differenzen in den Werten der Teststatistiken:

```
> dt<-e.a.b.c[1]-modelle1[1:3,1]
> dt
[1] 27.764666  9.662903  6.278387
```

und die Differenzen der Freiheitsgrade:

```
> ddf<-e.a.b.c[2]-modelle1[1:3,2]
> ddf
[1] 1 1 1
```

Die Überschreitungswahrscheinlichkeiten der Übergänge sind:

```
> pvalue<-1-pchisq(dt,ddf)
> pvalue
[1] 1.370056e-007 1.880263e-003 1.222193e-002
```

Wir sehen, dass alle Übergänge signifikant sind:

```
> pvalue<0.05
[1] T T T
```

Die Nummer des bei diesem Übergang besten Modells ist

```
> nbest<-(1:3)[dt==max(dt)]
> nbest
[1] 1
```

Es ist das Modell

```
> abc<-c("A","B","C")
> cat(c(abc[ind[-nbest]],",",abc[nbest]))
  B C , A
```

Wir überprüfen, ob dieses Modell geeignet ist:

```
> 1-pchisq(modelle1[nbest,1],modelle1[nbest,2])
[1] 0.007742962
```

Wir verwerfen dieses Modell und gehen weiter zu den Modellen AC, BC und AB, AC.

Mit der folgenden Befehlsfolge passt man die Modelle AC, BC und AB, AC an:

```
> modelle2<-matrix(0,2,2)
> for (i in 1:2)
 { m1<-ind[-nbest]
   m2<-c(nbest,ind[-nbest][i])
   e<-loglin(loglinbsp,list(m1,m2),print=F)
   modelle2[i,]<-c(e$lrt,e$df)
 }
> modelle2
          [,1] [,2]
 [1,] 5.619019    2
 [2,] 2.234499    2
```

Wir vergleichen die Modelle AB, AC und AC, BC mit dem Modell AC, B. Hierzu bestimmen wir zunächst die Differenzen in den Werten der Teststatistiken:

```
> dt<-modelle1[nbest,1]-modelle2[1:2,1]
> dt
[1] 6.278397 9.662917
```

und die Differenzen der Freiheitsgrade:

```
> ddf<-modelle1[nbest,2]-modelle2[1:2,2]
> ddf
[1] 1 1
```

Die Überschreitungswahrscheinlichkeiten der Übergänge sind

```
> pvalue<-1-pchisq(dt,ddf)
> pvalue
[1] 0.012221864 0.001880249
```

Beide Übergänge sind signifikant:

```
> pvalue<0.05
[1] T T
```

Die Nummer des bei diesem Übergang besten Modells ist

```
> nbestneu<-(1:2)[dt==max(dt)]
> nbestneu
[1] 2
```

Es ist das Modell

```
> cat(c(abc[ind[-nbest]],",",
      abc[c(nbest,ind[-nbest][nbestneu])]))
B C , A C
```

Wir überprüfen, ob dieses Modell geeignet ist:

```
> 1-pchisq(modelle2[nbestneu,1],modelle2[nbestneu,2])
[1] 0.3271785
```

Wir akzeptieren das Modell.

10.5 Ergänzungen und weiterführende Literatur

Wir haben uns mit loglinearen Modellen für zwei- und dreidimensionale Tabellen beschäftigt. Die beschriebene Vorgehensweise kann auf Modelle für mehr als drei Merkmale übertragen werden. Beispiele hierfür sind bei Agresti (1990), Andersen (1991), Christensen (1997) und Fahrmeir et al. (1996), Kapitel 10, zu finden. Die Interpretation der Modelle wird mit wachsender Dimension jedoch immer schwieriger. Es gibt eine Vielzahl unterschiedlicher Modellselektionsverfahren für loglineare Modelle. Diese sind in den oben genannten Quellen zu finden. Bei unserer Darstellung ist klar geworden, warum der Begriff 'hierarchisch' bei hierarchischen loglinearen Modellen verwendet wird. Wir sind aber nicht darauf eingegangen, warum der Begriff 'loglinear' verwendet wird. Der Grund ist ganz einfach. Man benötigt hierfür Kenntnisse über mehrfaktorielle Varianzanalyse. Da wir in diesem Buch aber nur die einfaktorielle Varianzanalyse betrachten, haben wir einen anderen Zugang zu hierarchischen loglinearen Modellen gewählt. Die Formulierung als varianzanalytisches Modell ist in den oben genannten Quellen zu finden.

10.6 Übungen

Exercise 25. Schreiben Sie in `S-PLUS` eine Funktion, die das beste loglineare Modell für eine dreidimensionale Kontingenztabelle liefert.

Exercise 26. Im Wintersemester 2001/2002 wurden an der Fakultät für Wirtschaftswissenschaften der Universität Bielefeld 299 Studenten befragt. Unter anderem wurde nach dem Merkmal `Geschlecht` gefragt. Die Studierenden wurden auch gefragt, ob sie bei den Eltern wohnen. Wir bezeichnen dieses Merkmal mit `Eltern`. Außerdem sollten die Studierenden angeben, ob sie in Bielefeld studieren wollten. Wir bezeichnen dieses Merkmal mit `Bielefeld`. In den Tabellen 10.22 und 10.23 sind die Kontingenztabellen der Merkmale `Eltern` und `Bielefeld` bei den Männern und Frauen zu finden.
Der Zusammenhang zwischen den Variablen soll mit Hilfe eines loglinearen Modells bestimmt werden. Im Folgenden entspricht A dem Merkmal `Eltern`, B dem Merkmal `Ausbildung` und C dem Merkmal `Geschlecht`.

1. Interpretieren Sie die folgenden Modelle:
   a) A, B, C,

Table 10.22. Kontingenztabelle der Merkmale `Eltern` und `Bielefeld` bei den Männern

Eltern	Bielefeld nein	ja
nein	35	67
ja	13	71

Table 10.23. Kontingenztabelle der Merkmale `Eltern` und `Bielefeld` bei den Frauen

Eltern	Bielefeld nein	ja
nein	27	37
ja	9	40

b) AB, C,
c) AC, BC.

2. Passen Sie die folgenden Modelle an:
   a) A, B, C,
   b) AB, C,
   c) AC, BC.
3. Tabelle 10.24 zeigt die Werte von $G(M)$ und die Freiheitsgrade der einzelnen Modelle. Welches Modell beschreibt den Zusammenhang am besten?

Table 10.24. Werte von $G(M)$ und Freiheitsgrade *df* loglinearer Modelle

Modell M	$G(M)$	*df*
A, B, C	17.73	4
BC, A	16.47	3
AC, B	17.64	3
AB, C	1.314	3
AB, AC	1.222	2
AB, BC	0.051	2
AC, BC	16.379	2
AB, AC, BC	0.0492	1
ABC	0	0

Exercise 27. Im Wintersemester 2001/2002 wurden im Rahmen einer Befragung der Hörer der Vorlesung Einführung in die Ökonometrie die Merkmale `Haarfarbe` und `Augenfarbe` der Teilnehmer erhoben. Die Ergebnisse der Befragung sind in Tabelle 10.25 zu finden.

Table 10.25. Haarfarbe und Augenfarbe von Studenten

Haarfarbe \ Augenfarbe	blau	graublau	grün	braun
blond	21	2	14	3
dunkelblond	10	5	6	6
braun	7	3	19	26
rot	0	3	0	3
schwarz	0	0	2	21

Das Merkmal `Haarfarbe` wird mit A und das Merkmal `Augenfarbe` mit B bezeichnet. Es soll ein geeignetes loglineares Modell gefunden werden, das den Zusammenhang zwischen den beiden Merkmalen beschreibt.

1. Interpretieren Sie die folgenden Modelle:
   a) 0,
   b) A,
   c) B,
   d) A, B,
   e) AB.
2. Passen Sie die folgenden Modelle an:
   a) 0,
   b) A,
   c) B,
   d) A, B,
   e) AB.
3. Welches Modell beschreibt den Sachverhalt am besten?

Part IV

Gruppenstruktur

11 Einfaktorielle Varianzanalyse

11.1 Problemstellung

Bisher sind wir davon ausgegangen, dass alle Objekte aus einer Grundgesamtheit stammen. Jetzt wollen wir die Grundgesamtheit hinsichtlich eines Merkmals in unterschiedliche Teilgesamtheiten zerlegen. Von Interesse ist dann, ob sich die Verteilung eines oder mehrerer Merkmale in diesen Teilgesamtheiten unterscheidet.

Example 43. Im Rahmen der PISA-Studie wurde auch der Zeitaufwand der Schüler für Hausaufgaben erhoben (vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S.417). Dort wird unterschieden zwischen sehr geringem, geringem, mittlerem, großem und sehr großem Aufwand. Wir fassen die Länder mit sehr geringem und geringem Aufwand und die Länder mit großem und sehr großem Aufwand zusammen. Somit liegen drei Gruppen vor. Die Gruppe der Länder mit wenig Zeitaufwand nennen wir im Folgenden Gruppe 1, die Gruppe der Länder mit mittlerem Zeitaufwand Gruppe 2 und die Gruppe der Länder mit großem Zeitaufwand Gruppe 3. Wir wollen vergleichen, ob sich die Verteilung des Merkmals `Mathematische Grundbildung` in den drei Gruppen unterscheidet. Wir könnten aber auch daran interessiert sein, ob sich die drei Merkmale `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung` in den drei Gruppen unterscheidet. □

Wird nur untersucht, ob sich die Verteilung eines Merkmals in mehreren Gruppen unterscheidet, so spricht man von *univariater Varianzanalyse*. Werden hingegen mehrere Merkmale gleichzeitig betrachtet, so hat man es mit *multivariater Varianzanalyse* zu tun.

11.2 Univariate einfaktorielle Varianzanalyse

11.2.1 Theorie

Es soll untersucht werden, ob die Verteilung einer Zufallsvariablen Y in mehreren Gruppen identisch ist. Ausgangspunkt sind die Realisationen y_{ij} der unabhängigen Zufallsvariablen Y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$. Dabei

bezieht sich der Index i auf die i-te Gruppe, während der Index j sich auf die j-te Beobachtung bezieht. In der i-ten Gruppe liegen also n_i Beobachtungen vor. Die einzelnen Gruppen können unterschiedlich groß sein. Die Gesamtzahl aller Beobachtungen bezeichnen wir wie bisher mit n. hmcounterend. (fortgesetzt)

Example 43. In Tabelle 11.1 sind die Werte des Merkmals `Mathematische Grundbildung` in den einzelnen Gruppen zu finden.

Table 11.1. Merkmal Mathematische Grundbildung in den Gruppen

Gruppe 1		Gruppe 2		Gruppe 3	
Land	Punkte	Land	Punkte	Land	Punkte
FIN	536	AUS	533	GR	447
J	557	B	520	GB	529
FL	514	BR	334	IRL	503
L	446	DK	514	I	457
A	515	D	490	LV	463
S	510	F	517	MEX	387
CH	529	IS	514	PL	470
CZ	498	CDN	533	RUS	478
		ROK	547	E	476
		NZ	537	H	488
		N	499		
		P	454		
		USA	493		

□

Wir unterstellen im Folgenden, dass die Y_{ij} normalverteilt sind mit Erwartungswert μ_i, $i = 1, \ldots, I$ und Varianz σ^2. Die Erwartungswerte der Gruppen können sich also unterscheiden, während die Varianz identisch sein muss. Es ist zu testen:

$$H_0 : \ \mu_1 = \ldots = \mu_I \tag{11.1}$$

gegen

$$H_1 : \ \mu_i \neq \mu_j \quad \text{für mind. ein Paar } (i,j) \text{ mit } i \neq j \, .$$

Es liegt nahe zur Überprüfung von (11.1) die Mittelwerte

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \tag{11.2}$$

der einzelnen Gruppen zu bestimmen und zu vergleichen. hmcounterend. (fortgesetzt)

Example 43. Es gilt $\bar{y}_1 = 513.125$, $\bar{y}_2 = 498.8462$ und $\bar{y}_3 = 469.8$. Die Mittelwerte unterscheiden sich. □

Der Vergleich von zwei Mittelwerten $\bar{y}_1$ und $\bar{y}_2$ ist einfach. Wir bilden die Differenz $\bar{y}_1 - \bar{y}_2$ der beiden Mittelwerte. Bei mehr als zwei Gruppen können wir alle Paare von Gruppen betrachten und $\bar{y}_i$ mit $\bar{y}_j$ für $i < j$ vergleichen. Hierdurch erhalten wir aber kein globales Maß für den Vergleich aller Gruppen. Um dieses zu erhalten, fassen wir die Mittelwerte $\bar{y}_i$, $i = 1, \ldots, I$ als eine Stichprobe auf und bestimmen, wie stark sie um den Mittelwert

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij} \tag{11.3}$$

aller Beobachtungen streuen. hmcounterend. (fortgesetzt)

Example 43. Es gilt $\bar{y} = 493.1613$. □

Es liegt nahe, die Streuung der Mittelwerte $\bar{y}_i$ um das Gesamtmittel $\bar{y}$ folgendermaßen zu bestimmen:

$$\sum_{i=1}^{I} (\bar{y}_i - \bar{y})^2.$$

Hierbei wird aber nicht berücksichtigt, dass die Gruppen unterschiedlich groß sein können. Eine große Gruppe sollte ein stärkeres Gewicht erhalten als eine kleine Gruppe. Wir bilden also

$$SS_B = \sum_{i=1}^{I} n_i \, (\bar{y}_i - \bar{y})^2. \tag{11.4}$$

Man bezeichnet SS_B als *Streuung zwischen den Gruppen*. hmcounterend. (fortgesetzt)

Example 43. Es gilt

$$\begin{aligned} SS_B &= 8(513.125 - 493.1613)^2 + 13(498.8462 - 493.1613)^2 \\ &\quad + 10(469.8 - 493.1613)^2 = 9066.03. \end{aligned}$$

□

Wie das folgende Beispiel zeigt, ist die Größe SS_B allein aber keine geeignete Teststatistik zur Überprüfung der Hypothese (11.1).

Example 44. In der Tabelle 11.2 sind die Werte eines Merkmals in drei Gruppen zu finden.

Es gilt

$$\bar{y}_1 = 49, \quad \bar{y}_2 = 56, \quad \bar{y}_3 = 51, \quad \bar{y} = 52.$$

Table 11.2. Werte eines Merkmals in drei Gruppen mit kleiner Streuung innerhalb der Gruppen

Gruppe	Werte
1	47 53 49 50 46
2	55 54 58 61 52
3	53 50 51 52 49

Table 11.3. Werte eines Merkmals in drei Gruppen mit großer Streuung innerhalb der Gruppen

Gruppe	Werte
1	50 42 53 45 55
2	48 57 65 59 51
3	57 59 48 46 45

In der Tabelle 11.3 sind ebenfalls die Werte eines Merkmals in drei Gruppen zu finden. Auch dort gilt

$$\bar{y}_1 = 49, \quad \bar{y}_2 = 56, \quad \bar{y}_3 = 51, \quad \bar{y} = 52.$$

Also ist auch in beiden Tabellen der Wert von SS_B identisch. Wie die Abbildungen 11.1 und 11.2 zeigen, unterscheiden sich die beiden Situationen beträchtlich. Die Boxplots in Abbildung 11.1 verdeutlichen, dass die Streuung innerhalb der Gruppen klein ist, während in Abbildung 11.2 die Streuung innerhalb der Gruppen groß ist. Abbildung 11.1 spricht für einen Lageunterschied zwischen den Gruppen, während die unterschiedlichen Mittelwerte in 11.2 eher durch die hohen Streuungen erklärt werden können.
Die Stichprobenvarianzen in den Gruppen für die Beobachtungen in Tabelle 11.2 sind

$$s_1^2 = 7.5, \quad s_2^2 = 12.5, \quad s_3^2 = 2.5\,.$$

Für die Gruppen in Tabelle 11.3 erhält man folgende Stichprobenvarianzen:

$$s_1^2 = 29.5, \quad s_2^2 = 45.0, \quad s_3^2 = 42.5\,.$$

□

Wir müssen also neben der Streuung zwischen den Gruppen die Streuung innerhalb der Gruppen berücksichtigen. Die Streuung innerhalb der i-ten Gruppe messen wir durch

$$\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2\,. \tag{11.5}$$

Summieren wir (11.5) über alle Gruppen, so erhalten wir

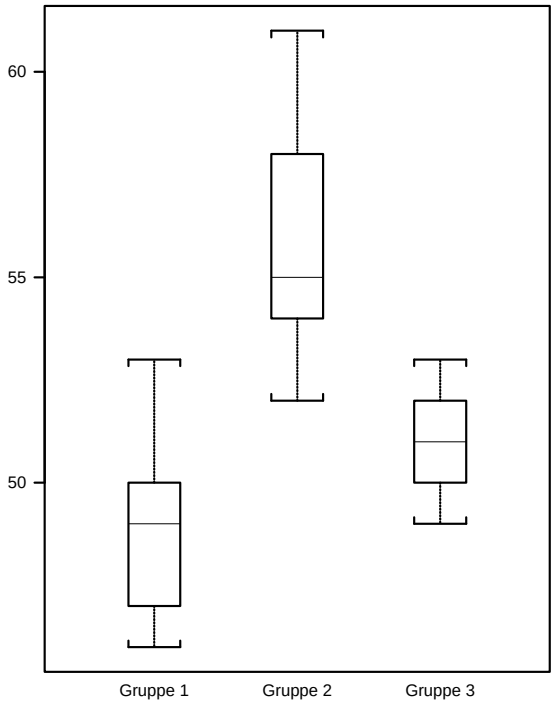


Fig. 11.1. Boxplot von drei Gruppen mit kleiner Streuung innerhalb der Gruppen

$$SS_W = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 . \tag{11.6}$$

Wir nennen SS_W auch *Streuung innerhalb der Gruppen.* hmcounterend. (fortgesetzt)

Example 43. Es gilt $SS_W = 56720.17$. □

Die Gesamtstreuung messen wir durch:

$$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 . \tag{11.7}$$

hmcounterend. (fortgesetzt)

Example 43. Es gilt $SS_T = 65786.2$. □

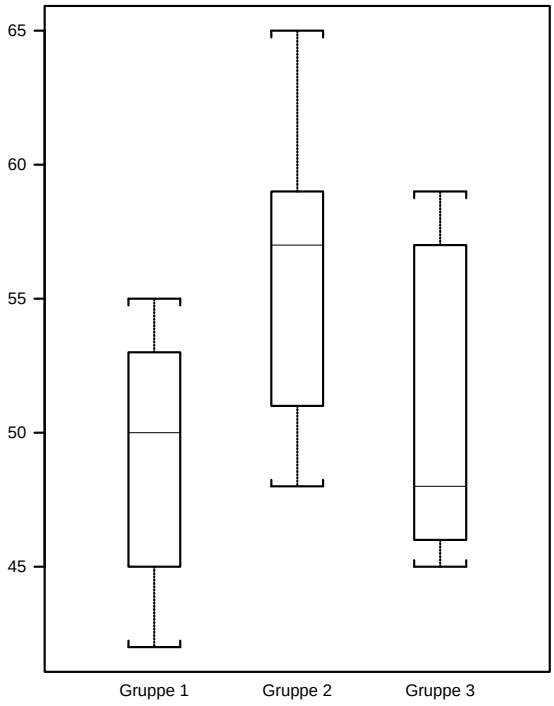


Fig. 11.2. Boxplot von drei Gruppen mit kleiner Streuung innerhalb der Gruppen

Im Beispiel gilt

$$SS_T = SS_B + SS_W \,. \tag{11.8}$$

Dies ist kein Zufall. Diese Beziehung gilt allgemein, wie man folgendermaßen sieht:

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{y})^2$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)(\bar{y}_i-\bar{y})$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{I} n_i(\bar{y}_i-\bar{y})^2 + 2\sum_{i=1}^{I}(\bar{y}_i-\bar{y})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{I} n_i(\bar{y}_i-\bar{y})^2$$

$$= SS_B + SS_W\,.$$

Hierbei haben wir die folgende Beziehung berücksichtigt:

$$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i) = \sum_{j=1}^{n_i} y_{ij} - \sum_{j=1}^{n_i}\bar{y}_i = n_i\,\bar{y}_i - n_i\,\bar{y}_i = 0\,.$$

Eine geeignete Teststatistik erhält man nun, indem man die mittleren Streuungen vergleicht, wobei der Mittelwert unter der Nebenbedingung bestimmt wird, wie viele der Summanden frei gewählt werden können. Die Streuung zwischen den Stichproben setzt sich aus I Summanden zusammen, von denen aber nur $I-1$ frei gewählt werden können, da sich der Mittelwert der I-ten Stichprobe aus

$$\bar{y}, \bar{y}_1, \ldots, \bar{y}_{I-1}$$

ergibt. Die Streuung innerhalb der Stichproben setzt sich aus n Summanden zusammen. In der i-ten Stichprobe ergibt sich aber y_{in_i} aus der Kenntnis von

$$y_{i1}, \ldots, y_{in_i-1}, \bar{y}_i\,.$$

Somit sind von den n Summanden nur $n - I$ frei wählbar. Wir erhalten also $MSS_B = SS_B/(I-1)$ und $MSS_W = SS_W/(n-I)$. hmcounterend. (fortgesetzt)

Example 43. Es gilt $MSS_B = 4533.013$ und $MSS_W = 2025.72$. □

Die Teststatistik ist

$$F = \frac{MSS_B}{MSS_W} = \frac{\frac{1}{I-1}\sum_{i=1}^{I} n_i \left(\bar{Y}_i - \bar{Y}\right)^2}{\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2}. \tag{11.9}$$

Ist die mittlere Streuung zwischen den Stichproben groß im Verhältnis zur mittleren Streuung innerhalb der Stichproben, so wird die Nullhypothese identischer Erwartungswerte abgelehnt. Unter der Nullhypothese ist die Teststatistik in (11.9) F-verteilt mit $I-1$ und $n-I$ Freiheitsgraden. Der Beweis ist bei Seber (1977), S. 97 zu finden.

Wir lehnen die Hypothese (11.1) zum Niveau α ab, wenn gilt $F > F_{I-1,n-I;1-\alpha}$, wobei $F_{I-1,n-I;1-\alpha}$ das $1-\alpha$-Quantil der F-Verteilung mit $I-1$ und $n-I$ Freiheitsgraden ist. hmcounterend. (fortgesetzt)

Example 43. Es gilt

$$F = \frac{4533.013}{2025.72} = 2.2377.$$

Der Tabelle C.6 auf Seite 508 entnehmen wir $F_{2,28;0.95} = 3.34$. Wir lehnen die Hypothese (11.1) also nicht ab. □

Man spricht auch vom F-Test. Da die Teststatistik das Verhältnis von zwei Schätzern der Varianz σ^2 ist, spricht man von *Varianzanalyse*. Die Ergebnisse einer Varianzanalyse werden in einer ANOVA-Tabelle zusammengestellt. Dabei steht ANOVA für Analysis Of Variance. Tabelle 11.4 zeigt den allgemeinen Aufbau einer ANOVA-Tabelle.

Table 11.4. Allgemeiner Aufbau einer ANOVA-Tabelle

Quelle der Variation	Quadratsummen	Freiheits-grade	Mittlere Quadratsummen	F
zwischen den Gruppen	SS_B	$I-1$	MSS_B	$\frac{MSS_B}{MSS_W}$
innerhalb der Gruppen	SS_W	$n-I$	MSS_W	
Gesamt	SS_T	$n-1$		

hmcounterend. (fortgesetzt)

Example 43. In Tabelle 11.5 ist die ANOVA-Tabelle zu finden.

Table 11.5. ANOVA-Tabelle für den Vergleich des Merkmals Mathematische Grundbildung in den 3 Gruppen

Quelle der Variation	Quadratsummen	Freiheits-grade	Mittlere Quadratsummen	F
zwischen den Gruppen	9066.03	2	4533.013	2.2377
innerhalb der Gruppen	56720.17	28	2025.720	
Gesamt	65786.2	30		

□

Wir wollen nun noch den Fall $I = 2$ betrachten. Man spricht auch vom *unverbundenen Zweistichprobenproblem*. Wir gehen aus von den Realisationen y_{ij} der unabhängigen Zufallsvariablen Y_{ij}, wobei wir unterstellen, dass Y_{ij} normalverteilt ist mit Erwartungswert μ_i und Varianz σ^2 für $i = 1, 2$, $j = 1, \ldots, n_i$. Es soll getestet werden:

$$H_0 : \ \mu_1 = \mu_2 \tag{11.10}$$

gegen

$$H_1 : \ \mu_1 \neq \mu_2 \, .$$

Unter der Annahme der Normalverteilung sollte man den t-Test anwenden. Dessen Teststatistik lautet

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{11.11}$$

mit

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 .$$

Wenn die Hypothese (11.10) zutrifft, ist die Teststatistik in (11.11) t-verteilt mit $n_1 + n_2 - 2$ Freiheitsgraden. Wir lehnen H_0 zum Signifikanzniveau α ab, wenn gilt

$$|t| > t_{1-\alpha/2;n_1+n_2-2} .$$

Dabei ist $t_{1-\alpha/2;n_1+n_2-2}$ das $1-\alpha$-Quantil der t-Verteilung mit $n_1 + n_2 - 2$ Freiheitsgraden.

hmcounterend. (fortgesetzt)

Example 43. Wir vergleichen die Gruppe 1 mit der Gruppe 2.

Das Testproblem lautet:

$$H_0 : \ \mu_1 = \mu_2$$

gegen

$$H_1 : \ \mu_1 \neq \mu_2.$$

Es gilt

$$t = 0.6596.$$

Der Tabelle C.4 auf Seite 506 entnehmen wir $t_{0.975;19} = 2.093$. Wir lehnen H_0 also nicht ab. □

11.2.2 Praktische Aspekte

Überprüfung der Normalverteilungsannahme Der F-Test beruht auf der Annahme der Normalverteilung. Es gibt eine Reihe von Möglichkeiten die Gültigkeit dieser Annahme zu überprüfen. Man kann einen Test auf Normalverteilung wie den *Kolmogorow-Smirnow-Test* durchführen. Dieser und viele andere Tests auf Normalverteilung sind bei Büning & Trenkler (1994) zu finden. Wir wollen eine graphische Darstellung betrachten, mit der man die Annahme der Normalverteilung überprüfen kann. Bei einem *Normal-Quantil-Plot* zeichnet man die geordneten Beobachtungen $y_{(1)}, \ldots, y_{(n)}$ gegen Quantile der Normalverteilung. Bei der Wahl der Quantile gibt es mehrere Möglichkeiten. Die empirische Verteilungsfunktion $\hat{F}(y)$ ist der Anteil der Beobachtungen, die kleiner oder gleich y sind. Sind keine identischen Beobachtungen in der Stichprobe, so gilt $\hat{F}(y_{(i)}) = i/n$. Somit wird $y_{(i)}$ über die empirische Verteilungsfunktion das i/n-Quantil zugeordnet. Diese Vorgehensweise hat aber den Nachteil, dass die beiden Ränder der Verteilung nicht gleich behandelt werden. Dieses Problem kann man dadurch umgehen, dass man $y_{(i)}$ das $(i-0.5)/n$-Quantil zuordnet. Bei einem Normal-Quantil-Plot zeichnet man also die geordneten Beobachtungen $y_{(1)}, \ldots, y_{(n)}$ gegen die Quantile $\Phi^{-1}((1-0.5)/n), \ldots, \Phi^{-1}((n-0.5)/n)$ der Standardnormalverteilung. Liegt Normalverteilung vor, so sollten die Punkte um eine Gerade streuen. Bei der einfaktoriellen Varianzanalyse werden die Residuen $e_{ij} = y_{ij} - \bar{y}_i$, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$ gegen die Quantile der Standardnormalverteilung gezeichnet. hmcounterend. (fortgesetzt)

Example 43. Abbildung 11.3 zeigt den Normal-Quantil-Plot der Residuen.

Der Plot deutet darauf hin, dass die Normalverteilungsannahme nicht gerechtfertigt ist.

□

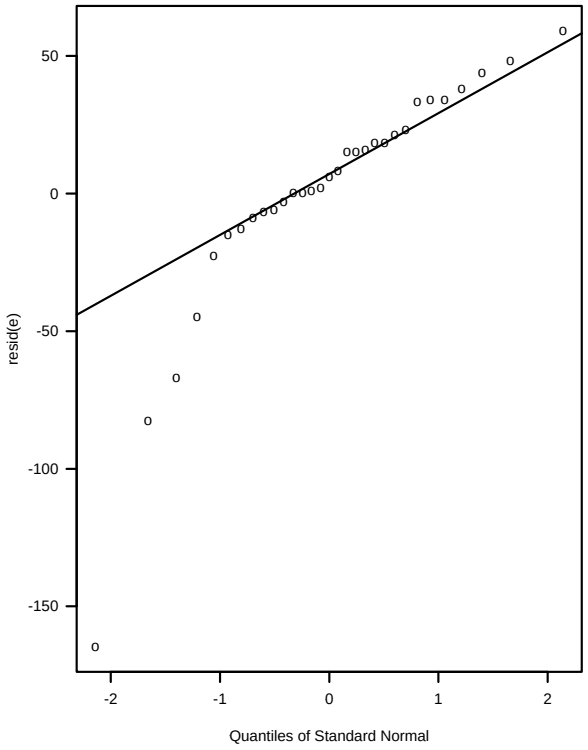


Fig. 11.3. Normal-Quantil-Plot bei einfaktorieller Varianzanalyse

Der Kruskal-Wallis-Test Ist die Annahme der Normalverteilung nicht gerechtfertigt, so sollte man einen nichtparametrischen Test durchführen. Am bekanntesten ist der *Kruskal-Wallis-Test*. Dieser beruht auf der Annahme, dass die Beobachtungen y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$ Realisationen von unabhängigen Zufallsvariablen Y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$ mit stetiger Verteilungsfunktion sind. Es ist zu testen

$$H_0: \text{ Die Verteilungen in allen Gruppen sind identisch} \qquad (11.12)$$

gegen

$H_1:$ Mindestens zwei Gruppen unterscheiden sich hinsichtlich der Lage.

Der Kruskal-Wallis-Test beruht auf den *Rängen* R_{ij} der y_{ij}, $i = 1, \ldots, I$, $j = 1 \ldots, n_i$, unter allen Beobachtungen. Dabei ist der Rang R_{ij} gleich der Anzahl der Beobachtungen, die kleiner oder gleich y_{ij} sind. Sind Beobach-

tungen identisch, so spricht man von *Bindungen*. In diesem Fall vergibt man für die gebundenen Werte *Durchschnittsränge*. hmcounterend. (fortgesetzt)

Example 43. Tabelle 11.6 zeigt die Länder mit den zugehörigen Rängen.

Table 11.6. Ränge des Merkmals Mathematische Grundbildung

Gruppe 1		Gruppe 2		Gruppe 3	
Land	Rang	Land	Rang	Land	Rang
FIN	28.0	AUS	26.5	GR	4.0
J	31.0	B	23.0	GB	24.5
FL	19.0	BR	1.0	IRL	16.0
L	3.0	DK	19.0	I	6.0
A	21.0	D	12.0	LV	7.0
S	17.0	F	22.0	MEX	2.0
CH	24.5	IS	19.0	PL	8.0
CZ	14.0	CDN	26.5	RUS	10.0
		ROK	30.0	E	9.0
		NZ	29.0	H	11.0
		N	15.0		
		P	5.0		
		USA	13.0		

□

Beim Kruskal-Wallis-Test werden nun für $i = 1, \ldots, I$ die Rangsummen R_i in den einzelnen Gruppen bestimmt:

$$R_i = \sum_{j=1}^{n_i} R_{ij}.$$

hmcounterend. (fortgesetzt)

Example 43. Es gilt

$$R_1 = 157.5, \quad R_2 = 241, \quad R_3 = 97.5.$$

□

Diese Rangsummen werden mit ihren Erwartungswerten $E(R_i)$ unter (11.12) verglichen. Wenn keine Bindungen vorliegen, so werden bei n Beobachtungen die Ränge $1, \ldots, n$ vergeben. Trifft (11.12) zu, so ist für eine Beobachtung jeder Rang gleichwahrscheinlich. Es gilt also

$$P(R_{ij} = k) = \frac{1}{n}$$

für $k = 1, \ldots, n$, $i = 1, \ldots, I$ und $j = 1, \ldots, n_i$. Der erwartete Rang $E(R_{ij})$ von Y_{ij} ist dann

$$E(R_{ij}) = \sum_{k=1}^{n} k \, \frac{1}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

Die erwartete Rangsumme der i-ten Gruppe ist somit

$$E(R_i) = E\left(\sum_{j=1}^{n_i} R_{ij}\right) = \sum_{j=1}^{n_i} E(R_{ij}) = \sum_{j=1}^{n_i} \frac{n+1}{2} = \frac{n_i(n+1)}{2}.$$

hmcounterend. (fortgesetzt)

Example 43. Mit $n = 31$, $n_1 = 8$, $n_2 = 13$ und $n_3 = 10$ gilt

$$E(R_1) = 128, \quad E(R_2) = 208, \quad E(R_3) = 160.$$

□

Die Teststatistik des Kruskal-Wallis-Tests vergleicht die Rangsummen R_i mit ihren Erwartungswerten $E(R_i)$. Sie lautet:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{1}{n_i} \left(R_i - \frac{n_i(n+1)}{2}\right)^2. \tag{11.13}$$

hmcounterend. (fortgesetzt)

Example 43. Es gilt

$$H = \frac{12}{31 \cdot 32} \left[\frac{(157.5 - 128)^2}{8} + \frac{(241 - 208)^2}{13} + \frac{(97.5 - 160)^2}{10}\right]$$

$$= 7.054542.$$

□

Wir lehnen die Hypothese (11.12) ab, wenn gilt $H \geq h_{1-\alpha}$. Dabei ist $h_{1-\alpha}$ das $1-\alpha$-Quantil der Verteilung von H. Die Verteilung von H ist für kleine Werte von n bei Büning & Trenkler (1994) tabelliert.

Für große Stichprobenumfänge ist H approximativ χ^2-verteilt mit $I-1$ Freiheitsgraden. Wir lehnen (11.12) ab, wenn gilt $H \geq \chi^2_{I-1,1-\alpha}$. Dabei ist $\chi^2_{I-1,1-\alpha}$ das $1-\alpha$-Quantil der χ^2-Verteilung mit $I-1$ Freiheitsgraden.

Im Beispiel liegen Bindungen vor. In diesem Fall wird H modifiziert zu

$$H^* = \frac{H}{1 - \frac{1}{n^3 - n} \sum_{l=1}^{r} (b_l^3 - b_l)}. \tag{11.14}$$

Dabei ist r die Anzahl der Gruppen mit identischen Beobachtungen und b_l die Anzahl der Beobachtungen in der l-ten Bindungsgruppe. Wir lehnen (11.12) im Fall von Bindungen ab, wenn gilt $H^* \geq \chi^2_{I-1,1-\alpha}$. hmcounterend. (fortgesetzt)

Example 43. Der Wert 514 kommt dreimal und die Werte 529 und 533 kommen jeweils zweimal vor. Somit gibt es 2 Bindungsgruppen mit zwei Beobachtungen und eine Bindungsgruppe mit drei Beobachtungen. Hieraus folgt

$$1 - \frac{1}{n^3 - n} \sum_{l=1}^{r} (b_l^3 - b_l) = 0.99879.$$

Also ist $H^* = 7.0631$. Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{2,0.95} = 5.99$. Wir lehnen die Hypothese (11.12) zum Niveau 0.05 also ab. □

Im Beispiel wurde die Nullhypothese beim F-Test nicht abgelehnt, während sie beim Kruskal-Wallis-Test abgelehnt wurde. Welcher Testentscheidung kann man trauen? Da der Normal-Quantil-Plot darauf hindeutet, dass die Annahme der Normalverteilung nicht gerechtfertigt ist, ist der Kruskal-Wallis-Test für die Daten besser geeignet, sodass man dessen Entscheidung berücksichtigen sollte. Wir haben hier nur deshalb beide Tests auf den gleichen Datensatz angewendet, um die Vorgehensweise beider Tests zu illustrieren. In der Praxis muss man sich vor der Durchführung für einen Test entscheiden. Es besteht die Möglichkeit, datengestützt einen Test auszuwählen. Man spricht dann von einem *adaptiven Test*. Bei Büning (1996) werden adaptive Tests für die univariate einfaktorielle Varianzanalyse beschrieben.
Sollen nur zwei Gruppen hinsichtlich der Lage miteinander verglichen werden, so sollte man den *Wilcoxon-Test* anwenden. Das Testproblem lautet

$$H_0: \text{ Die Verteilungen in beiden Gruppen sind identisch}, \quad (11.15)$$

$$H_1: \text{ Die beiden Gruppen unterscheiden sich hinsichtlich der Lage}.$$

Beim Wilcoxon-Test werden wie beim Kruskal-Wallis-Test die Ränge R_{ij}, $i = 1, 2$, $j = 1, \ldots, n_i$ der Beobachtungen in den beiden Stichproben bestimmt. Die Teststatistik ist die Summe der Ränge der ersten Stichprobe:

$$W = \sum_{j=1}^{n_1} R_{1j}. \quad (11.16)$$

Die Hypothese (11.15) wird zum Signifikanzniveau α abgelehnt, wenn gilt

$$W \leq w_{\alpha/2}$$

oder

$$W \geq w_{1-\alpha/2}.$$

Dabei ist $w_{\alpha/2}$ das $\alpha/2$-Quantil und $w_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Teststatistik des Wilcoxon-Tests. Die Verteilung von W ist in Büning & Trenkler (1994) tabelliert. Für große Stichprobenumfänge kann die Verteilung von W durch die Normalverteilung approximiert werden. Wir bilden die Teststatistik

$$Z = \frac{W - 0.5\, n_1\, (N+1)}{\sqrt{n_1\, n_2\, (N+1)/12}} \tag{11.17}$$

mit $N = n_1 + n_2$. Wir lehnen die Hypothese (11.15) zum Signifikanzniveau α ab, wenn gilt $|Z| \geq z_{1-\alpha/2}$. Dabei ist $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung. Die Approximation durch die Normalverteilung ist auch für kleinere Stichprobenumfänge gut, wenn man in der Teststatistik berücksichtigt, dass die Verteilung der diskreten Zufallsvariablen W durch die stetige Normalverteilung approximiert wird. Man bildet

$$Z = \frac{W - 0.5 - 0.5\, n_1\, (N+1)}{\sqrt{n_1\, n_2\, (N+1)/12}}. \tag{11.18}$$

Eine Begründung für eine derartige *Stetigkeitskorrektur* ist bei Schlittgen (2000), S. 241-242, zu finden.

Liegen Bindungen vor, so muss man den Nenner in Gleichung (11.17) beziehungsweise (11.18) modifizieren. Man ersetzt ihn durch

$$\sqrt{\frac{n_1\, n_2}{12}\left[N + 1 - \frac{1}{N^2 - N}\sum_{l=1}^{r}(b_l^3 - b_l)\right]}.$$

Dabei ist r die Anzahl der Gruppen mit identischen Beobachtungen und b_l die Anzahl der Beobachtungen in der l-ten Bindungsgruppe. Wir bezeichnen die modifizierte Teststatistik mit Z^* und lehnen die Hypothese (11.15) zum Signifikanzniveau α ab, wenn gilt $|Z^*| \geq z_{1-\alpha/2}$. hmcounterend. (fortgesetzt)

Example 43. Wir vergleichen die Gruppe 1 mit der Gruppe 2. Die Beobachtungen in der ersten Stichprobe sind

```
536 557 514 446 515 510 529 498.
```

Die Beobachtungen in der zweiten Stichprobe sind

```
533 520 334 514 490 517 514 533 547 537 499 454 493.
```

Die Ränge der Beobachtungen in der ersten Stichprobe sind

```
18 21 10 2 12 8 15 6.
```

Die Ränge der Beobachtungen in der zweiten Stichprobe sind

```
16.5 14 1 10 4 13 10 16.5 20 19 7 3 5.
```

Somit gilt

$$W = 18 + 21 + 10 + 2 + 12 + 8 + 15 + 6 = 92.$$

Der Wert 514 kommt dreimal und der Wert 533 kommt zweimal vor. Somit gibt es eine Bindungsgruppe mit zwei Beobachtungen und eine Bindungsgruppe mit drei Beobachtungen. Es gilt

$$\frac{1}{N^2 - N} \sum_{l=1}^{r} (b_l^3 - b_l) = 0.0714.$$

Wenn wir die Stetigkeitskorrektur verwenden, erhalten wir folgenden Wert der Teststatistik:

$$Z^* = \frac{92 - 0.5 - 0.5 \cdot 8 \cdot 22}{\sqrt{\frac{8 \cdot 13}{12} [22 - 0.0714]}} = 0.2539.$$

Wegen $z_{0.975} = 1.96$ lehnen wir die Nullhypothese (11.15) zum Signifikanzniveau 0.05 nicht ab. Wir sehen, dass der Wilcoxon-Test zum gleichen Ergebnis wie der t-Test kommt. □

11.3 Multivariate einfaktorielle Varianzanalyse

Bisher haben wir beim Vergleich der Gruppen nur ein Merkmal betrachtet. Oft werden an jedem Objekt p quantitative Merkmale erhoben. Es liegen also die Realisationen $\mathbf{y}_{ij}$ der unabhängigen Zufallsvariablen $\mathbf{Y}_{ij}$ für $i = 1, \ldots, I$ und $j = 1, \ldots, n_i$ vor. Dabei bezieht sich der Index i wieder auf die Gruppe und der Index j auf die j-te Beobachtung in der jeweiligen Gruppe. Es gilt

$$\mathbf{Y}_{ij} = \begin{pmatrix} Y_{ij1} \\ \vdots \\ Y_{ijp} \end{pmatrix}.$$

Wir unterstellen, dass $\mathbf{Y}_{ij}$ für $i = 1, \ldots, I$ und $j = 1, \ldots, n_i$ multivariat normalverteilt ist mit Erwartungswert $\boldsymbol{\mu}_i$ und Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$. Wir gehen also wie bei der univariaten einfaktoriellen Varianzanalyse davon aus, dass die Gruppen unterschiedliche Erwartungswerte haben können, die Varianz-Kovarianz-Matrizen aber identisch sind.

Es ist zu testen:

$$H_0 : \ \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_I \tag{11.19}$$

gegen

$$H_1 : \ \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ für mind. ein Paar } (i, j) \text{ mit } i \neq j\,.$$

Wir bestimmen wie bei der einfaktoriellen Varianzanalyse das Gesamtmittel

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$$

und die Mittelwerte der Gruppen

$$\bar{\mathbf{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$$

für $i = 1, \ldots, I$. hmcounterend. (fortgesetzt)

Example 43. Wir betrachten die Merkmale `Lesekompetenz`, `Mathematische Grundbildung` und `Naturwissenschaftliche Grundbildung`. Es gilt

$$\bar{\mathbf{y}} = \begin{pmatrix} 493.452 \\ 493.161 \\ 492.613 \end{pmatrix}, \ \bar{\mathbf{y}}_1 = \begin{pmatrix} 500.125 \\ 513.125 \\ 505.625 \end{pmatrix},$$

$$\bar{\mathbf{y}}_2 = \begin{pmatrix} 499.308 \\ 498.846 \\ 494.615 \end{pmatrix}, \ \bar{\mathbf{y}}_3 = \begin{pmatrix} 480.500 \\ 469.800 \\ 479.600 \end{pmatrix}.$$

□

Wir ermitteln die Streuung innerhalb der Stichproben und die Streuung zwischen diesen. Wir bestimmen also die *Zwischen-Gruppen-Streumatrix*

$$\mathbf{B} = \sum_{i=1}^{I} n_i \, (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})' \tag{11.20}$$

und die *Inner-Gruppen-Streumatrix*

$$\mathbf{W} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)'. \tag{11.21}$$

hmcounterend. (fortgesetzt)

Example 43. Es gilt

$$\mathbf{B} = \begin{pmatrix} 2479.53 & 4524.25 & 2532.51 \\ 4524.25 & 9066.03 & 5266.13 \\ 2532.51 & 5266.13 & 3100.00 \end{pmatrix},$$

$$\mathbf{W} = \begin{pmatrix} 30802.14 & 38325.49 & 33335.91 \\ 38325.49 & 56720.17 & 44054.81 \\ 33335.91 & 44054.81 & 39469.35 \end{pmatrix}.$$

□

Es gibt eine Reihe von Vorschlägen für Teststatistiken, die auf $\mathbf{B}$ und $\mathbf{W}$ beruhen. Wir betrachten hier nur Wilks' Λ:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}. \tag{11.22}$$

hmcounterend. (fortgesetzt)

Example 43. Es gilt $\Lambda = 0.7009$. □

Die Nullhypothese (11.19) wird abgelehnt, wenn $\Lambda \leq \Lambda_{p,n-I,I-1;\alpha}$. Dabei ist $\Lambda_{p,n-I,I-1;\alpha}$ das α-Quantil der Λ-Verteilung mit den Parametern p, $n-I$ und $I-1$. Für bestimmte Parameterkonstellationen besteht ein Zusammenhang zwischen der Λ-Verteilung und der F-Verteilung. Diese Konstellationen sind in Tabelle 6.3 bei Johnson & Wichern (1998) zu finden. Ist speziell $I = 3$, so ist

$$\Lambda^* = \frac{n-p-2}{p} \, \frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \tag{11.23}$$

F-verteilt mit $2p$ und $2(n-p-2)$ Freiheitsgraden. Wir lehnen die Nullhypothese (11.19) ab, wenn $\Lambda^* > F_{2p,2(n-p-2);1-\alpha}$. Dabei ist $F_{2p,2(n-p-2);1-\alpha}$ das $1-\alpha$-Quantil der F-Verteilung mit $2p$ und $2(n-p-2)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 43. Es gilt

$$\Lambda^* = \frac{31-3-2}{3}\,\frac{1-\sqrt{0.7009}}{\sqrt{0.7009}} = 1.685\,.$$

Der Tabelle C.6 auf Seite 508 entnehmen wir $F_{6,52;0.95} = 2.28$. Wir lehnen (11.19) also nicht ab. □

Für große Werte von n ist $-(n-1-0.5(p+I))\ln\Lambda$ approximativ χ^2-verteilt mit $p(I-1)$ Freiheitsgraden. hmcounterend. (fortgesetzt)

Example 43. Wir schauen uns auch hier den Test an. Mit $n = 31$, $p = 3$ und $I = 3$ gilt

$$-(n-1-0.5(p+I))\ln\Lambda = 9.596.$$

Der Tabelle C.3 auf Seite 505 entnehmen wir $\chi^2_{6;0.95} = 12.59$. Wir lehnen die Nullhypothese (11.19) zum Niveau 0.05 also nicht ab. □

11.4 Einfaktorielle Varianzanalyse in S-PLUS

Wir wollen das Beispiel 43 in S-PLUS betrachten. Die Daten stehen in der Matrix `PISA`. Wir benötigen noch einen Vektor, der für jedes Land angibt, zu welcher der drei Gruppen es gehört. Hierzu erzeugen wir einen Vektor `gruppe`, der aus 31 Leerzeichen besteht:

```
> gruppe<-rep("",31)
```

Die Indizes der Länder, die wenig Zeit für Hausaufgaben aufwenden, sind `6`, `13`, `17`, `18`, `22`, `26`, `27` und `29`. Wir weisen dem Vektor `gruppe` an diesen Stellen das `"w"` zu:

```
> gruppe[c(6,13,17,18,22,26,27,29)]<-"w"
```

Entsprechend verfahren wir mit den anderen Gruppen. Die Indizes der Länder, deren Zeitaufwand für Hausaufgaben im mittleren Bereich liegt, sind `1`, `2`, `3`, `4`, `5`, `7`, `11`, `14`, `15`, `20`, `21`, `24` und `31`. Wir weisen `gruppe` an diesen Positionen ein `"m"` zu. Den restlichen Komponenten wird ein `"v"` zugewiesen. Nun müssen wir `gruppe` nur noch zu einem Faktor machen. Dies geschieht mit der Funktion `factor`:

```
> gruppe<-factor(gruppe)
> gruppe
 [1] m m m m m w m v v v m v w m m v w w v m m w v m v w
     w v w v m
```

Nach diesen Vorbereitungen können wir mit der Analyse beginnen. Wir schauen uns zuerst die univariate einfaktorielle Varianzanalyse an. Wir wollen das Merkmal `Mathematische Grundbildung` analysieren. Die Werte stehen in der zweiten Spalte der Matrix `PISA`. In S-PLUS gibt es eine Funktion `aov`, die folgendermaßen aufgerufen wird:

```
aov(formula, data = <<see below>>, projections = F,
    qr = F, contrasts = NULL, ...)
```

Man gibt, wie bei der Regressionsanalyse auf Seite 241 beschrieben wurde, die Beziehung als Formel ein.

Für das Beispiel heißt dies:

```
> e<-aov(PISA[,2]~gruppe)
```

Die ANOVA-Tabelle erhält man mit der Funktion `summary`:

```
> summary(e)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
   gruppe  2   9066.03 4533.013 2.237729 0.1254407
Residuals 28  56720.17 2025.720
```

In der ersten Spalte stehen die Quellen der Variation. Dabei steht `Residuals` für 'innerhalb der Gruppen'. In der zweiten spalte stehen die Freiheitsgrade. Es folgen in der dritten und vierten Spalte die Quadratsummen und die mittleren Quadratsummen. Neben dem Wert der Teststatistik des F-Tests gibt S-PLUS gibt noch den Wert der Überschreitungswahrscheinlichkeit aus. Sie beträgt 0.1254407. Wir lehnen also zum Signifikanzniveau $\alpha = 0.05$ die Nullhypothese (11.1) auf Seite 328 nicht ab.

Den Normal-Quantil-Plot in Abbildung 11.3 auf Seite 337 erhält man durch

```
> qqnorm(resid(e))
> qqline(resid(e))
```

Wir wollen die Gruppen 1 und 2 mit dem t-Test vergleichen. Wir erzeugen zunächst die Vektoren `gruppe1` und `gruppe2` mit den Werten in den beiden Gruppen:

```
> gruppe1<-PISA[gruppe=="w",2]
> gruppe1
 FIN   J  FL   L   A   S  CH  CZ
 536 557 514 446 515 510 529 498
> gruppe2<-PISA[gruppe=="m",2]
> gruppe2
 AUS   B  BR  DK   D   F  IS CDN ROK  NZ   N   P USA
 533 520 334 514 490 517 514 533 547 537 499 454 493
```

Mit der Funktion `t.test` kann man den t-Test durchführen. Der Aufruf

```
> t.test(gruppe1,gruppe2)
```

liefert folgendes Ergebnis:

```
        Standard Two-Sample t-Test

data:  gruppe1 and gruppe2
t = 0.6596, df = 19, p-value = 0.5174
alternative hypothesis:
true difference in means is not equal to 0
```

```
95 percent confidence interval:
 -31.02794  59.58563
sample estimates:
 mean of x mean of y
   513.125  498.8462
```

Der Wert der Teststatistik t in Gleichung (11.11) auf Seite 335 beträgt 0.6596, die Anzahl der Freiheitsgrade 19 und die Überschreitungswahrscheinlichkeit 0.5174. Somit wird die Nullhypothese (11.10) auf Seite 335 zum Signifikanzniveau 0.05 nicht abgelehnt.

Für den Kruskal-Wallis-Test gibt es die Funktion `kruskal.test`, die folgendermaßen aufgerufen wird:

```
kruskal.test(y, groups)
```

Die Daten stehen im Vektor `y`. Die i-te Komponente des Vektors `groups` gibt an, zu welcher Gruppe die i-te Beobachtung gehört. Wir geben also ein

```
> kruskal.test(PISA[,2],gruppe)
```

und erhalten folgendes Ergebnis:

```
    Kruskal-Wallis rank sum test

data:  PISA[, 2] and gruppe
Kruskal-Wallis chi-square = 7.0631, df = 2,
p-value = 0.0293
alternative hypothesis: two.sided
```

`S-PLUS` berücksichtigt das Vorhandensein von Bindungen und bestimmt die Teststatistik H^* in Gleichung (11.14) auf Seite 339. Die Überschreitungswahrscheinlichkeit beträgt 0.0293. Somit wird die Nullhypothese (11.12) auf Seite 337 zum Signifikanzniveau $\alpha = 0.05$ abgelehnt.

Um die Gruppen 1 und 2 mit dem Wilcoxon-Test zu vergleichen, ruft man die Funktion `wilcox.test` folgendermaßen auf:

```
> wilcox.test(gruppe1,gruppe2)
```

Man erhält folgendes Ergebnis:

```
    Wilcoxon rank-sum test

data:  gruppe1 and gruppe2
rank-sum normal statistic with correction Z = 0.2539,
p-value = 0.7996
alternative hypothesis: true mu is not equal to 0

Warning messages:
 cannot compute exact p-value with ties in:
 wil.rank.sum(x, y, alternative, exact, correct)
```

S-PLUS berücksichtigt die Bindungen und arbeitet mit der Stetigkeitskorrektur. Somit wird die Teststatistik in Gleichung (11.18) auf Seite 341 bestimmt. Die Überschreitungswahrscheinlichkeit beträgt 0.7996. Somit wird die Nullhypothese (11.15) auf Seite 340 zum Signifikanzniveau $\alpha = 0.05$ nicht abgelehnt.

Für die multivariate einfaktorielle Varianzanalyse gibt es in S-PLUS die Funktion manova. Sie wird aufgerufen durch

```
manova(formula, data=<<see below>>, qr=F,
       contrasts=NULL, ...)
```

Wir geben also ein

```
> e<-manova(PISA~gruppe)
```

Wilks' Λ erhalten wir durch

```
> summary(e,test="wilks")
          Df Wilks Lambda approx. F num df  den df  P-value
   gruppe 2  0.70094      1.68501        6      52  0.14329
Residuals 28
```

S-PLUS gibt den transformierten Wert in Gleichung (11.23) aus, der einer F-Verteilung folgt. Die Überschreitungswahrscheinlichkeit beträgt 0.14329. Somit wird die Nullhypothese (11.19) auf Seite 343 zum Signifikanzniveau $\alpha = 0.05$ nicht abgelehnt.

11.5 Ergänzungen und weiterführende Literatur

Wir haben uns bei der einfaktoriellen Varianzanalyse auf die klassischen parametrischen und nichtparametrischen Verfahren beschränkt. Weitere nichtparametrische Tests für die univariate einfaktorielle Varianzanalyse sind bei Büning & Trenkler (1994) zu finden. Diese beschreiben auch Tests zur Überprüfung der Annahme identischer Varianzen. Verfahren der univariaten und multivariaten mehrfaktoriellen Varianzanalyse sind bei Fahrmeir et al. (1996), Mardia et al. (1979) und Johnson & Wichern (1998) zu finden.

Wird die Hypothese abgelehnt, dass alle I Erwartungswerte identisch sind, so stellt sich die Frage, welche Gruppen sich unterscheiden. Wie man hierbei vorzugehen hat, wird bei Miller (1981) ausführlich dargestellt.

11.6 Übungen

Exercise 28. Betrachten Sie das Merkmal Lesekompetenz in Tabelle 1.1 auf Seite 4 und die Gruppen in Tabelle 11.1 auf Seite 328.

1. Führen Sie eine univariate einfaktorielle Varianzanalyse durch.

2. Führen Sie den Kruskal-Wallis-Test durch.
3. Welchen Test halten Sie für besser geeignet?

Exercise 29. Betrachten Sie die Merkmale `Ermitteln von Informationen`, `Textbezogenes Interpretieren` und `Reflektieren und Bewerten` in Tabelle 2.12 auf Seite 69 und die Gruppen in Tabelle 11.1 auf Seite 328.

1. Führen Sie eine multivariate einfaktorielle Varianzanalyse durch.
2. Betrachten Sie jedes einzelne Merkmal.
   a) Führen Sie eine univariate einfaktorielle Varianzanalyse durch.
   b) Führen Sie den Kruskal-Wallis-Test durch.

Exercise 30. In der PISA-Studie wurde die durchschnittliche Klassengröße in den einzelnen Ländern bestimmt. (vgl. Deutsches PISA-Konsortium (Hrsg.) (2001), S.422). Wir bilden drei Gruppen. Die erste Gruppe umfasst die Länder, bei denen in einer Klasse weniger als 22 Kinder unterrichtet werden. Das sind die folgenden Länder:

```
B DK FIN IS LV FL L S CH.
```

Die zweite Gruppe bilden die Länder, bei denen die durchschnittliche Klassengröße mindestens 22 aber weniger als 25 beträgt:

```
AUS D GR GB IRL I N A P RUS E CZ USA.
```

Die letzte Gruppe besteht aus den Ländern mit einer Klassengröße von mindestens 25:

```
BR F J CDN ROK MEX NZ PL H.
```

Betrachten Sie das Merkmal `Lesekompetenz` in Tabelle 1.1 auf Seite 4.

1. Führen Sie eine univariate einfaktorielle Varianzanalyse durch.
2. Führen Sie den Kruskal-Wallis-Test durch.
3. Welchen Test halten Sie für besser geeignet?

12 Diskriminanzanalyse

12.1 Problemstellung und theoretische Grundlagen

In diesem Kapitel gehen wir wie bei der Varianzanalyse davon aus, dass die Gruppen bekannt sind. Im Gegensatz zur Varianzanalyse ist aber nicht bekannt, zu welcher Gruppe ein Objekt gehört. Gesucht ist eine Entscheidungsregel, die es erlaubt, ein Objekt einer der Gruppen zuzuordnen. Man spricht in diesem Fall von *Diskriminanzanalyse*. Ein klassisches Beispiel ist die Einschätzung der Kreditwürdigkeit eines Kunden. Bei der Vergabe des Kredites ist nicht bekannt, ob der Kunde die Verpflichtungen einhalten wird. Man kennt aber eine Reihe von Merkmalen wie das Alter, das Einkommen und das Vermögen. Auf Basis dieser Informationen ordnet man die Person entweder der Gruppe der Kunden zu, die kreditwürdig sind, oder der Gruppe der Kunden, die nicht kreditwürdig sind. Auch Ärzte klassifizieren Patienten anhand einer Reihe von Symptomen als krank oder gesund.

Wir wollen uns damit beschäftigen, wie man datengestützt eine Entscheidungsregel finden kann, die ein Objekt mit dem p-dimensionalen Merkmalsvektor $\mathbf{x}$ genau einer der Gruppen zuordnet. Dabei werden wir ausschließlich den Fall betrachten, dass ein Objekt einer von zwei Gruppen zugeordnet werden soll.

Example 45. Im Beispiel 2 auf Seite 3 ist das Ergebnis eines Tests zu finden, der vor einem Brückenkurs in Mathematik im Wintersemester 1988/1989 am Fachbereich Wirtschaftswissenschaft der FU Berlin durchgeführt wurde. Bei dem Test mussten 26 Aufgaben bearbeitet werden. Wir kodieren `w` und `j` mit 1 und `m` und `n` mit 0. Mit Hilfe des Merkmals `Punkte` bilden wir zwei Gruppen von Studenten. Die Gruppe 1 besteht aus den Studenten, die mindestens 14 Punkte erreicht haben und damit den Test bestanden haben. In Gruppe 2 sind die Studenten, die den Test nicht bestanden haben. In Tabelle 12.1 sind die Daten zu finden.

Table 12.1. Ergebnisse von Studienanfängern bei einem Mathematik-Test

Geschlecht	MatheLK	MatheNote	Abitur88	Gruppe
0	0	3	0	2
0	0	4	0	2
0	0	4	0	2
0	0	4	0	2
0	0	3	0	2
1	0	3	0	2
1	0	4	1	2
1	0	3	1	2
1	0	4	1	1
0	1	3	0	1
0	1	3	0	1
0	1	2	0	1
0	1	3	0	2
1	1	3	0	1
1	1	2	0	1
1	1	2	0	1
0	1	1	1	1
1	1	2	1	2
1	1	2	1	2
1	1	4	1	1

Wir gehen im Folgenden davon aus, dass wir nur an diesen Studenten interessiert sind. Wir fassen sie also als eine Grundgesamtheit auf. In Gruppe 1 sind 9 Studierende und in Gruppe 2 sind 11. Wir wollen nun einen Studierenden der Gruppe zuordnen, zu der er gehört, ohne zu wissen, um welche Gruppe es sich handelt. Wir gehen zunächst davon aus, dass nur eines der Merkmale bekannt ist. Stellen wir uns vor, die ausgewählte Person ist weiblich, das Merkmal `Geschlecht` nimmt also den Wert 1 an. Welcher der beiden Gruppen sollen wir sie zuordnen? Um diese Frage zu beantworten, schauen wir uns die Kontingenztabelle der Merkmale `Geschlecht` und `Gruppe` an, die in Tabelle 12.2 zu finden ist.

Table 12.2. Kontingenztabelle der Merkmale Geschlecht und Gruppe

	`Gruppe` 1	2
`Geschlecht`		
`0`	4	6
`1`	5	5

Wir können die Information in Tabelle 12.2 auf zwei Arten zur Beantwortung der Frage benutzen. Wir können uns zum einen die Verteilung des

Merkmals `Geschlecht` in den beiden Gruppen anschauen. Wir können aber auch die Verteilung des Merkmals `Gruppe` bei den Frauen und bei den Männern betrachten. Auf den ersten Blick sieht es so aus, als ob nur die zweite Vorgehensweise sinnvoll ist. Wir kennen die Ausprägung des Merkmals `Geschlecht` und fragen nach dem Merkmal `Gruppe`. In der Praxis ist man bei der Datenerhebung aber mit der ersten Situation konfrontiert. Es werden in der Regel zuerst die beiden Gruppen gebildet und dann in diesen die Merkmale bestimmt, auf deren Basis die Entscheidungsregel angegeben werden soll. Schauen wir uns also deshalb zunächst die erste Situation an. In Tabelle 12.3 ist die Verteilung des Merkmals `Geschlecht` in den beiden Gruppen zu finden.

Table 12.3. Verteilung des Merkmales Geschlecht in den Gruppen 1 und 2

	Gruppe	1	2
`Geschlecht`			
`0`		0.44	0.55
`1`		0.56	0.45

Liegt die Gruppe 1 vor, so beträgt die Wahrscheinlichkeit, eine Person auszuwählen, die weiblich ist, $5/9 = 0.56$. Liegt hingegen die Gruppe 2 vor, so beträgt die Wahrscheinlichkeit, eine Person auszuwählen, die weiblich ist, $5/11 = 0.45$. Es ist also wahrscheinlicher, aus der Gruppe 1 eine weibliche Person auszuwählen als aus der Gruppe 2. Wurde also eine weibliche Person ausgewählt, so ist es plausibler, dass sie aus der Gruppe 1 kommt. Wir entscheiden uns damit für die Gruppe, bei der die Merkmalsausprägung `weiblich` wahrscheinlicher ist. □

Die Entscheidung im Beispiel 45 beruht auf dem *Likelihood-Prinzip*. Wir nennen sie deshalb die *Maximum-Likelihood-Entscheidungsregel*. Schauen wir uns diese formal an.

Eine Population bestehe aus den Gruppen 1 und 2. Anhand der Ausprägung $\mathbf{x}$ der p-dimensionalen Zufallsvariablen $\mathbf{X}$ soll ein Objekt einer der beiden Gruppen zugeordnet werden. Wir gehen im Folgenden davon aus, dass die p-dimensionale Zufallsvariable $\mathbf{X}$ die Wahrscheinlichkeits- bzw. Dichtefunktion $f_i(\mathbf{x})$ besitzt, wenn $\mathbf{X}$ zur i-ten Gruppe gehört, $i = 1, 2$. hmcounterend. (fortgesetzt)

Example 45. Sei X die Anzahl der Frauen in einer Stichprobe vom Umfang 1. Die Zufallsvariable X kann also die Werte 0 und 1 annehmen. Die Wahrscheinlichkeitsverteilung von X hängt davon ab, aus welcher Gruppe die Person kommt. Sei

$$f_i(x) = P(X = x | Person\ kommt\ aus\ Gruppe\ i).$$

Es gilt

$$f_1(0) = \frac{4}{9} = 0.44, \quad f_1(1) = \frac{5}{9} = 0.56$$

und

$$f_2(0) = \frac{6}{11} = 0.55, \quad f_2(1) = \frac{5}{11} = 0.45.$$

□

Definition 21. *Ein Objekt mit Merkmalsausprägung* $\mathbf{x}$ *wird nach der Maximum-Likelihood-Entscheidungsregel der Gruppe 1 zugeordnet, wenn gilt*

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1. \tag{12.1}$$

Es wird der Gruppe 2 zugeordnet, wenn gilt

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1. \tag{12.2}$$

Gilt

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = 1, \tag{12.3}$$

so kann man es willkürlich einer der beiden Gruppen zuordnen.

hmcounterend. (fortgesetzt)

Example 45. Wählt man also das Geschlecht als Entscheidungsvariable, so ordnet man eine Person der Gruppe 1 zu, wenn sie weiblich ist. Man ordnet sie der Gruppe 2 zu, wenn sie männlich ist. □

Die Entscheidungsregel ist fehlerbehaftet. Es gibt zwei Fehlentscheidungen: Man kann ein Objekt der Gruppe 1 zuordnen, obwohl es zur Gruppe 2 gehört, und man kann ein Objekt der Gruppe 2 zuordnen, obwohl es zur Gruppe 1 gehört. Die Wahrscheinlichkeiten dieser Fehlentscheidungen heißen *individuelle Fehlerraten* oder *Verwechslungswahrscheinlichkeiten*. Die Summe der beiden individuellen Fehlerraten nennt man *Fehlerrate*. hmcounterend. (fortgesetzt)

Example 45. Wir ordnen eine Person der Gruppe 2 zu, wenn sie männlich ist. Wird also ein Mann aus Gruppe 1 beobachtet, so ordnen wir diesen fälschlicherweise Gruppe 2 zu. Die Wahrscheinlichkeit, eine Person aus Gruppe 1 irrtümlich der Gruppe 2 zuzuordnen, beträgt also $4/9 = 0.44$. Entsprechend beträgt die Wahrscheinlichkeit, eine Person aus Gruppe 2 irrtümlich der Gruppe 1 zuzuordnen, $5/11 = 0.45$. Die individuellen Fehlerraten betragen also 0.44 und 0.45. Die Fehlerrate ist gleich 0.89, also sehr hoch. Das Merkmal `Geschlecht` diskriminiert also sehr schlecht zwischen den beiden Gruppen. Wählt man hingegen als Kriterium, ob jemand den Leistungskurs

Table 12.4. Kontingenztabelle der Merkmale MatheLK und Gruppe

	Gruppe 1	2
MatheLK		
0	1	8
1	8	3

Mathematik besucht hat oder nicht, kann man viel besser zwischen den beiden Gruppen diskriminieren. In Tabelle 12.4 ist die Kontingenztabelle der Merkmale `MatheLK` und `Gruppe` zu finden.

Tabelle 12.5 enthält die Verteilung des Merkmals `MatheLK` in den beiden Gruppen.

Table 12.5. Verteilung des Merkmals MatheLK in der Gruppe der Teilnehmer, die den Test bestanden haben, und in der Gruppe der Teilnehmer, die den Test nicht bestanden haben

	Gruppe	1	2
MatheLK			
0		0.11	0.73
1		0.89	0.27

Aufgrund der Maximum-Likelihood-Entscheidungsregel ordnen wir einen Studierenden der Gruppe 1 zu, wenn er den Mathematik-Leistungskurs besucht hat. Hat er den Mathematik-Leistungskurs hingegen nicht besucht, so ordnen wir die Person der Gruppe 2 zu. Unter den Personen, die den Test bestanden haben, beträgt der Anteil derjenigen, die keinen Leistungskurs Mathematik besucht haben, 0.11. Unter den Personen, die den Test nicht bestanden haben, beträgt der Anteil derjenigen, die den Leistungskurs Mathematik besucht haben, 0.27. Somit betragen die individuellen Fehlerraten 0.11 und 0.27. Die Fehlerrate beträgt also 0.38. Wir sehen, dass diese Fehlerrate beträchtlich niedriger als beim Merkmal `Geschlecht` ist. □

Wir haben die Maximum-Likelihood-Entscheidungsregel für p Merkmale definiert, im Beispiel aber nur ein Merkmal berücksichtigt. Schauen wir uns exemplarisch zwei Merkmale an. hmcounterend. (fortgesetzt)

Example 45. Wir betrachten die Merkmale `Geschlecht` und `MatheLK` gleichzeitig. Bei jedem Studierenden beobachten wir also ein Merkmalspaar (x_1, x_2), wobei x_1 das Merkmal `Geschlecht` und x_2 das Merkmal `MatheLK` ist. In Tabelle 12.6 sind alle Merkmalsausprägungen mit den absoluten Häufigkeiten in den beiden Gruppen zu finden. Tabelle 12.7 gibt die Wahrscheinlichkeitsverteilung der beiden Merkmale in den beiden Gruppen wieder.

Table 12.6. Absolute Häufigkeiten der Merkmale Geschlecht und MatheLK in den beiden Gruppen

(x_1, x_2)	Gruppe 1	2
$(0,0)$	0	5
$(0,1)$	4	1
$(1,0)$	1	3
$(1,1)$	4	2

Table 12.7. Wahrscheinlichkeitsverteilung der Merkmale Geschlecht und MatheLK in den beiden Gruppen

(x_1, x_2)	Gruppe 1	2
$(0,0)$	0.00	0.45
$(0,1)$	0.44	0.09
$(1,0)$	0.11	0.27
$(1,1)$	0.44	0.18

Auf Grund der Maximum-Likelihood-Entscheidungsregel ordnen wir einen Studierenden der Gruppe 1 zu, wenn er die Merkmalsausprägungen $(0,1)$ oder $(1,1)$ besitzt. Wir ordnen ihn der Gruppe 2 zu, wenn er die Merkmalsausprägungen $(0,0)$ oder $(1,0)$ besitzt. Wir sehen, dass die Entscheidungsregel im Beispiel nur vom Merkmal `MatheLK` abhängt. Die Fehlerrate beträgt 0.89. □

Die Maximum-Likelihood-Entscheidungsregel berücksichtigt nicht, dass die beiden Populationen unterschiedlich groß sein können. Ist die eine Population größer als die andere, so sollte auch die Chance größer sein, dass wir ein Objekt der größeren Population zuordnen. Schauen wir uns an, wie wir die Größe der Populationen berücksichtigen können.

Wir betrachten eine Zufallsvariable Y, die den Wert 1 annimmt, wenn ein Objekt aus Gruppe 1 kommt, und den Wert 0 annimmt, wenn ein Objekt aus Gruppe 2 kommt. Sei $P(Y = 1) = \pi_1$ die Wahrscheinlichkeit, dass ein Objekt aus Population 1, und $P(Y = 0) = \pi_2$ die Wahrscheinlichkeit, dass ein Objekt aus Population 2 kommt. Man nennt π_1 und π_2 auch *a priori-Wahrscheinlichkeiten*. Um den Umfang der Populationen bei der Entscheidungsregel zu berücksichtigen, multiplizieren wir die Likelihood-Funktionen mit den Wahrscheinlichkeiten der Populationen. Je größer ein π_i, $i = 1, 2$ ist, umso größer wird auch die Chance, dass ein Objekt dieser Population zugeordnet wird. Wir erhalten die sogenannte *Bayes-Entscheidungsregel*.

Definition 22. *Ein Objekt mit Merkmalsausprägung* $\mathbf{x}$ *wird nach der Bayes-Entscheidungsregel der Gruppe 1 zugeordnet, wenn gilt*

$$\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x}). \tag{12.4}$$

Es wird der Gruppe 2 zugeordnet, wenn gilt

$$\pi_1 \, f_1(\mathbf{x}) < \pi_2 \, f_2(\mathbf{x}). \tag{12.5}$$

Gilt

$$\pi_1 \, f_1(\mathbf{x}) = \pi_2 \, f_2(\mathbf{x}), \tag{12.6}$$

so kann man es willkürlich einer der beiden Gruppen zuordnen.

hmcounterend. (fortgesetzt)

Example 45. Es gilt $\pi_1 = 0.45$ und $\pi_2 = 0.55$. Wählen wir das Merkmal `Geschlecht` als Entscheidungsvariable, so gilt

$$\begin{aligned}
\pi_1 \, f_1(0) &= 0.45 \cdot \frac{4}{9} = 0.2, \\
\pi_2 \, f_2(0) &= 0.55 \cdot \frac{6}{11} = 0.3, \\
\pi_1 \, f_1(1) &= 0.45 \cdot \frac{5}{9} = 0.25, \\
\pi_2 \, f_2(1) &= 0.55 \cdot \frac{5}{11} = 0.25.
\end{aligned}$$

Wir ordnen also nach der Bayes-Entscheidungsregel eine Person der Gruppe 2 zu, wenn sie männlich ist. Ist sie weiblich, können wir sie willkürlich einer der beiden Gruppen zuordnen. Die Bayes-Entscheidungsregel kommt also im Beispiel zu einer anderen Entscheidung als die Maximum-Likelihood-Entscheidungsregel. □

Wir können die Gleichungen (12.4), (12.5) und (12.6) so umformen, dass man sie einfacher mit der Maximum-Likelihood-Entscheidungsregel vergleichen kann. Hierzu schauen wir uns nur (12.4) an. Die beiden anderen Entscheidungen ändern sich analog.

Ein Objekt mit Merkmalsausprägung $\mathbf{x}$ wird nach der Bayes-Entscheidungsregel der Gruppe 1 zugeordnet, wenn gilt

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1}. \tag{12.7}$$

Wir sehen, dass sich bei der Bayes-Entscheidungsregel gegenüber der Maximum-Likelihood-Entscheidungsregel nur die Grenze ändert, auf der die Entscheidung basiert.

Die Bayes-Entscheidungsregel besitzt unter allen Entscheidungsregeln die kleinste Fehlerrate. Ein Beweis dieser Tatsache ist bei Fahrmeir et al. (1996) zu finden.

Man kann die Bayes-Entscheidungsregel in einer Form darstellen, durch die ihr Name verdeutlicht wird. Wir betrachten die *a posteriori-Wahrscheinlichkeiten* $P(Y = 1|\mathbf{x})$ und $P(Y = 0|\mathbf{x})$. Es liegt nahe, ein Objekt der Gruppe

1 zuzuordnen, wenn $P(Y = 1|\mathbf{x})$ größer ist als $P(Y = 0|\mathbf{x})$. Diese Vorschrift entspricht der Bayes-Entscheidungsregel. Aufgrund des Satzes von Bayes gilt

$$P(Y = 1|\mathbf{x}) = \frac{\pi_1 \, f_1(\mathbf{x})}{f(\mathbf{x})} \tag{12.8}$$

und

$$P(Y = 0|\mathbf{x}) = \frac{\pi_2 \, f_2(\mathbf{x})}{f(\mathbf{x})} \tag{12.9}$$

mit

$$f(\mathbf{x}) = \pi_1 \, f_1(\mathbf{x}) + \pi_2 \, f_2(\mathbf{x}).$$

Aus (12.8) und (12.9) folgt

$$\pi_1 \, f_1(\mathbf{x}) = P(Y = 1|\mathbf{x}) \, f(\mathbf{x}) \tag{12.10}$$

und

$$\pi_2 \, f_2(\mathbf{x}) = P(Y = 0|\mathbf{x}) \, f(\mathbf{x}). \tag{12.11}$$

Setzen wir (12.10) und (12.11) in (12.4) ein, so folgt:

Ein Objekt mit Merkmalsausprägung $\mathbf{x}$ wird nach der Bayes-Entscheidungsregel der Gruppe 1 zugeordnet, wenn gilt

$$P(Y = 1|\mathbf{x}) > P(Y = 0|\mathbf{x}). \tag{12.12}$$

hmcounterend. (fortgesetzt)

Example 45. Wählen wir das Merkmal `Geschlecht` als Entscheidungsvariable, so können wir die Bayes-Entscheidungsregel mit Tabelle 12.2 problemlos über (12.12) bestimmen. Es gilt

$$P(Y = 0|0) = 0.6,$$
$$P(Y = 1|0) = 0.4$$

und

$$P(Y = 0|1) = 0.5,$$
$$P(Y = 1|1) = 0.5.$$

□

Bei der Diskriminanzanalyse gibt es im Zweigruppenfall zwei Fehlentscheidungen. Man kann ein Objekt irrtümlich der Gruppe 1 oder irrtümlich der Gruppe 2 zuordnen. Bisher sind wir davon ausgegangen, dass diese beiden Fehlklassifikationen gleichgewichtig sind. Dies ist aber nicht immer der Fall. So ist es in der Regel sicherlich schlimmer, einen Kranken als gesund einzustufen als einen Gesunden als krank. Wir wollen also die Entscheidungsregel noch um Kosten erweitern.

Definition 23. *Seien $C(1|2)$ die Kosten, die entstehen, wenn man ein Objekt, das in Gruppe 2 gehört, irrtümlich in Gruppe 1 einstuft, und $C(2|1)$ die Kosten, die entstehen, wenn man ein Objekt, das in Gruppe 1 gehört, irrtümlich in Gruppe 2 einstuft. Ein Objekt mit Merkmalsausprägung $\mathbf{x}$ wird nach der kostenminimalen Entscheidungsregel der Gruppe 1 zugeordnet, wenn gilt*

$$\pi_1 \, C(2|1) \, f_1(\mathbf{x}) > \pi_2 \, C(1|2) \, f_2(\mathbf{x}). \tag{12.13}$$

Es wird der Gruppe 2 zugeordnet, wenn gilt

$$\pi_1 \, C(2|1) \, f_1(\mathbf{x}) < \pi_2 \, C(1|2) \, f_2(\mathbf{x}). \tag{12.14}$$

Gilt

$$\pi_1 \, C(2|1) \, f_1(\mathbf{x}) = \pi_2 \, C(1|2) \, f_2(\mathbf{x}), \tag{12.15}$$

so kann man es willkürlich einer der beiden Gruppen zuordnen.

Ein Beweis der Kostenoptimalität ist bei Krzanowski (2000), S.335-336 zu finden. Wir formen (12.13), (12.14) und (12.15) so um, dass wir sie einfacher mit der Maximum-Likelihood-Entscheidungsregel und der Bayes-Entscheidungsregel vergleichen können. Wir beschränken uns auch hier auf (12.13).

Ein Objekt mit Merkmalsausprägung $\mathbf{x}$ wird nach der kostenminimalen Entscheidungsregel der Gruppe 1 zugeordnet, wenn gilt

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2 \, C(1|2)}{\pi_1 \, C(2|1)}. \tag{12.16}$$

Wir sehen, dass die kostenminimale Entscheidungsregel mit der Bayes-Entscheidungsregel zusammenfällt, wenn $C(1|2) = C(2|1)$ gilt.

Wir verwenden bei der Formulierung der Entscheidungsregel die Maximum-Likelihood-Entscheidungsregel. Bei den beiden anderen Regeln muss man nur die rechte Seite der Ungleichung entsprechend modifizieren. Außerdem geben wir bei der Entscheidungsregel nur den Teil an, der beschreibt, wann man eine Beobachtung der ersten Gruppe zuordnen soll.

Bisher sind wir davon ausgegangen, dass die Verteilung der Grundgesamtheit bekannt ist. Ist sie nicht bekannt, so fassen wir die Beobachtungen als Stichprobe auf und schätzen die jeweiligen Parameter. So schätzen wir zum Beispiel die a priori-Wahrscheinlichkeiten über die Anteile der beiden Gruppen an der Stichprobe. Ein solcher Schätzer setzt natürlich voraus, dass eine Zufallsstichprobe aus der Grundgesamtheit vorliegt.

12.2 Diskriminanzanalyse bei normalverteilten Grundgesamtheiten

In der Einleitung dieses Kapitels haben wir die Grundprinzipien der Diskriminanzanalyse am Beispiel eines qualitativen Merkmals klargemacht. Oft werden quantitative Merkmale erhoben. In diesem Fall nimmt man an, dass das Merkmal aus einer normalverteilten Grundgesamtheit kommt.

12.2.1 Diskriminanzanalyse bei Normalverteilung mit bekannten Parametern

Schauen wir uns zunächst den univariaten Fall an. Wir gehen von einer Zufallsvariablen X aus, die in Gruppe 1 normalverteilt ist mit den Parametern μ_1 und σ_1^2, und die in Gruppe 2 normalverteilt ist mit den Parametern μ_2 und σ_2^2. Für $i = 1, 2$ gilt also

$$f_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\} \quad \text{für} \quad x \in I\!R\,.$$

Theorem 12 gibt die Entscheidungsregel bei univariater Normalverteilung an.

Theorem 12. *Die Zufallsvariable X sei in Gruppe 1 normalverteilt mit den Parametern μ_1 und σ_1^2 und in Gruppe 2 normalverteilt mit den Parametern μ_2 und σ_2^2. Dann lautet die Maximum-Likelihood-Entscheidungsregel:*
Ordne das Objekt mit Merkmalsausprägung x der Gruppe 1 zu, wenn gilt

$$x^2\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) - 2\,x\left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) + \left(\frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2}\right) > 2\,\ln\frac{\sigma_1}{\sigma_2}. \quad (12.17)$$

Beweis:
(12.1) auf Seite 354 ist äquivalent zu

$$\ln f_1(\mathbf{x}) - \ln f_2(\mathbf{x}) > 0.$$

Es gilt

$$\ln f_i(x) = -0.5\,\ln\sigma_i^2 - 0.5\,\ln 2\pi - \frac{(x-\mu_i)^2}{2\sigma_i^2}.$$

Hieraus folgt

$$\begin{aligned}\ln f_1(x) - \ln f_2(x) &= -0.5 \ln \sigma_1^2 - 0.5 \ln 2\pi - \frac{(x-\mu_1)^2}{2\sigma_1^2} \\ &\quad + 0.5 \ln \sigma_2^2 + 0.5 \ln 2\pi + \frac{(x-\mu_2)^2}{2\sigma_2^2} \\ &= 0.5 \ln \sigma_2^2 - 0.5 \ln \sigma_1^2 \\ &\quad + \frac{x^2 - 2x\mu_2 + \mu_2^2}{2\sigma_2^2} - \frac{x^2 - 2x\mu_1 + \mu_1^2}{2\sigma_1^2} \\ &= 0.5\, x^2 \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) - x \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) \\ &\quad + 0.5 \left(\frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2}\right) - \ln \frac{\sigma_1}{\sigma_2}.\end{aligned}$$

Hieraus folgt (12.17).

Man sieht, dass die Entscheidungsregel auf einer in x quadratischen Funktion basiert.

Example 46. In Gruppe 1 liege Normalverteilung mit $\mu_1 = 3$ und $\sigma_1 = 1$ und in Gruppe 2 Normalverteilung mit $\mu_2 = 4$ und $\sigma_2 = 2$ vor. Die Entscheidungsregel lautet also:
Ordne ein Objekt der Gruppe 1 zu, wenn gilt

$$1.152 \leq x \leq 4.181.$$

Ordne ein Objekt der Gruppe 2 zu, wenn gilt

$$x < 1.152$$

oder

$$x > 4.181.$$

Abbildung 12.1 zeigt die Dichtefunktion $f_1(x)$ einer Normalverteilung mit $\mu = 3$ und $\sigma = 1$ (durchgezogene Linie) und die Dichtefunktion $f_2(x)$ einer Normalverteilung mit $\mu = 4$ und $\sigma = 2$ (gestrichelte Linie). Der Wertebereich von x, bei dem wir ein Objekt der Gruppe 1 zuordnen, ist fett dargestellt. Wir sehen, dass wir im Fall ungleicher Varianzen drei Bereiche erhalten.

□

Ein wichtiger Spezialfall liegt vor, wenn die beiden Varianzen gleich sind. Wir betrachten den Fall $\mu_1 > \mu_2$. In diesem Fall erhalten wir folgende Entscheidungsregel:

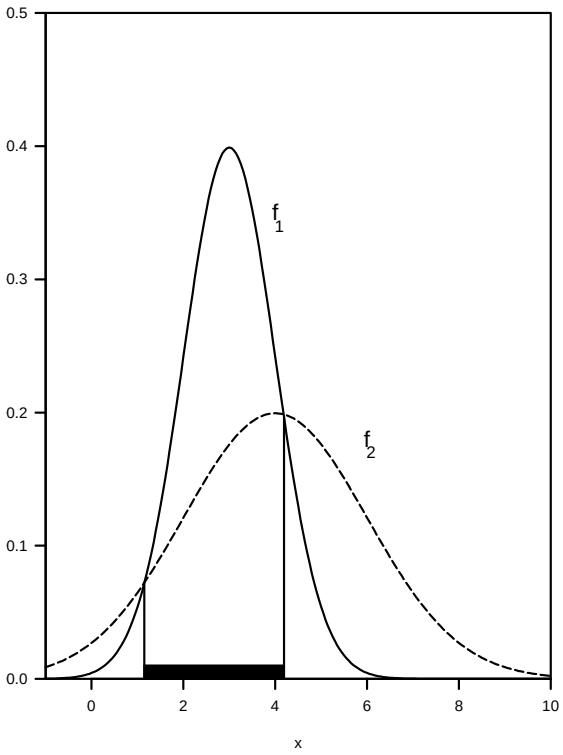


Fig. 12.1. Veranschaulichung der Maximum-Likelihood-Entscheidungsregel bei univariater Normalverteilung mit ungleichen Varianzen

Ordne das Objekt mit Merkmalsausprägung x der Gruppe 1 zu, wenn gilt

$$x > \frac{\mu_1 + \mu_2}{2}.$$

Setzen wir nämlich in (12.17) $\sigma_1 = \sigma_2 = \sigma$, so ergibt sich

$$x^2 \left(\frac{1}{\sigma^2} - \frac{1}{\sigma^2}\right) - 2\,x \left(\frac{\mu_2}{\sigma^2} - \frac{\mu_1}{\sigma^2}\right) + \left(\frac{\mu_2^2}{\sigma^2} - \frac{\mu_1^2}{\sigma^2}\right) > 2 \ln \frac{\sigma}{\sigma}.$$

Dies ist äquivalent zu

$$-2\,x \left(\frac{\mu_2}{\sigma^2} - \frac{\mu_1}{\sigma^2}\right) + \left(\frac{\mu_2^2}{\sigma^2} - \frac{\mu_1^2}{\sigma^2}\right) > 0.$$

Multiplizieren wir beide Seiten von (12.18) mit σ^2, so erhalten wir

$$-2\,x\,(\mu_2 - \mu_1) + (\mu_2^2 - \mu_1^2) > 0.$$

Wegen

$$\mu_1^2 - \mu_2^2 = (\mu_1 - \mu_2)\,(\mu_1 + \mu_2)$$

und

$$\mu_1 > \mu_2$$

folgt

$$x > \frac{\mu_1 + \mu_2}{2}.$$

Im Fall gleicher Varianzen erhält man also eine in x lineare Entscheidungsregel.

Example 47. In Gruppe 1 liege Normalverteilung mit $\mu = 5$ und $\sigma = 1$ und in Gruppe 2 Normalverteilung mit $\mu = 3$ und $\sigma = 1$ vor. Wir erhalten also folgende Entscheidungsregel:
Ordne das Objekt mit Merkmalsausprägung x der Gruppe 1 zu, wenn gilt

$$x > 4.$$

Ordne das Objekt mit Merkmalsausprägung x der Gruppe 2 zu, wenn gilt

$$x < 4.$$

Ein Objekt mit Merkmalsausprägung 4 ordnen wir zufällig einer der beiden Gruppen zu. Abbildung 12.2 zeigt die Dichtefunktion $f_1(x)$ einer Normalverteilung mit $\mu = 5$ und $\sigma = 1$ (durchgezogene Linie) und die Dichtefunktion $f_2(x)$ einer Normalverteilung mit $\mu = 3$ und $\sigma = 1$ (gestrichelte Linie). Der Wertebereich von x, bei dem wir ein Objekt der Gruppe 1 zuordnen, ist fett dargestellt. Wir sehen, dass wir im Fall gleicher Varianzen zwei Bereiche erhalten.

□

Betrachten wir nun den multivariaten Fall. Wir gehen davon aus, dass die p-dimensionale Zufallsvariable $\mathbf{X}$ in der i-ten Gruppe multivariat normalverteilt ist mit Parametern $\boldsymbol{\mu}_i$ und $\boldsymbol{\Sigma}_i$, $i = 1, 2$. In der i-ten Gruppe liegt also folgende Dichtefunktion vor:

$$f_i(\mathbf{x}) = (2\,\pi)^{-p/2}\,|\boldsymbol{\Sigma}_i|^{-0.5}\,\exp\left\{-0.5\,(\mathbf{x} - \boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right\}.$$

Theorem 13. *Die p-dimensionale Zufallsvariable $\mathbf{X}$ sei in der i-ten Gruppe multivariat normalverteilt mit Parametern $\boldsymbol{\mu}_i$ und $\boldsymbol{\Sigma}_i$, $i = 1, 2$. Dann lautet die Maximum-Likelihood-Entscheidungsregel:*
Ordne das Objekt mit Merkmalsausprägung x der Gruppe 1 zu, wenn gilt

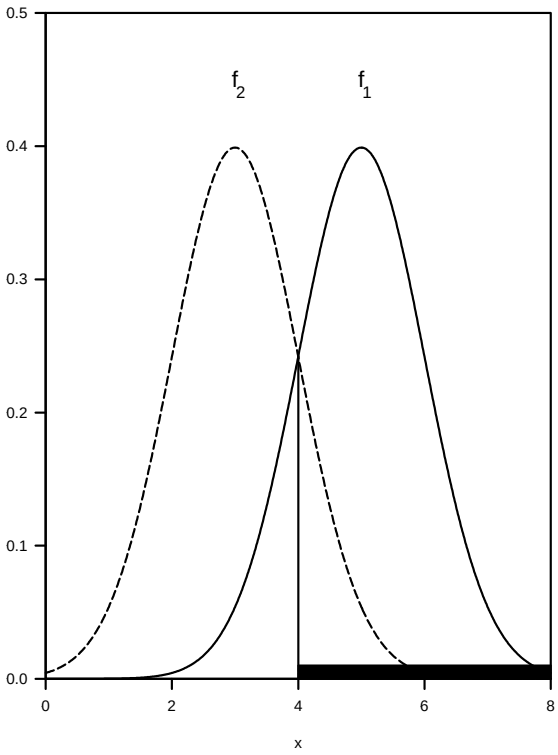


Fig. 12.2. Veranschaulichung der Maximum-Likelihood-Entscheidungsregel bei univariater Normalverteilung mit gleichen Varianzen

$$-0.5\,\mathbf{x}'\,(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\,\mathbf{x}+(\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1})\,\mathbf{x}>k. \qquad (12.18)$$

Dabei ist

$$k=0.5\,\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}+0.5\,(\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1-\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2).$$

Beweis:
Gleichung (12.1) auf Seite 354 ist äquivalent zu

$$\ln f_1(\mathbf{x})-\ln f_2(\mathbf{x})>0.$$

Es gilt

$$\ln f_i(\mathbf{x})=-\ln(2\,\pi)^{p/2}-0.5\,\ln|\boldsymbol{\Sigma_i}|-0.5\,(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\,.$$

Hieraus folgt:

$$\begin{aligned}
\ln f_1(\mathbf{x}) - \ln f_2(\mathbf{x}) &= -0.5 \ln|\boldsymbol{\Sigma}_1| - 0.5\,(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \\
&\quad + 0.5 \ln|\boldsymbol{\Sigma}_2| + 0.5\,(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \\
&= -0.5\,(\mathbf{x}'\,\boldsymbol{\Sigma}_1^{-1}\,\mathbf{x} - \mathbf{x}'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\mathbf{x} \\
&\quad + \boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1) + 0.5\,(\mathbf{x}'\,\boldsymbol{\Sigma}_2^{-1}\,\mathbf{x} - \mathbf{x}'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2 \\
&\quad - \boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\mathbf{x} + \boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2) - 0.5 \ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \\
&= -0.5\,(\mathbf{x}'\,\boldsymbol{\Sigma}_1^{-1}\,\mathbf{x} - 2\,\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\mathbf{x} + \boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1) \\
&\quad + 0.5\,(\mathbf{x}'\,\boldsymbol{\Sigma}_2^{-1}\,\mathbf{x} - 2\,\boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\mathbf{x} + \boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2) \\
&\quad - 0.5 \ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \\
&= -0.5\,\mathbf{x}'\,(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\,\mathbf{x} + (\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1})\,\mathbf{x} \\
&\quad - 0.5 \ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - 0.5\,(\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2).
\end{aligned}$$

Wir sehen, dass die Entscheidungsregel wie im univariaten Fall quadratisch in $\mathbf{x}$ ist. Man spricht auch von *quadratischer Diskriminanzanalyse*. Sind die beiden Varianz-Kovarianz-Matrizen identisch, gilt also $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, so ordnen wir eine Beobachtung $\mathbf{x}$ der ersten Gruppe zu, falls gilt

$$(\boldsymbol{\mu}_1' - \boldsymbol{\mu}_2')\,\boldsymbol{\Sigma}^{-1}\,\mathbf{x} - 0.5\,(\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}_2) > 0. \tag{12.19}$$

Dies ergibt sich sofort aus (12.18), wenn man $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ setzt.

Wir sehen, dass die Entscheidungsregel linear in $\mathbf{x}$ ist. Man spricht auch von der *linearen Diskriminanzanalyse*. Man kann diese Entscheidungsregel auch folgendermaßen darstellen:
Ordne eine Beobachtung $\mathbf{x}$ der ersten Gruppe zu, falls gilt

$$\mathbf{a}'\,\mathbf{x} > 0.5\,\mathbf{a}'\,(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \tag{12.20}$$

Dabei ist

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{12.21}$$

Es gilt nämlich

$$\boldsymbol{\mu}_1'\,\boldsymbol{\Sigma}_1^{-1}\,\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\,\boldsymbol{\Sigma}_2^{-1}\,\boldsymbol{\mu}_2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \tag{12.22}$$

Dies sieht man folgendermaßen:

$$\begin{aligned}
\boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_1^{-1} \, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \, \boldsymbol{\Sigma}_2^{-1} \, \boldsymbol{\mu}_2 &= \boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_1^{-1} \, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \, \boldsymbol{\Sigma}_2^{-1} \, \boldsymbol{\mu}_2 \\
&\quad + \boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_1^{-1} \, \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_2^{-1} \, \boldsymbol{\mu}_2 \\
&= \boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_1^{-1} \, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \, \boldsymbol{\Sigma}_2^{-1} \, \boldsymbol{\mu}_2 \\
&\quad + \boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_1^{-1} \, \boldsymbol{\mu}_2 - \boldsymbol{\mu}_2' \, \boldsymbol{\Sigma}_2^{-1} \, \boldsymbol{\mu}_1 \\
&= \boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_1^{-1} \, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \boldsymbol{\mu}_2' \, \boldsymbol{\Sigma}_2^{-1} \, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \\
&= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).
\end{aligned}$$

Setzt man (12.22) in die linke Seite von (12.19) ein, so ergibt sich:

$$\begin{aligned}
&(\boldsymbol{\mu}_1' - \boldsymbol{\mu}_2') \, \boldsymbol{\Sigma}^{-1} \, \mathbf{x} - 0.5 \, (\boldsymbol{\mu}_1' \, \boldsymbol{\Sigma}_1^{-1} \, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \, \boldsymbol{\Sigma}_2^{-1} \, \boldsymbol{\mu}_2) = \\
&(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \boldsymbol{\Sigma}^{-1} \, \mathbf{x} - 0.5 \, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \\
&(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \boldsymbol{\Sigma}^{-1} \, \left(\mathbf{x} - 0.5 \, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right).
\end{aligned}$$

Setzt man

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{12.23}$$

so erhält man Beziehung (12.20).

12.2.2 Diskriminanzanalyse bei Normalverteilung mit unbekannten Parametern

Bisher sind wir davon ausgegangen, dass alle Parameter der zugrunde liegenden Normalverteilung bekannt sind. Dies ist in der Praxis meist nicht der Fall. Wir müssen die Parameter also schätzen. Hierbei unterstellen wir, dass die Varianz-Kovarianz-Matrizen in den beiden Gruppen identisch sind. Ausgangspunkt sind also im Folgenden eine Stichprobe $\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$ aus Gruppe 1 und eine Stichprobe $\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$ aus Gruppe 2.

Example 48. Im Beispiel 11 auf Seite 10 haben wir 20 Zweigstellen eines Kreditinstituts in Baden-Württemberg betrachtet. Die Filialen können in zwei Gruppen eingeteilt werden. Die Filialen der ersten Gruppe haben einen hohen Marktanteil und ein überdurchschnittliches Darlehens- und Kreditgeschäft. Es sind die ersten 14 Zweigstellen in Tabelle 1.12 auf Seite 11. Die restlichen 6 Filialen sind technisch gut ausgestattet, besitzen ein überdurchschnittliches Einlage- und Kreditgeschäft und eine hohe Mitarbeiterzahl. Sie bilden die zweite Gruppe. Wir wollen auf der Basis der Merkmale `Einwohnerzahl` und `Gesamtkosten` eine Entscheidungsregel angeben. Abbildung 12.3 zeigt das Streudiagramm der beiden Merkmale. Jede Beobachtung wird durch das Symbol ihrer Gruppe dargestellt.
Man kann die beiden Gruppen sehr gut erkennen. □

In (12.20) haben wir gesehen, dass wir eine Beobachtung der ersten Gruppe zuordnen, wenn gilt

$$\mathbf{a}'\,\mathbf{x} > 0.5\,\mathbf{a}'\,(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Dabei ist

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Diese Entscheidungsregel hängt von den Parametern $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ und $\boldsymbol{\Sigma}$ ab. Diese sind unbekannt. Es liegt nahe, sie zu schätzen. Wir schätzen die unbekannten Erwartungswerte durch die entsprechenden Mittelwerte und erhalten für $i = 1, 2$

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij}\,.$$

hmcounterend. (fortgesetzt)

Example 48. In Gruppe 1 gilt

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 3510.4 \\ 390.2 \end{pmatrix}.$$

In Gruppe 2 gilt

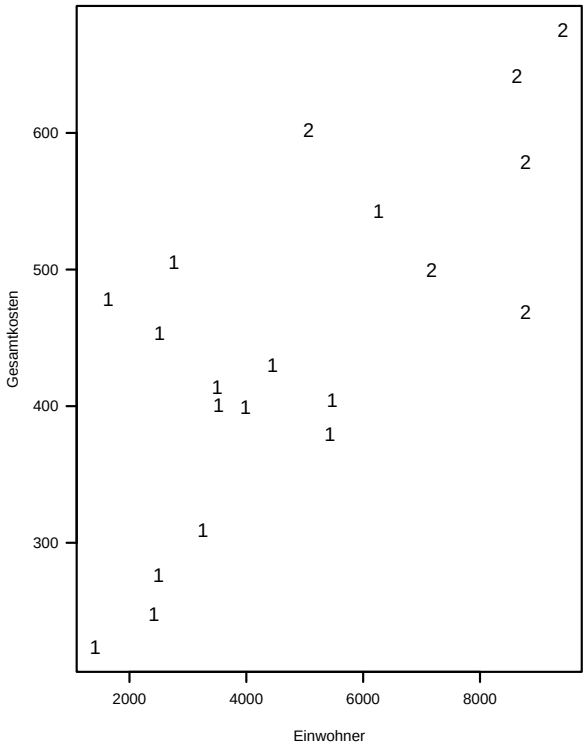


Fig. 12.3. Streudiagramm der Merkmale Einwohner und Gesamtkosten bei 20 Zweigstellen eines Kreditinstituts

$$\bar{\mathbf{x}}_2 = \begin{pmatrix} 7975.2 \\ 577.5 \end{pmatrix}.$$

Wir sehen, dass sich die Mittelwerte der beiden Gruppen stark unterscheiden. □

Die gemeinsame Varianz-Kovarianz-Matrix $\boldsymbol{\Sigma}$ schätzen wir durch die sogenannte *gepoolte Varianz-Kovarianz-Matrix*:

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1)\, \mathbf{S}_1 + (n_2 - 1)\, \mathbf{S}_2 \right). \tag{12.24}$$

Dabei gilt für $i = 1, 2$:

$$\mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)\, (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'.$$

hmcounterend. (fortgesetzt)

Example 48. In Gruppe 1 gilt

$$\mathbf{S}_1 = \begin{pmatrix} 2147306.26 & 61126.41 \\ 61126.41 & 9134.82 \end{pmatrix}.$$

In Gruppe 2 gilt

$$\mathbf{S}_2 = \begin{pmatrix} 2578808.97 & 17788.39 \\ 17788.39 & 6423.85 \end{pmatrix}.$$

Hieraus ergibt sich

$$\mathbf{S} = \begin{pmatrix} 2267168.12 & 49088.07 \\ 49088.07 & 8381.77 \end{pmatrix}.$$

□

Wir ordnen eine Beobachtung $\mathbf{x}$ der ersten Gruppe zu, falls gilt

$$\mathbf{a}'\,\mathbf{x} > 0.5\,\mathbf{a}'\,(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \tag{12.25}$$

mit

$$\mathbf{a} = \mathbf{S}^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{12.26}$$

hmcounterend. (fortgesetzt)

Example 48. Es gilt

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.00000051 & -0.00000296 \\ -0.00000296 & 0.00013663 \end{pmatrix}.$$

Mit

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \begin{pmatrix} -4464.8 \\ -187.2 \end{pmatrix}$$

gilt

$$\mathbf{a} = \begin{pmatrix} -0.00170 \\ -0.01237 \end{pmatrix}.$$

Wir klassifizieren eine Zweigstelle mit dem Merkmalsvektor

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

zur Gruppe 1, falls gilt

$$\begin{pmatrix} -0.0017 & -0.01237 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} > \begin{pmatrix} -0.0017 & -0.01237 \end{pmatrix} \begin{pmatrix} 5742.8 \\ 483.8 \end{pmatrix}.$$

Dies können wir noch vereinfachen zu

$$0.0017\,x_1 + 0.01237\,x_2 < 15.747\,. \tag{12.27}$$

Wir können diese Entscheidungsregel zusammen mit den Daten auf zweierlei Art visualisieren. Wir zeichnen in das Streudiagramm die Gerade, die sich auf Grund der Linearkombination ergibt. Dies wurde in Abbildung 12.4 gemacht. Wir sehen, dass die Zweigstellen duch die Gerade gut getrennt werden. Wir können aber auch für jede der 20 Zweigstellen den Wert abtragen, der sich auf Grund der Linearkombination ergibt. Dies zeigt das folgende Bild:

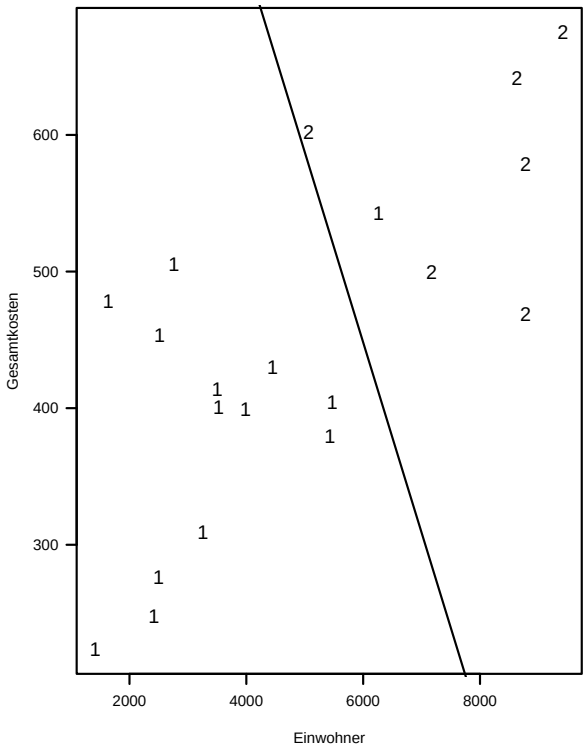


Fig. 12.4. Streudiagramm der Merkmale Einwohner und Gesamtkosten bei 20 Zweigstellen eines Kreditinstituts mit der Gerade, die die Gruppen trennt

□

12.3 Fishers lineare Diskriminanzanalyse

Die bisher betrachteten Verfahren beruhen auf der Annahme der Normalverteilung. Von Fisher wurde eine Vorgehensweise vorgeschlagen, die ohne diese Annahme auskommt. hmcounterend. (fortgesetzt)

Example 48. Wir wollen möglichst gut zwischen den beiden Arten von Zweigstellen unterscheiden. Die Annahmen der Normalverteilung und gleicher Varianz-Kovarianz-Matrizen liefern eine lineare Entscheidungsregel. Wodurch zeichnet sich diese Entscheidungsregel aus? Schauen wir uns dazu noch einmal an, wie sich die 20 Beobachtungen verteilen, wenn man die Linearkombina-

tion der beiden Merkmale bildet, wobei die Mittelwerte fett eingezeichnet sind:

Die beiden Gruppen sind nahezu perfekt getrennt. Dies zeigt sich dadurch, dass zum einen die Mittelwerte der beiden Gruppen weit voneinander entfernt sind und zum anderen die Streuung in den Gruppen klein ist. □

Fishers Ziel ist es, eine lineare Entscheidungsregel zu finden, bei der die Gruppen die eben beschriebenen Eigenschaften besitzen. Er geht aus von den Beobachtungen $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}$,$\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}$ und sucht eine Linearkombination

$$\mathbf{y}_{ij} = \mathbf{d}'\mathbf{x}_{ij}$$

der Beobachtungen, sodass die dadurch gewonnenen eindimensionalen Beobachtungen $y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}$ die Gruppenstruktur möglichst gut wiedergeben. Dies beinhaltet, dass die Streuung zwischen den Gruppen

$$(\bar{y}_1 - \bar{y}_2)^2 \tag{12.28}$$

möglichst groß ist. Außerdem sollte die Streuung innerhalb der Gruppen

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \tag{12.29}$$

möglichst klein sein. Dabei ist

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

für $i = 1, 2$.

Hat man den Vektor $\mathbf{d}$ gefunden, so bildet man für eine Beobachtung $\mathbf{x}$ den Wert

$$y = \mathbf{d}' \mathbf{x}$$

und ordnet die Beobachtung y der ersten Gruppe zu, wenn y näher an $\bar{y}_1$ liegt. Liegt y näher an $\bar{y}_2$, so ordnet man die Beobachtung der zweiten Gruppe zu. Wir ordnen eine Beobachtung y also der ersten Gruppe zu, falls gilt

$$|y - \bar{y}_1| < |y - \bar{y}_2|. \tag{12.30}$$

Wie findet man den Gewichtungsvektor $\mathbf{d}$? Da die Streuung zwischen den Gruppen möglichst groß und die Streuung innerhalb der Gruppen möglichst klein sein soll, bildet man den Quotienten aus (12.28) und (12.29):

$$F = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$$

und sucht den Vektor $\mathbf{d}$, für den F maximal wird. Mit

$$\mathbf{W} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$$

gilt

$$F = \frac{(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{\mathbf{d}' \mathbf{W} \mathbf{d}} .$$

Dies sieht man folgendermaßen:

$$F = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2} = \frac{(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{d}' \mathbf{x}_{ij} - \mathbf{d}' \bar{\mathbf{x}}_i)^2}$$

$$= \frac{(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{d}' \mathbf{x}_{ij} - \mathbf{d}' \bar{\mathbf{x}}_i) (\mathbf{d}' \mathbf{x}_{ij} - \mathbf{d}' \bar{\mathbf{x}}_i)}$$

$$= \frac{(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{\sum_{i=2}^{c} \sum_{j=1}^{n_i} (\mathbf{d}' \mathbf{x}_{ij} - \mathbf{d}' \bar{\mathbf{x}}_i) (\mathbf{d}' \mathbf{x}_{ij} - \mathbf{d}' \bar{\mathbf{x}}_i)'}$$

$$= \frac{(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{\sum_{i=2}^{c} \sum_{j=1}^{n_i} \mathbf{d}' (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \mathbf{d}}$$

$$= \frac{(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{\mathbf{d}' \left(\sum_{i=2}^{c} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i) (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' \right) \mathbf{d}}$$

$$= \frac{(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{\mathbf{d}' \mathbf{W} \mathbf{d}}.$$

Hierbei wurde folgende Beziehung berücksichtigt:

$$\bar{y}_i = \mathbf{d}' \bar{\mathbf{x}}_i.$$

Wir bilden die partielle Ableitung von F nach $\mathbf{d}$:

$$\frac{\partial}{\partial \mathbf{d}} F = \frac{2 (\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2) (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \mathbf{d}' \mathbf{W} \mathbf{d} - 2 \mathbf{W} \mathbf{d} (\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{(\mathbf{d}' \mathbf{W} \mathbf{d})^2}.$$

Die notwendigen Bedingungen für einen Extremwert lauten also

$$\frac{2 (\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2) (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \mathbf{d}' \mathbf{W} \mathbf{d} - 2 \mathbf{W} \mathbf{d} (\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)^2}{(\mathbf{d}' \mathbf{W} \mathbf{d})^2} = 0.$$

Wir multiplizieren diese Gleichung mit

$$\frac{(\mathbf{d}' \mathbf{W} \mathbf{d})^2}{2(\mathbf{d}' \bar{\mathbf{x}}_1 - \mathbf{d}' \bar{\mathbf{x}}_2)}$$

und erhalten

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\, \mathbf{d}'\mathbf{W}\mathbf{d} - \mathbf{W}\mathbf{d}(\mathbf{d}'\bar{\mathbf{x}}_1 - \mathbf{d}'\bar{\mathbf{x}}_2) = 0.$$

Es muss also gelten

$$\mathbf{W}\mathbf{d}\left(\frac{\mathbf{d}'\,\bar{\mathbf{x}}_1 - \mathbf{d}'\,\bar{\mathbf{x}}_2}{\mathbf{d}'\,\mathbf{W}\,\mathbf{d}}\right) = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\,.$$

Dabei ist der Ausdruck

$$\frac{\mathbf{d}'\,\bar{\mathbf{x}}_1 - \mathbf{d}'\,\bar{\mathbf{x}}_2}{\mathbf{d}'\mathbf{W}\mathbf{d}}$$

für gegebenes $\mathbf{d}$ eine Konstante, die die Richtung von $\mathbf{d}$ nicht beeinflusst. Somit ist $\mathbf{d}$ proportional zu

$$\mathbf{W}^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\,.$$

Wir wählen

$$\mathbf{d} = \mathbf{W}^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\,. \tag{12.31}$$

Dieser Ausdruck weist die gleiche Struktur wie (12.26) auf. Schauen wir uns $\mathbf{W}$ an. Es gilt

$$\mathbf{W} = \sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)\,(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)' + \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)\,(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'.$$

Mit (2.16) gilt

$$\mathbf{W} = (n_1 - 1)\,\mathbf{S}_1 + (n_2 - 1)\,\mathbf{S}_2.$$

Mit (12.24) gilt also

$$\mathbf{W} = (n_1 + n_2 - 2)\,\mathbf{S}.$$

Bis auf eine multiplikative Konstante ist $\mathbf{d}$ in Gleichung (12.31) gleich $\mathbf{a}$ in Gleichung (12.26). Also liefert der Ansatz von Fisher die gleiche Entscheidungsregel wie bei Normalverteilung mit gleichen Varianz-Kovarianz-Matrizen. Der Ansatz von Fisher kommt ohne die Annahme der Normalverteilung und identischer Varianzen aus, wobei er ein sinnvolles Zielkriterium verwendet. Dies deutet darauf hin, dass man die lineare Diskriminanzanalyse in vielen Situationen anwenden kann.

Example 49. Wir betrachten die Daten in Tabelle 12.1 auf Seite 352 und benutzen alle vier Merkmale zur Klassifikation.
In Gruppe 1 gilt

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 0.556 \\ 0.889 \\ 2.667 \\ 0.333 \end{pmatrix}.$$

In Gruppe 2 gilt

$$\bar{\mathbf{x}}_2 = \begin{pmatrix} 0.455 \\ 0.273 \\ 3.182 \\ 0.364 \end{pmatrix}.$$

Die Mittelwerte der binären Merkmale können wir sehr schön interpretieren, da wir diese Merkmale mit 0 und 1 kodiert haben. In diesem Fall ist der Mittelwert gleich dem Anteil der Personen, die die Eigenschaft besitzen, die wir mit 1 kodiert haben. So sind also 55.6 Prozent in der ersten Gruppe und 45.5 Prozent in der zweiten Gruppe weiblich. Dies haben wir bereits in Tabelle 12.3 auf Seite 353 gesehen. Nun benötigen wir noch die gepoolte Varianz-Kovarianz-Matrix.
In Gruppe 1 gilt

$$\mathbf{S}_1 = \begin{pmatrix} 0.278 & -0.056 & 0.208 & 0.042 \\ -0.056 & 0.111 & -0.167 & -0.083 \\ 0.208 & -0.167 & 1.000 & 0.125 \\ 0.042 & -0.083 & 0.125 & 0.250 \end{pmatrix}.$$

In Gruppe 2 gilt

$$\mathbf{S}_2 = \begin{pmatrix} 0.273 & 0.064 & -0.191 & 0.218 \\ 0.064 & 0.218 & -0.255 & 0.091 \\ -0.191 & -0.255 & 0.564 & -0.173 \\ 0.218 & 0.091 & -0.173 & 0.255 \end{pmatrix}.$$

Hieraus ergibt sich

$$\mathbf{W} = \begin{pmatrix} 4.954 & 0.192 & -0.246 & 2.516 \\ 0.192 & 3.068 & -3.886 & 0.246 \\ -0.246 & -3.886 & 13.640 & -0.730 \\ 2.516 & 0.246 & -0.730 & 4.550 \end{pmatrix}.$$

Es gilt

$$\mathbf{W}^{-1} = \begin{pmatrix} 0.2812 & -0.0145 & -0.0074 & -0.1559 \\ -0.0145 & 0.5108 & 0.1455 & 0.0037 \\ -0.0074 & 0.1455 & 0.1154 & 0.0147 \\ -0.1559 & 0.0037 & 0.0147 & 0.3082 \end{pmatrix}.$$

Mit

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \begin{pmatrix} 0.101 \\ 0.616 \\ -0.515 \\ -0.031 \end{pmatrix}$$

gilt

$$\mathbf{d} = \begin{pmatrix} 0.028 \\ 0.238 \\ 0.029 \\ -0.030 \end{pmatrix}.$$

Somit gilt $\bar{y}_1 = \mathbf{d}'\bar{\mathbf{x}}_1 = 0.295$ und $\bar{y}_2 = \mathbf{d}'\bar{\mathbf{x}}_2 = 0.159$. Eine Beobachtung mit Merkmalsvektor $\mathbf{x}$ ordnen wir also Gruppe 1 zu, wenn gilt

$$|\mathbf{d}'\mathbf{x} - \bar{y}_1| < |\mathbf{d}'\mathbf{x} - \bar{y}_2|. \tag{12.32}$$

Der erste Student hat den Merkmalsvektor

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 4 \\ 1 \end{pmatrix}.$$

Es gilt $\mathbf{d}'\mathbf{x} = 0.114$. Wir ordnen ihn also Gruppe 2 zu. □

12.4 Logistische Diskriminanzanalyse

Die Bayes-Entscheidungsregel hängt ab von $P(Y = 1|\mathbf{x})$ und $P(Y = 0|\mathbf{x})$. Es liegt nahe, diese Wahrscheinlichkeiten zu schätzen und ein Objekt der Gruppe 1 zuzuordnen, wenn gilt

$$\hat{P}(Y = 1|\mathbf{x}) > 0.5.$$

Wie soll man die Wahrscheinlichkeit $P(Y = 1|\mathbf{x})$ schätzen? Der einfachste Ansatz besteht darin, $\hat{P}(Y = 1|\mathbf{x})$ als Linearkombination der Komponenten $x_1, \ldots, x_p$ von $\mathbf{x}$ darzustellen:

$$P(Y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Es gilt

$$E(Y|\mathbf{x}) = 1 \cdot P(Y = 1|\mathbf{x}) + 0 \cdot (1 - P(Y = 1|\mathbf{x})) = P(Y = 1|\mathbf{x}).$$

Wir können also (12.33) als lineares Regressionsmodell auffassen und die Kleinste-Quadrate-Schätzer der Parameter $\beta_0, \beta_1, \ldots, \beta_p$ bestimmen.

Example 50. Wir betrachten das Beispiel 11 auf Seite 10. Wir wollen den Typ der Filiale auf der Basis des Merkmals `Gesamtkosten` klassifizieren. Wir kodieren die Zweigstellen, die einen hohen Marktanteil und ein überdurchschnittliches Darlehens- und Kreditgeschäft besitzen, mit dem Wert 1, die restlichen mit dem Wert 0. Abbildung 12.5 zeigt das Streudiagramm der Gruppenvariablen Y und des Merkmals `Gesamtkosten`. Außerdem ist noch die Kleinste-Quadrate-Gerade eingezeichnet.

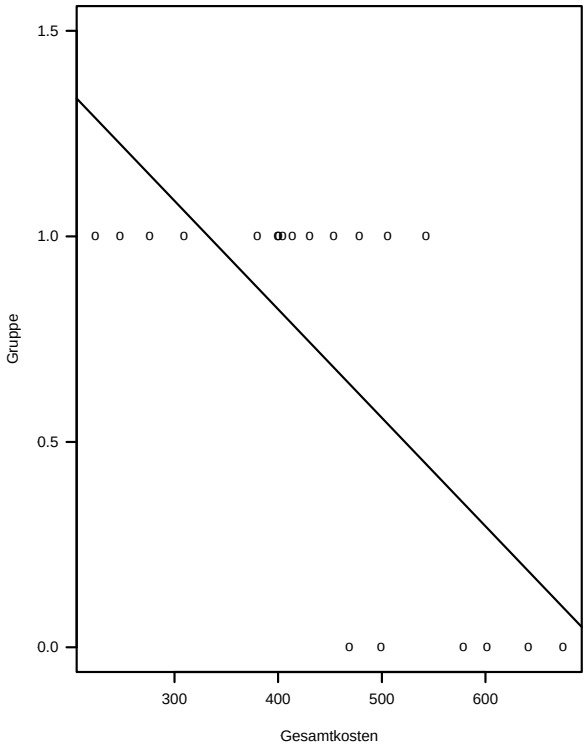


Fig. 12.5. Streudiagramm des Merkmals Gesamtkosten und der Gruppenvariablen

□

Abbildung 12.5 zeigt den Nachteil dieses Ansatzes. Die geschätzten Wahrscheinlichkeiten können aus dem Intervall $[0, 1]$ herausfallen. Das kann man durch folgenden Ansatz verhindern:

$$P(Y = 1|\mathbf{x}) = F(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p).$$

Dabei ist $F(x)$ die Verteilungsfunktion einer stetigen Zufallsvariablen. Wählt man die Verteilungsfunktion

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}$$

der logistischen Verteilung, so erhält man folgendes Modell:

$$P(Y = 1|\mathbf{x}) = \frac{\exp\left(\beta_0 + \beta_1\, x_1 + \ldots + \beta_p\, x_p\right)}{1 + \exp\left(\beta_0 + \beta_1\, x_1 + \ldots + \beta_p\, x_p\right)}. \tag{12.33}$$

Man spricht von *logistischer Regression*. Die logistische Regression ist detailliert beschrieben bei Hosmer & Lemeshow (1989) und Kleinbaum (1994). Hier ist auch eine Herleitung der Maximum-Likelihood-Schätzer der Parameter β_0, β_1,...,β_p zu finden. Wir zeigen in Kapitel 12.7, wie man die Schätzer mit `S-PLUS` gewinnt.

hmcounterend. (fortgesetzt)

Example 50. Wir bezeichnen die Gesamtkosten mit x_1. Wir unterstellen folgendes Modell:

$$P(Y = 1|x_1) = \frac{\exp\left(\beta_0 + \beta_1\, x_1\right)}{1 + \exp\left(\beta_0 + \beta_1\, x_1\right)}$$

und erhalten folgende Schätzer:

$$\hat{\beta}_0 = 15.2, \qquad \hat{\beta}_1 = -0.0294\,.$$

Abbildung 12.6 zeigt die Funktion $\hat{P}(Y = 1|x_1)$.
In der Graphik ist an der Stelle, an der die geschätzte Wahrscheinlichkeit gleich 0 ist, eine senkrechte Linie eingetragen. Alle Punkte links von der Linie werden der ersten Gruppe zugeordnet. □

Ein großer Vorteil der logistischen Diskriminanzanalyse ist, dass sie auch angewendet werden kann, wenn die Merkmale qualitatives Messniveau aufweisen. Wir haben Fishers lineare Diskriminanzanalyse auf das Beispiel 2 auf Seite 3 angewendet. Hierbei haben wir die binären Merkmale mit den Werten 0 und 1 kodiert. Die logistische Diskriminanzanalyse kann man auf diese Daten anwenden, ohne sie vorher zu transformieren.

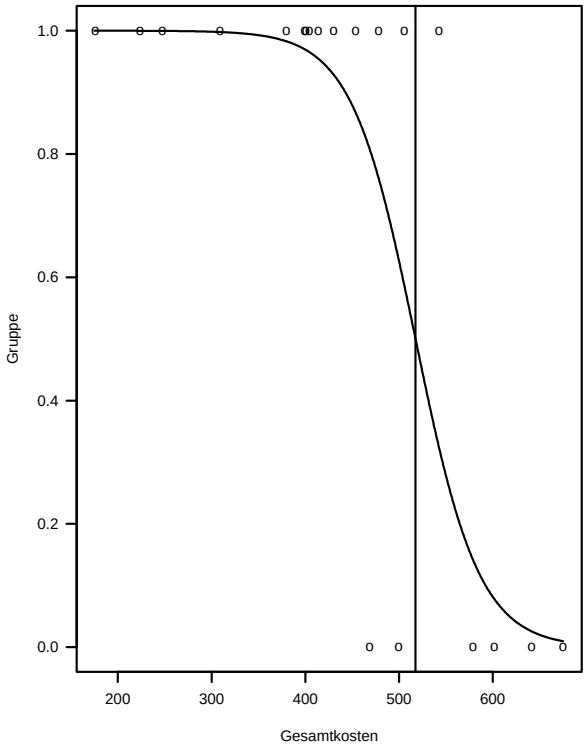


Fig. 12.6. Streudiagramm des Merkmals Gesamtkosten und der Gruppenvariablen mit geschätzter logistischer Funktion

12.5 Klassifikationsbäume

Wir betrachten wieder das Problem, ein Objekt entweder der Gruppe 1 oder der Gruppe 2 zuzuordnen. Um dieses Ziel zu erreichen, werden eine Reihe von Merkmalen erhoben. Bei den klassischen Verfahren der Diskriminanzanalyse werden alle Merkmale gleichzeitig zur Entscheidungsfindung benutzt. Man kann bei der Entscheidungsfindung aber auch sequentiell vorgehen. Hierbei werden nacheinander Ja-Nein-Fragen gestellt, wobei eine Frage in Abhängigkeit von der vorhergehenden Antwort ausgewählt wird. Das Ergebnis ist ein sogenannter *Klassifikationsbaum*.

Example 51. Wir schauen uns wieder die Daten in Tabelle 12.1 auf Seite 352 an. Dabei betrachten wir nur die binären Merkmale `Geschlecht`, `MatheLK` und `Abitur88`. Ein Student soll auf Basis dieser Merkmale einer der beiden

Gruppen zugeordnet werden. Abbildung 12.7 zeigt einen Klassifikationsbaum der Daten.

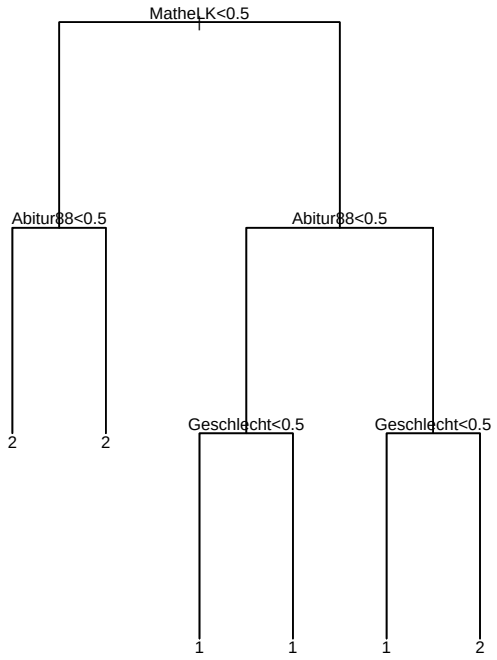


Fig. 12.7. Klassifikationsbaum der Daten des Mathematik-Tests

□

Der Baum besteht aus *Knoten* und *Ästen*. Bei den Knoten unterscheidet man *Entscheidungsknoten* und *Endknoten*. Zu jedem Entscheidungsknoten gehört eine Frage, die mit ja oder nein beantwortet werden kann. Wird die Frage mit ja beantwortet, geht man im linken Ast des Baumes zum nächsten Knoten. Wird die Frage hingegen mit nein beantwortet, geht man im rechten Ast des Baumes zum nächsten Knoten. Der oberste Knoten heißt auch *Wurzelknoten*. hmcounterend. (fortgesetzt)

Example 51. Im Baum in Abbildung 12.7 lautet die erste Frage:

$$\text{MatheLK} < 0.5?$$

Diese können wie übersetzen mit:

Hat der Studierende den Mathematik-Leistungskurs nicht besucht?

Wird diese Frage mit nein beantwortet, so gehen wir im rechten Ast zum nächsten Knoten lautet. Hier lautet die Frage:

$$\text{Abitur88} < 0.5?$$

Wir fragen also, ob der Studierende sein Abitur nicht 1988 gemacht hat. Wird diese Frage verneint, so landen wir im rechten Ast. Die letzte Frage lautet hier:

$$\text{Geschlecht} < 0.5?$$

Wir fragen also, ob der Studierende männlich ist. Wird die Frage mit ja beantwortet, so landen wir im linken Ast in einem Endknoten. Diesem ist die Zahl 1 zugeordnet. Dies bedeutet, dass wir den Studierenden der Gruppe 1 zuordnen. Ein männlicher Studierender, der den Mathematik-Leistungskurs besucht hat und sein Abitur 1988 gemacht hat, wird also der Gruppe zugeordnet, die den Test besteht. □

Beginnend beim Wurzelknoten werden also so lange Ja-Nein-Fragen gestellt, bis man in einem Endknoten landet, dem eine der beiden Gruppen zugeordnet ist. Wie konstruiert man einen Klassifikationsbaum? Zur Beantwortung dieser Frage bezeichnen wir in Anlehnung an Breiman et al. (1984) den i-ten Knoten mit t_i, wobei wir die Knoten einer Hierarchiestufe von links nach rechts durchnummerieren. hmcounterend. (fortgesetzt)

Example 51. Abbildung 12.8 zeigt den obigen Baum in dieser Notation. □

Der Konstruktionsprozess beginnt beim Wurzelknoten. Diesem ist die gesamte Population zugeordnet. Sei p_{t_1} die Wahrscheinlichkeit, dass ein Objekt im Wurzelknoten zur Gruppe 1 gehört. Ist p_{t_1} größer als 0.5, so ordnen wir das Objekt der Gruppe 1 zu, ist p_{t_1} hingegen kleiner als 0.5, so ordnen wir es der Gruppe 2 zu. Ist p_{t_1} gleich 0.5, so ordnen wir es zufällig einer der beiden Gruppen zu. hmcounterend. (fortgesetzt)

Example 51. Wegen $p_{t_1} = 0.45$ ordnen wir einen zufällig ausgewählten Studenten der Gruppe 2 zu. □

Je näher p_{t_1} an 0.5 liegt, umso fehlerhafter ist die Entscheidung. Wie können wir die Unsicherheit durch eine Maßzahl quantifizieren? Um diese Frage zu beantworten, betrachten wir die Zufallsvariable Y mit

$$Y = \begin{cases} 1 & \text{falls der Studierende zur Gruppe 1 gehört,} \\ 0 & \text{falls der Studierende zur Gruppe 2 gehört.} \end{cases}$$

Die Zufallsvariable Y ist bernoulliverteilt mit Parameter p_{t_1}. Die Varianz von Y ist $p_{t_1}(1 - p_{t_1})$. Dies wird auf Seite 78 bewiesen. Je näher p_{t_1} an 0.5

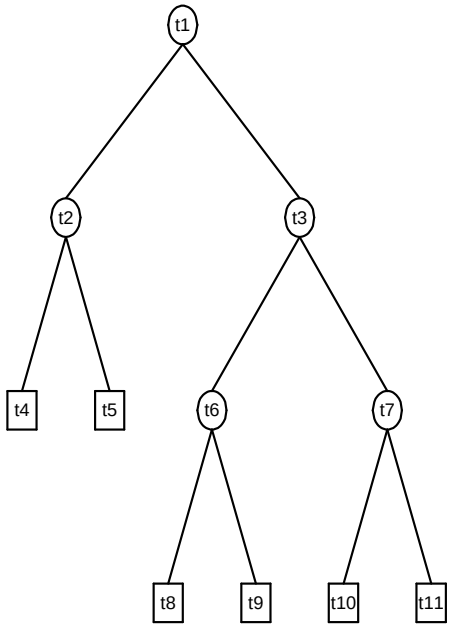


Fig. 12.8. Klassifikationsbaum in allgemeiner Notation

liegt, umso größer ist die Varianz. Die Varianz ist minimal, wenn p_{t_1} gleich 0 oder 1 ist. Die Varianz ist also geeignet, die Unsicherheit eines Knotens zu quantifizieren. Breiman et al. (1984) verwenden sie als *Unreinheitsmaß* eines Knotens t und bezeichnen dieses Unreinheitsmaß mit $i(t)$. Es gilt also

$$i(t) = p_t(1 - p_t).$$

Dabei ist p_t die Wahrscheinlichkeit, dass sich ein Objekt im Knoten t in Gruppe 1 befindet. Breiman et al. (1984) betrachten noch folgendes Maß für die Unreinheit eines Knotens t:

$$-\,p_t\,\ln p_t - (1 - p_t)\,\ln\,(1 - p_t). \tag{12.34}$$

Man nennt (12.34) auch *Entropie*. Wir wollen im Folgenden $i(t) = p_t\,(1 - p_t)$ betrachten. hmcounterend. (fortgesetzt)

Example 51. Im Wurzelknoten t_1 gilt

$$i(t_1) = 0.45 \cdot 0.55 = 0.2475.$$

□

Im Beispiel ist die Unreinheit des Wurzelknotens t_1 sehr groß. Das liegt daran, dass beide Gruppen nahezu gleich häufig in der Population vertreten sind. Die Population ist also sehr heterogen bezüglich des Merkmals, das die Gruppen definiert. Wir wollen die Population in zwei Teilpopulationen zerlegen, die homogener sind. Hierzu benutzen wir Informationen über die Objekte. Die erste Frage liefert die erste Information. Hinter dieser Frage steht eine Regel, die die Population in zwei Teilpopulationen t_2 und t_3 zerlegt. Wir bezeichnen im Folgenden eine Regel, die eine Population zerlegt, mit dem Symbol s. Dabei steht s für split. Im allgemeinen Fall zerlegen wir einen Knoten t in einen linken Knoten t_L und einen rechten Knoten t_R. hmcounterend. (fortgesetzt)

Example 51. Die erste Teilpopulation besteht aus den Studierenden, die keinen Mathematik-Leistungskurs besucht haben, und ist dem Knoten t_2 zugeordnet. Die zweite Teilpopulation besteht aus den Studierenden, die den Mathematik-Leistungskurs besucht haben, und ist dem Knoten t_3 zugeordnet. In Tabelle 12.1 auf Seite 352 gehören die ersten 9 Studierenden zum Knoten t_2, während die restlichen 11 Studierenden dem Knoten t_3 zugeordnet sind. □

Um zu entscheiden, ob wir durch die Zerlegung der Population eines Knotens t in die Knoten t_L und t_R besser zwischen den Gruppen diskriminieren können, bestimmen wir zuerst die Unreinheiten $i(t_L)$und $i(t_R)$. hmcounterend. (fortgesetzt)

Example 51. Schauen wir uns den Knoten t_2 an. Wir sehen, dass einer von den ersten 9 Studenten in Tabelle 12.1 in Gruppe 1 ist. Somit ist $p_{t_2} = \frac{1}{9}$. Es gilt also

$$i(t_2) = p_{t_2}(1 - p_{t_2}) = \frac{1}{9}\frac{8}{9} = 0.09877.$$

Die Unreinheit dieses Knotens ist viel kleiner als die Unreinheit des Wurzelknotens. Im Knoten t_3 sind 8 von 11 Studierenden in Gruppe 1. Somit ist $p_{t_3} = \frac{8}{11}$. Es gilt also

$$i(t_3) = p_{t_3}(1 - p_{t_3}) = \frac{8}{11}\frac{3}{11} = 0.1983.$$

Die Unreinheit des Knotens t_3 ist kleiner als die Unreinheit des Wurzelknotens. Wir haben die Population in zwei Teilpopulationen aufgeteilt, die hinsichtlich des Merkmals `Bestanden` homogener sind. □

Sei p_{t_L} die Wahrscheinlichkeit, vom Knoten t aus im linken Knoten zu landen, und p_{t_R} die Wahrscheinlichkeit, vom Knoten t_1 im rechten Knoten zu landen. hmcounterend. (fortgesetzt)

Example 51. Es gilt $p_{t_2} = 0.45$ und $p_{t_3} = 0.55$ □

Breiman et al. (1984) definieren die Verminderung $\Delta(s,t)$ der Unreinheit, die sich durch Zerlegung s eines Knotens t in den linken Knoten t_L und den rechten Knoten t_R ergibt, durch

$$\Delta(s,t) = i(t) - p(t_L)\, i(t_L) - p(t_R)\, i(t_R).$$

hmcounterend. (fortgesetzt)

Example 51. Es gilt

$$\begin{aligned}\Delta(s,t_1) &= 0.2475 - 0.45 \cdot 0.09877 - 0.55 \cdot 0.1983 \\ &= 0.2475 - 0.1535115 = 0.0939885.\end{aligned}$$

□

Welches der Merkmale soll man wählen, um einen Knoten zu zerlegen? Breiman et al. (1984) schlagen vor, die Zerlegung auf Basis des Merkmals durchzuführen, bei dem die Verminderung der Unreinheit am größten ist. hmcounterend. (fortgesetzt)

Example 51. Tabelle 12.8 zeigt die Werte von $\Delta(s,t_1)$ für die Merkmale `Geschlecht`, `MatheLK` und `Abitur88`.

Table 12.8. Werte von $\Delta(s,t_1)$ für die Merkmale Geschlecht, MatheLK und Abitur88

Merkmal	$\Delta(s,t_1)$
Geschlecht	0.00250
MatheLK	0.09399
Abitur88	0.00025

Wir sehen, dass das Merkmal `MatheLK` die größte Verbesserung bewirkt. □

Die eben beschriebene Vorgehensweise können wir auf jeden Knoten t anwenden. Wann endet dieser Prozess? Breiman et al. (1984) schlagen vor, einen Knoten t nicht weiter zu zerlegen, wenn $i(t)$ gleich 0 ist. In einem Endknoten wird dann ein Objekt der Gruppe zugeordnet, die am häufigsten in diesem Knoten vertreten ist.
Wodurch zeichnet sich diese Regel aus? Sei $\tilde{T}$ die Menge der Endknoten und $\tilde{p}(t)$ die Wahrscheinlichkeit, dass ein Objekt im Endknoten t landet. Breiman et al. (1984) definieren die Unreinheit $I(T)$ des Baumes durch

$$I(T) = \sum_{t \in \tilde{T}} i(t)\, \tilde{p}(t). \tag{12.35}$$

Breiman et al. (1984), S. 32-33, zeigen, dass $I(T)$ minimiert wird, wenn die Zerlegung s jedes Knotens t so gewählt wird, dass $\Delta(s,t)$ minimal ist.

Wir haben die Klassifikationsbäume bisher nur für den Fall betrachtet, dass zwei Gruppen vorliegen, alle Merkmale binär sind und es sich um eine Vollerhebung handelt. Liegen mehr als zwei Gruppen vor, so ergeben sich zwei Modifikationen. Wir ordnen in einem Endknoten ein Objekt der Gruppe zu, die im Endknoten am häufigsten vertreten ist. Außerdem ändert sich die Bestimmung des Unreinheitsmaßes in einem Knoten. Seien p_{it} für $i = 1, \ldots, k$ die Wahrscheinlichkeiten, dass ein Objekt im Knoten t zur i-ten Gruppe gehört. Es liegt nahe, $i(t) = p_t\,(1 - p_t)$ folgendermaßen zu verallgemeinern:

$$i(t) = \sum_{i=1}^{k} p_{it}\,(1 - p_{it}) = \sum_{i=1}^{k} p_{it} - \sum_{i=1}^{k} p_{it}^2 = 1 - \sum_{i=1}^{k} p_{it}^2. \qquad (12.36)$$

Man nennt (12.36) den *Gini-Index*. Als Verallgemeinerung der Entropie erhält man

$$i(t) = -\sum_{i=1}^{k} p_{it}\,\ln p_{it}. \qquad (12.37)$$

Bei einem binären Merkmal ist eine Verzweigung eindeutig definiert. Bei einem qualitativen Merkmal mit mehr als zwei Ausprägungen gibt es mehr Möglichkeiten. Sei $A = \{a_1, \ldots, a_k\}$ die Menge der Ausprägungsmöglichkeiten. Dann werden alle Zerlegungen in S und $\bar{S}$ mit $S \subset A$ betrachtet. Sind die Merkmalsausprägungen eines Merkmals X geordnet, so betrachtet man alle Zerlegungen der Form $X \leq c$ und $X > c$ mit $c \in \mathbb{R}$.

Bisher haben wir die Konstruktion eines Klassifikationsbaums für den Fall betrachtet, dass die Grundgesamtheit vollständig bekannt ist. Handelt es sich um eine Stichprobe, so ersetzt man die Wahrscheinlichkeiten durch die korrespondierenden relativen Häufigkeiten.

Liegen die Daten in Form einer Stichprobe vor, so muss man sich Gedanken über die Größe des Baumes machen. Ist der Baum zu groß, so werden Objekte, die nicht zur Stichprobe gehören, unter Umständen falsch klassifiziert, da die Struktur des Baumes sehr stark von der Stichprobe abhängt. Ist der Baum hingegen zu klein, so wird vielleicht nicht die ganze Struktur der Grundgesamtheit abgebildet. Man muss also einen Mittelweg finden. Von Breiman et al. (1984) wurde vorgeschlagen, zunächst den vollständigen Baum zu konstruieren und ihn dann geeignet zu beschneiden. Ist $\tilde{T}$ die Menge aller Endknoten eines Baumes, so betrachten Breiman et al. (1984) eine Schätzung $R(\tilde{T})$ der Fehlerrate. Es wird der Baum ausgewählt, bei dem

$$R(\tilde{T}) + \alpha\,|\tilde{T}|$$

minimal ist. Dabei ist $|\tilde{T}|$ die Anzahl der Endknoten des Baumes und α eine Konstante. Details zu dieser Vorgehensweise sind bei Breiman et al. (1984) zu finden.

Bei einigen Verfahren zur Konstruktion von Klassifikationsbäumen beruht die Entscheidung, durch welches Merkmal ein Knoten zerlegt werden soll,

auf Teststatistiken. Schauen wir uns die Vorgehensweise von Clark & Pregibon (1992) für die Zerlegung eines Knotens t in die Knoten t_L und t_R an. Sei n_t die Anzahl der Beobachtungen im Knoten t und n_{1t} die Anzahl der Beobachtungen im Knoten t, die zur Gruppe 1 gehören. Außerdem sei p_{1t} die Wahrscheinlichkeit, dass eine Beobachtung im Knoten t zur Gruppe 1 gehört. Die *Devianz* des Knotens t ist

$$D_t = -2\left[n_{1t}\ln p_{1t} + (n - n_{1t})\ln\left(1 - p_{1t}\right)\right]. \tag{12.38}$$

Die unbekannte Wahrscheinlichkeit p_{1t} wird geschätzt durch n_{1t}/n_t. Setzen wir dies ein in Gleichung 12.38, so erhalten wir die geschätzte Devianz

$$\hat{D}_t = -2\left[n_{1t}\ln n_{1t}/n_t + (n - n_{1t})\ln\left(1 - n_{1t}/n_t\right)\right]. \tag{12.39}$$

hmcounterend. (fortgesetzt)

Example 51. Im Wurzelknoten t gilt $n_t = 20$ und $n_{1t} = 9$. Somit gilt

$$\hat{D}_t = -2\left[9\ln 9/20 + (20 - 9)\ln\left(1 - 9/20\right)\right] = 27.53.$$

□

Man kann mit der Devianz die Entscheidung fällen, auf Basis welchen Merkmals ein Knoten t in zwei Knoten t_L und t_R verzweigt werden soll. Hierzu bestimmt man die Devianzen $\hat{D}_t$, $\hat{D}_{t_L}$ und $\hat{D}_{t_R}$ der Knoten t, t_L und t_R für jedes der Merkmale und wählt das Merkmal, bei dem die Verminderung der Devianz

$$\hat{D}_t - \hat{D}_{t_L} - \hat{D}_{t_R}$$

am größten ist. hmcounterend. (fortgesetzt)

Example 51. Schauen wir uns das Merkmal `MatheLK` als Kriterium an. Im linken Ast sind die Studenten, die keinen Mathematik-Leistungskurs besucht haben. Dies sind 9. Von diesen gehört einer zur Gruppe 1 der Studierenden, die den Test bestehen. Somit gilt

$$\hat{D}_{t_L} = -2\left[1\ln 1/9 + (9 - 1)\ln\left(1 - 1/9\right)\right] = 6.28.$$

Im linken Ast sind die Studenten, die einen Mathematik-Leistungskurs besucht haben. Dies sind 11. Von diesen gehören 8 zur Gruppe 1. Somit gilt

$$\hat{D}_{t_R} = -2\left[8\ln 8/11 + (11 - 8)\ln\left(1 - 8/11\right)\right] = 12.89.$$

Die Verminderung der Devianz beträgt

$$\hat{D}_t - \hat{D}_{t_L} - \hat{D}_{t_R} = 27.53 - 6.28 - 12.89 = 8.36.$$

Beim Merkmal `Geschlecht` beträgt die Verminderung der Devianz 0.21 und beim Merkmal `Abitur88` 0.03. Es wird also das Merkmal `MatheLK` für die erste Zerlegung gewählt. □

Auf Basis der Devianz kann man ein Kriterium angeben, wann ein Baum nicht weiter verzweigt werden soll. Ist die Devianz eines Knotens kleiner als 1 Prozent der Devianz des Wurzelknotens, so wird dieser Knoten nicht weiter verzweigt.

12.6 Praktische Aspekte

Schätzung der Fehlerrate Die Güte eines Verfahrens der Diskriminanzanalyse beurteilt man anhand der Fehlerrate, die man auf Basis der Daten schätzen muss.

Example 52. Wir betrachten das Beispiel 48 auf Seite 368. Wir unterstellen Normalverteilung und identische Varianzen. Die Entscheidungsregel ist in Gleichung (12.27) auf Seite (371) zu finden. □

Das einfachste Verfahren zur Schätzung der Fehlerrate besteht darin, mit der geschätzten Entscheidungsregel jede der Beobachtungen zu klassifizieren. Als Schätzwert der Fehlerrate dient der Anteil der fehlklassifizierten Beobachtungen. Man spricht auch von der *Resubstitutionsfehlerrate*. hmcounterend. (fortgesetzt)

Example 52. Es werden zwei quantitative Merkmale verwendet, sodass man die Fehlerrate schätzen kann, indem man die Punkte zählt, die auf der falschen Seite der Geraden liegen. Abbildung 12.4 auf Seite 372 zeigt, dass nur eine Beobachtung fehlklassifiziert wird. Die geschätzte Fehlerrate beträgt somit 0.05. □

Durch diese Schätzung wird die Fehlerrate meist unterschätzt. Das liegt daran, dass die Entscheidungsregel aus den gleichen Daten geschätzt wurde, die zur Schätzung der Fehlerrate benutzt werden. Die Daten, die man zur Schätzung der Entscheidungsregel verwendet, sollten aber unabhängig von den Daten sein, mit denen man die Fehlerrate schätzt. Ist der Stichprobenumfang groß, so kann man die Stichprobe in eine *Lernstichprobe* und eine *Teststichprobe* aufteilen. Mit den Beobachtungen der Lernstichprobe wird die Entscheidungsregel geschätzt. Die Beobachtungen der Teststichprobe werden mit dieser Entscheidungsregel klassifiziert. Der Anteil der fehlklassifizierten Beobachtungen dient als Schätzung der Fehlerrate. hmcounterend. (fortgesetzt)

Example 52. Wir wählen aus den 20 Filialen 10 Filialen zufällig für die Lernstichprobe aus. Es sind dies die folgenden Filialen:

```
1  6  8 10 11 15 16 17 19 20 .
```

Die restlichen Filialen bilden die Teststichprobe. Klassifizieren wir die Beobachtungen der Teststichprobe, so werden zwei Beobachtungen fehlklassifiziert. Die geschätzte Fehlerrate beträgt somit 0.2. □

Wir haben im Beispiel die Beobachtungen in eine Lern- und eine Teststichprobe aufgeteilt, um die Schätzung der Fehlerrate über die Resubstitutionsfehlerrate zu illustrieren. Eigentlich ist aber die Anzahl der Beobachtungen viel zu klein, um den Datensatz in eine Lern- und Teststichprobe aufzuteilen. Lachenbruch & Mickey (1968) haben eine Vorgehensweise zur Schätzung der

Fehlerrate vorgeschlagen, bei der die Fehlerrate auch für kleine Datensätze geschätzt werden kann. Die Entscheidungsregel wird dabei ohne die i-te Beobachtung geschätzt. Anschließend wird die i-te Beobachtung auf Basis der Entscheidungsregel klassifiziert. Dies geschieht für alle Beobachtungen. Als Schätzer der Fehlerrate dient die Anzahl der fehlklassifizierten Beobachtungen. Man bezeichnet das Verfahren als *Leaving-one-out-Methode*. hmcounterend. (fortgesetzt)

Example 52. Wir entfernen also jeweils eine Beobachtung aus der Stichprobe, schätzen die Entscheidungsregel und klassifizieren die weggelassene Bobachtung. Die 7-te und die 18-te Beobachtung werden falsch klassifiziert. Somit beträgt die geschätzte Fehlerrate 0.1. □

Ein Vergleich der linearen Diskriminanzanalyse mit Klassifikationsbäumen Wir wollen in diesem Abschnitt die lineare Diskriminanzanalyse und Klassifikationsbäume betrachten und anhand von zwei artifiziellen Beispielen Konstellationen aufzeigen, in denen eines der beiden Verfahren dem anderen überlegen ist.

Example 53. Wir betrachten ein Beispiel aus Breiman et al. (1984), S. 39. Abbildung 12.9 zeigt das Streudiagramm von zwei Merkmalen.
Die Beobachtungen stammen aus den Gruppen 1 und 2. Die Gruppenzugehörigkeit jeder Beobachtung ist im Streudiagramm markiert. Außerdem ist die Gerade eingezeichnet, die man erhält, wenn man Fishers lineare Diskriminanzanalyse anwendet. Wir sehen, dass die lineare Diskriminanzanalyse die Gruppen sehr gut trennt. Abbildung 12.10 zeigt den vollständigen Klassifikationsbaum. Wir sehen, dass sehr viele Fragen notwendig sind, um ein Objekt zu klassifizieren.
Abbildung 12.11 zeigt die Zerlegung der (x, y)-Ebene, die sich aus dem Klassifikationsbaum ergibt.
Die zugrunde liegende Struktur kann durch den Klassifikationsbaum nur durch eine Vielzahl von Fragen ermittelt werden. Beschneidet man den Baum, so wird die Fehlerrate hoch sein. Breiman et al. (1984) schlagen vor, die Entscheidungen im Klassifikationsbaum auf der Basis von Linearkombinationen der Merkmale zu fällen. Von Loh & Shih (1997) wurde dieser Ansatz weiterentwickelt und im Programm QUEST implementiert. Ein Nachteil dieses Ansatzes ist die Interpretierbarkeit der Entscheidungen, da diese auf Linearkombinationen der Merkmale beruhen. □

Example 54. Abbildung 12.12 zeigt das Streudiagramm von zwei Merkmalen, wobei die Beobachtungen aus zwei Gruppen stammen. Auch hier ist die Gerade eingezeichnet, die auf Grund von Fishers linearer Diskriminanzanalyse die Gruppen trennt.
Wir sehen, dass die Resubstitutionsfehlerrate sehr hoch ist. Die lineare Diskriminanzanalyse ist für diese Konstellation nicht geeignet. Abbildung 12.13 zeigt den vollständigen Klassifikationsbaum.

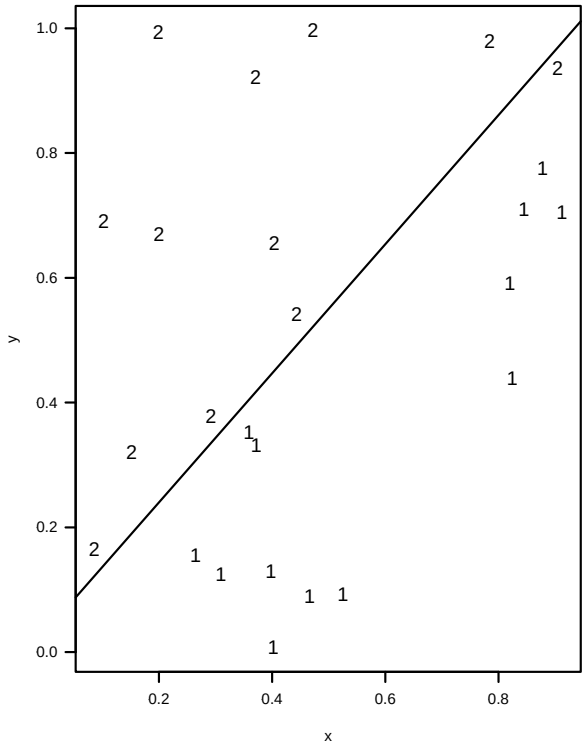


Fig. 12.9. Streudiagramm von 25 artifiziellen Beobachtungen mit der Geraden, die man durch Fishers lineare Diskriminanzanalyse erhält

Berücksichtigt man die Entscheidungsregeln des Klassifikationsbaums im Streudiagramm, so erkennt man, dass die beiden Gruppen durch den Klassifikationsbaum nahezu perfekt getrennt werden. Dies kann man in Abbildung 12.14 erkennen.

□

12.7 Diskriminanzanalyse in S-PLUS

Wir wollen Fishers lineare Diskiminanzanalyse auf das Beispiel 11 auf Seite 10 anwenden. Die Daten mögen in der Matrix `bank` stehen:

```
> bank
    Einwohner Gesamtkosten
1        1642         478.2
```

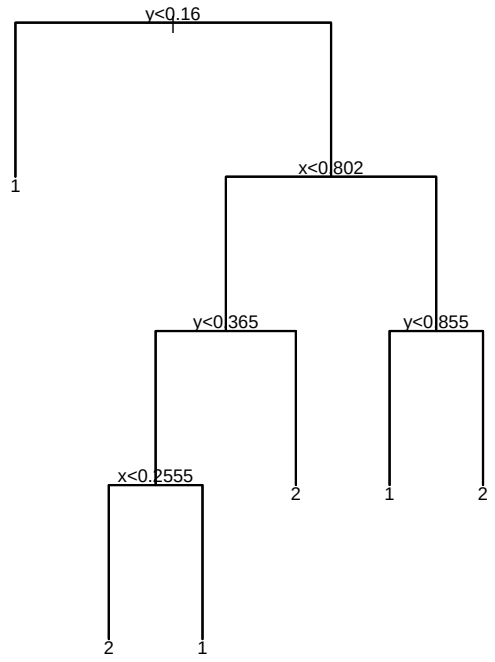


Fig. 12.10. Klassifikationsbaum von 25 artifiziellen Beobachtungen in zwei Gruppen

```
 2      2418        247.3
 3      1417        223.6
 4      2761        505.6
 5      3991        399.3
 6      2500        276.0
 7      6261        542.5
 8      3260        308.9
 9      2516        453.6
10      4451        430.2
11      3504        413.8
12      5431        379.7
13      3523        400.5
14      5471        404.1
15      7172        499.4
```

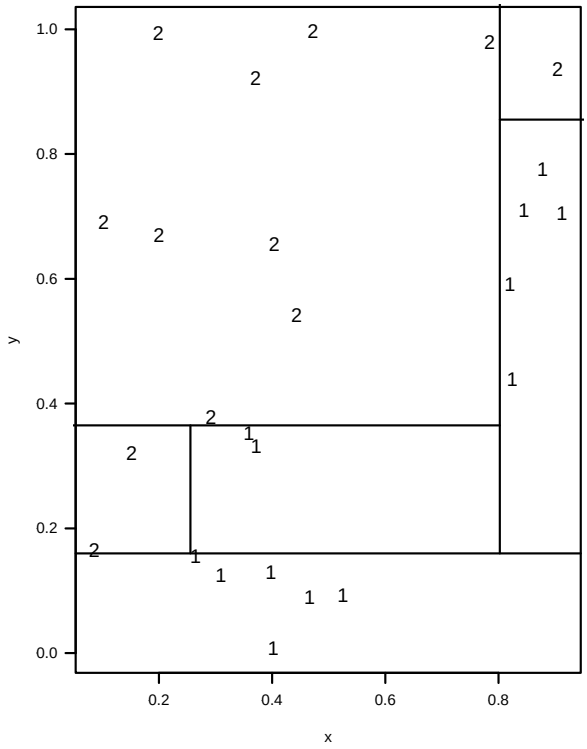


Fig. 12.11. Streudiagramm von 25 artifiziellen Beobachtungen mit Zerlegung der Ebene, die sich aus dem Klassifikationsbaum ergibt

```
16      9419        674.9
17      8780        468.6
18      5070        601.5
19      8780        578.8
20      8630        641.5
```

Der Vektor `bankgr` gibt für jede Beobachtung die Nummer der Gruppe an:

```
> bankgr
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2
```

In `S-PLUS` gibt es eine Funktion `discr`, mit der man Fishers lineare Diskriminanzanalyse durchführen kann. Der Aufruf von `discr` ist

```
discr(x, k)
```

In der Datenmatrix `x` sind die Beobachtungen so angeordnet, dass die ersten n_1 Zeilen von `x` die erste Gruppe bilden, und die restlichen Zeilen zur zweiten

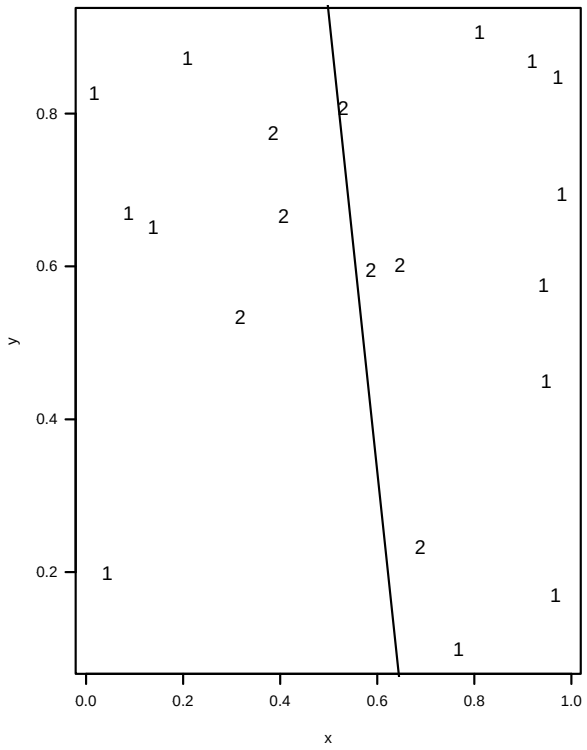


Fig. 12.12. Streudiagramm von 20 artifiziellen Beobachtungen mit der Geraden, die man durch Fishers lineare Diskriminanzanalyse erhält

Gruppe gehören. Das Argument k gibt die Anzahl der Gruppen an, wenn alle Gruppen gleich groß sind. Unterscheiden sich die Größen der Gruppen, so ist `k` ein Vektor, dessen i-te Komponente die Größe der i-ten Gruppe enthält. Das Ergebnis der Funktion `discr` ist eine Liste, deren zweite Komponente für uns wichtig ist. Die zweite Komponente ist eine Matrix. Der erste Spaltenvektor dieser Matrix ist proportional zum Vektor $\mathbf{d}$ in Gleichung (12.31) auf Seite 376. Schauen wir uns dies für das Beispiel an:

```
> n1<-sum(bankgr==1)
> n2<-sum(bankgr==2)
> d<-discr(bank,c(n1,n2))[[2]][,1]
> d
[1] -0.0005552187 -0.0040370515
```

Um ein Objekt klassifizieren zu können, benötigen wir $\bar{\mathbf{x}}_1$ und $\bar{\mathbf{x}}_2$. Den Mittelwert der ersten Gruppe erhält man durch

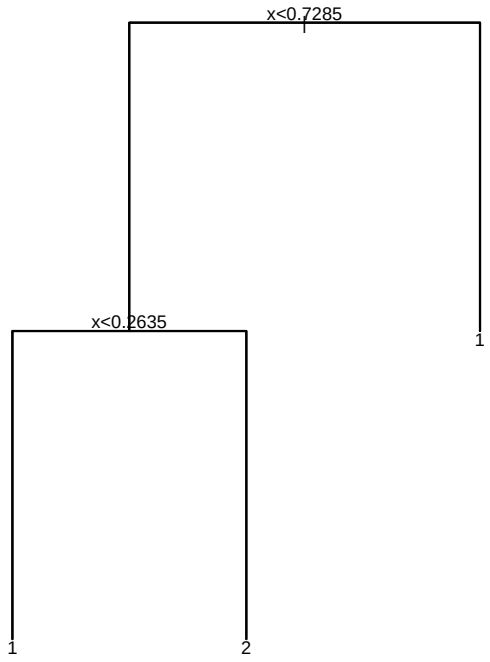


Fig. 12.13. Klassifikationsbaum von 20 artifiziellen Beobachtungen in zwei Gruppen

```
> xq1<-apply(bank[bankgr==1,],2,mean)
> xq1
 Einwohner Gesamtkosten
  3510.429     390.2357
```

und den Mittelwert der zweiten Gruppe durch

```
> xq2<-apply(bank[bankgr==2,],2,mean)
> xq2
 Einwohner Gesamtkosten
  7975.167       577.45
```

Wir ordnen ein Objekt mit Merkmalsvektor $\mathbf{x}$ der Gruppe 1 zu, wenn gilt

$$|\mathbf{d}'\mathbf{x} - \mathbf{d}'\mathbf{x}_1| < |\mathbf{d}'\mathbf{x} - \mathbf{d}'\mathbf{x}_2|.$$

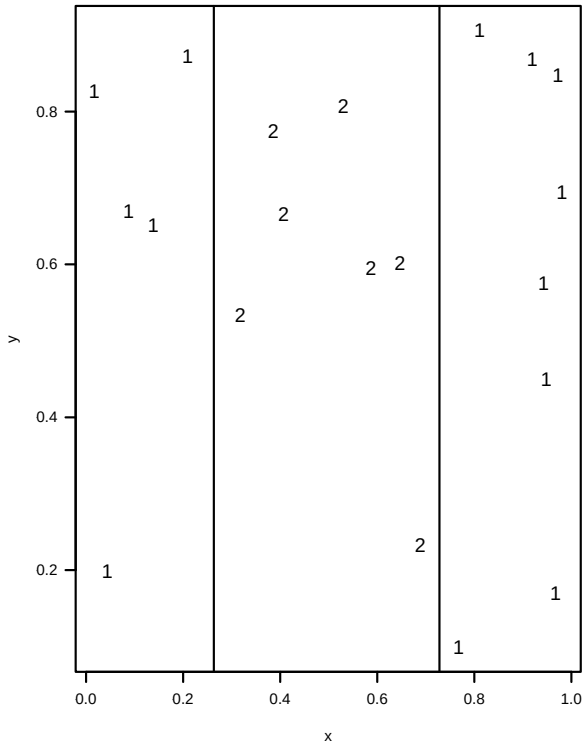


Fig. 12.14. Streudiagramm von 20 artifiziellen Beobachtungen mit Zerlegung der Ebene, die sich aus dem Klassifikationsbaum ergibt

Schauen wir uns dies examplarisch für die erste Filiale an. Wir bestimmen $\mathbf{d}'\mathbf{x}$:

```
> x<-bank[1,]
> dx<-d%*%x
> dx
          [,1]
[1,] -2.842187
```

Wir müssen $\mathbf{d}'\mathbf{x}$ mit $\mathbf{d}'\mathbf{x}_1$ und $\mathbf{d}'\mathbf{x}_2$ vergleichen. Schauen wir uns zunächst deren Werte an:

```
> d%*%xq1
          [,1]
[1,] -3.524457
> d%*%xq2
```

```
           [,1]
[1,] -6.759157
```

Der Ausdruck

```
> abs(dx-d%*%xq1)>abs(dx-d%*%xq2)
```

liefert `F`, wenn die Beobachtung der ersten Gruppe zugeordnet werden soll, und `T`, wenn sie der zweiten Gruppe zugeordnet werden soll. Addieren wir zu diesem Ausdruck den Wert 1, so wird `F` vor der Addition in die 0 und `T` in die 1 konvertiert. Der Ausdruck

```
> 1+(abs(dx-d%*%xq1)>abs(dx-d%*%xq2))
```

liefert also die Gruppe, der die Beobachtung zugeordnet wird. Im Beispiel erhalten wir das Ergebnis

```
     [,1]
[1,]    1
```

Die Resubstitutionsfehlerrate liefert also folgende Befehlsfolge:

```
> yq1<-d%*%xq1
> yq2<-d%*%xq2
> y<-as.vector(bank%*%d)
> g<-1+(abs(y-yq1)>abs(y-yq2))
> g
 [1] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 2 2 2 2 2
> mean(bankgr!=g)
[1] 0.05
```

Wir zeigen nun, wie man die Fehlerrate schätzt, wenn man die Daten in eine Lern- und eine Teststichprobe aufteilt. Wir wählen zunächst aus den natürlichen Zahlen $1, \ldots, 20$ zehn Zahlen zufällig ohne Zurücklegen aus. Dies leistet die Funktion `sample`:

```
> ilern<-sample(20,10)
> ilern
 [1] 20 19  1 16 15 11 17 10  8  6
```

Wir sortieren diese Werte noch:

```
> ilern<-sort(ilern)
> ilern
 [1]  1  6  8 10 11 15 16 17 19 20
```

Die Indizes der Teststichprobe erhalten wir durch

```
> itest<-(1:20)[-ilern]
> itest
 [1]  2  3  4  5  7  9 12 13 14 18
```

Wir bestimmen die Größen der Gruppen in der Lernstichprobe und schätzen den Vektor **d** für die Lernstichprobe:

```
> n1<-sum(bankgr[ilern]==1)
> n1
[1] 5
> n2<-sum(bankgr[ilern]==2)
> n2
[1] 5
> d<-discr(bank[ilern,],c(n1,n2))[[2]][,1]
> d
[1] 0.001063026 0.001758229
```

Anschließend bestimmen wir die Mittelwerte in den Gruppen der Lernstichprobe:

```
> xq1<-apply(bank[ilern,][bankgr[ilern]==1,],2,mean)
> xq1
 Einwohner Gesamtkosten
    3071.4       381.42
> xq2<-apply(bank[ilern,][bankgr[ilern]==2,],2,mean)
> xq2
 Einwohner Gesamtkosten
    8556.2       572.64
```

Nun müssen wir nur doch den Vektor **d** auf jedes Element der Teststichprobe anwenden:

```
> test<-bank[itest,]
> y<-as.vector(test%*%d)
> g<-1+(abs(y-d%*%xq1)>abs(y-d%*%xq2))
> g
 [1] 1 1 1 1 2 1 1 1 1 1
```

Die Schätzung der Fehlerrate ist:

```
> mean(g!=bankgr[itest])
[1] 0.2
```

Um die Leaving-one-out-Methode anzuwenden, muss man die Funktion `discr` auf den Datensatz ohne die i-te Beobachtung anwenden und die i-te Beobachtung klassifizieren. Die folgende Befehlsfolge bestimmt die Schätzung der Fehlerrate mit der Leaving-one-out-Methode.

```
> n1<-sum(bankgr==1)
> n2<-sum(bankgr==2)
> xq1<-apply(bank[bankgr==1,],2,mean)
> xq2<-apply(bank[bankgr==2,],2,mean)
> g<-rep(0,20)
```

```
> for (i in 1:20)
 {if(i<=n1)
    {d<-discr(bank[-i,],c(n1-1,n2))[[2]][,1]
     x1<-(n1*xq1-bank[i,])/(n1-1)
     y<-d%*%bank[i,]
     g[i]<-1+(abs(y-d%*%x1)>abs(y-d%*%xq2))
    }
  else
    {d<-discr(bank[-i,],c(n1,n2-1))[[2]][,1]
     x2<-(n2*xq2-bank[i,])/(n2-1)
     y<-d%*%bank[i,]
     g[i]<-1+(abs(y-d%*%xq1)>abs(y-d%*%x2))
    }
  }
> mean(g!=bankgr)
[1] 0.1
```

Die Indizes der Filialen, die falsch klassifiziert werden, erhalten wir durch

```
> (1:20)[g!=bankgr]
[1]  7 18
```

Schauen wir uns die logistische Diskriminanzanalyse in `S-PLUS` an. Da die logistische Regression ein spezielles verallgemeinertes lineares Modell ist, verwendet man die Funktion `glm`. Wir wollen hier nicht auf alle Aspekte der Funktion `glm` eingehen, sondern nur zeigen, wie man mit ihr eine logistische Regression durchführt. Hierzu gibt man ein

```
glm(formula,family=binomial)
```

Man gibt, wie bei der Regressionsanalyse auf Seite 241 beschrieben wird, die Beziehung als Formel ein. Wir wollen im Beispiel 11 auf Seite 10 den Typ der Filiale auf Basis der Gesamtkosten klassifizieren. Wir erzeugen einen Vektor `typ`, der den Wert 1 annimmt, wenn die erste Gruppe vorliegt. Ansonsten ist er 0.

```
> typ<-2-bankgr
```

Wir rufen dann die Funktion `glm` auf und weisen das Ergebnis der Variablen `e` zu:

```
> e<-glm(typ~bank[,2],family=binomial)
```

Die Koeffizienten des Regressionsmodells (12.33) liefert die Funktion `coefficients`:

```
> coefficients(e)
 (Intercept)  bank[, 2]
    15.21896 -0.0294073
```

Die Resubstitutionsfehlerrate können wir mit Hilfe der Funktion `fitted` bestimmen. Diese liefert die geschätzten Wahrscheinlichkeiten. Ist eine geschätzte Wahrscheinlichkeit größer als 0.5, so ordnen wir die Beobachtung der Gruppe 1 zu, ansonsten der Gruppe 2. Der Befehl

```
> g<-2-(fitted(e)>0.5)
```

liefert die geschätzte Gruppenzugehörigkeit:

```
> g
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
 1 1 1 1 1 1 2 1 1  1  1  1  1  1  1  2  1  2  2  2
```

Die Resubstitutionsfehlerrate erhalten wir durch

```
> mean(g!=bankgr)
[1] 0.15
```

Dies ist im Einklang mit Abbildung 12.6. Dort sind drei Beobachtungen fehlklassifiziert. Schauen wir uns noch die Leaving-one-out-Methode an:

```
> for (i in 1:20)
 {e<-coefficients(glm(typ[-i]~bank[-i,2],family=binomial))
 g[i]<-2-(exp(e[1]+e[2]*bank[i,2])/
      (1+exp(e[1]+e[2]*bank[i,2]))>0.5)
 }
> mean(g!=bankgr)
[1] 0.15
```

Wir wollen für die binären Merkmale in Beispiel 12.1 auf Seite 352 einen Klassifikationsbaum erstellen. Die Daten mögen in den Vektoren `Geschlecht`, `MatheLK` und `Abitur88` stehen:

```
> Geschlecht
 [1] 0 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 0 1 1 1
> MatheLK
 [1] 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
> Abitur88
 [1] 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 1 1 1
```

Die Gruppenzugehörigkeit möge im Vektor `Gruppe` stehen:

```
> Gruppe
 [1] 2 2 2 2 2 2 2 2 1 1 1 1 2 1 1 1 1 2 2 1
```

In `S-PLUS` gibt es die Funktion `tree`, mit der man einen Klassifikationsbaum erstellen kann. Die Entscheidungen beruhen in dieser Funktion auf der Devianz. Vor dem Aufruf von `tree` muss man aus der Variablen `Gruppe` einen Faktor machen.

```
> Gruppe<-factor(Gruppe)
```

Die Funktion `tree` wird folgendermaßen aufgerufen:

```
tree(formula, data=<<see below>>, weights=<<see below>>,
      subset=<<see below>>, na.action=na.fail,
      method="recursive.partition", control=<<see below>>,
      model=NULL, x=F, y=T, ...)
```

Für uns sind die Argumente `formula` und `control` wichtig. Man gibt wie bei der Regressionsanalyse auf Seite 241 die Beziehung als Formel ein. Für das Beispiel heißt dies:

```
> e<-tree(Gruppe~Geschlecht+MatheLK+Abitur88)
```

Schauen wir uns `e` an:

```
> e
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 20 27.530 2 ( 0.4500 0.5500 )
  2) MatheLK<0.5 9  6.279 2 ( 0.1111 0.8889 ) *
  3) MatheLK>0.5 11 12.890 1 ( 0.7273 0.2727 )
    6) Geschlecht<0.5 5  5.004 1 ( 0.8000 0.2000 ) *
    7) Geschlecht>0.5 6  7.638 1 ( 0.6667 0.3333 ) *
```

Jede Zeile enthält Informationen über einen Knoten. Nach dem Namen des Knotens folgen die Anzahl der Beobachtungen im Knoten, der Wert der Devianz und die Gruppe, der ein Objekt zugeordnet wird, wenn dieser Knoten ein Endknoten ist oder wäre. Als letzte Informationen stehen in runden Klammern die geschätzten Wahrscheinlichkeiten der beiden Gruppen in diesem Knoten. So ist der erste Knoten der Wurzelknoten. Er enthält 20 Beobachtungen und die Devianz beträgt 27.53. Ein Objekt würde der Gruppe 2 zugeordnet. Die geschätzte Wahrscheinlichkeit von Gruppe 1 beträgt in diesem Knoten 0.45. Die geschätzte Wahrscheinlichkeit von Gruppe 2 beträgt in diesem Knoten 0.55. Die folgende Befehlsfolge zeichnet den Baum, der in Abbildung 12.15 zu finden ist:

```
> plot(e,type="u")
> text(e)
```

Um den vollständigen Klassifikationsbaum in Abbildung 12.7 auf Seite 382 zu erstellen, benötigen wir das Argument `control`. Der Aufruf

```
> e<-tree(gruppe~Geschlecht+MatheLK+Abitur88,
          control=tree.control(nobs=20,minsize=1))
> plot(e,type="u")
> text(e)
```

erstellt diesen Baum.

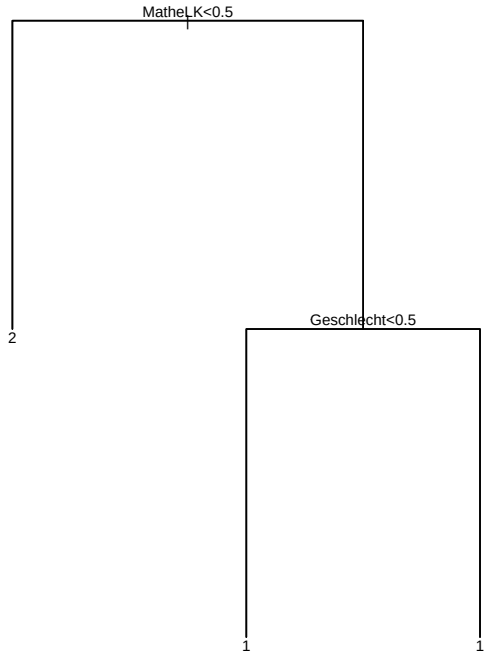


Fig. 12.15. Ein Klassifikationsbaum

12.8 Ergänzungen und weiterführende Literatur

Wir haben in diesem Kapitel nur den Fall betrachtet, dass ein Objekt einer von zwei Gruppen zugeordnet werden soll. Die Vorgehensweise ist nahezu identisch, wenn mehr als zwei Gruppen betrachtet werden. Die Details sind bei Fahrmeir et al. (1996), Huberty (1994) und McLachlan (1992) zu finden. Die beiden letztgenannten Bücher beschäftigen sich ausschließlich mit der Diskriminanzanalyse. Hier sind auch andere Verfahren zur Schätzung der Fehlerrate zu finden. Einen hervorragenden Überblick über die Schätzung der Fehlerrate liefert Hand (1997). Neben den hier betrachteten Verfahren werden neuronale Netze zur Klassifikation benutzt. Smith (1993) gibt eine einfache Einführung in die Theorie und Praxis neuronaler Netze. Umfassende Einführungen in die Diskriminanzanalyse unter Berücksichtigung moderner Verfahren liefern Ripley (1996) und Hastie et al. (2001).

12.9 Übungen

Exercise 31. Es werden zwei Merkmale bei jeweils drei Objekten in zwei Gruppen beobachtet. Die Datenmatrizen $\mathbf{X}_1$ und $\mathbf{X}_2$ enthalten die Merkmalsausprägungen in den Gruppen.

$$\mathbf{X}_1 = \begin{pmatrix} 2 & 5 \\ 1 & 7 \\ 3 & 6 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 5 & 1 \\ 6 & 3 \\ 7 & 2 \end{pmatrix}.$$

1. Erstellen Sie das Streudiagramm aller Beobachtungen.
2. Bestimmen Sie die Entscheidungsregel von Fishers linearer Diskriminanzanalyse für die Daten.
3. Zeichnen Sie die Gerade ein, die sich auf Grund Fishers linearer Diskriminanzanalyse ergibt.
4. Bestimmen Sie die Resubstitutionsfehlerrate.
5. Bestimmen Sie die Schätzung der Fehlerrate mit der Leaving-one-out-Methode unter Verwendung von `S-PLUS`.

Exercise 32. Betrachten Sie das Beispiel 52 auf Seite 389 und wählen Sie die Lern- und Teststichprobe wie auf Seite 389. Es soll die lineare Diskriminanzanalyse von Fisher durchgeführt werden.

1. Verwenden Sie zunächst zur Klassifizierung das Merkmal `Gesamtkosten`.
   a) Schätzen Sie die Entscheidungsregel auf Basis der Lernstichprobe.
   b) Schätzen Sie die Fehlerrate auf Basis der Teststichprobe.
2. Verwenden Sie nun das Merkmal `Einwohner` zur Klassifizierung.
   a) Schätzen Sie die Entscheidungsregel auf Basis der Lernstichprobe.
   b) Schätzen Sie die Fehlerrate auf Basis der Teststichprobe.
3. Verwenden Sie beide Merkmale zur Klassifizierung.
   a) Schätzen Sie die Entscheidungsregel auf Basis der Lernstichprobe.
   b) Schätzen Sie die Fehlerrate auf Basis der Teststichprobe.

Exercise 33. Betrachten Sie das Beispiel 45 auf Seite 351. Ein Student soll der Gruppe, die den Test besteht, oder der Gruppe, die den Test nicht besteht, zugeordnet werden.

1. Verwenden Sie zunächst nur das Merkmal `MatheNote`.
   a) Geben Sie die Klassifikationsregel von Fisher für dieses Beispiel an.
   b) Bestimmen Sie die Klassifikationsregel der logistischen Diskriminanzanalyse mit `S-PLUS`.
   c) Bestimmen Sie die Fehlerrate mit der Resubstitutionsmethode und der Leaving-one-out-Methode.
2. Verwenden Sie nun die Merkmale `Geschlecht`, `MatheLK`, `MatheNote` und `Abitur88`. Führen Sie die folgenden Aufgaben in `S-PLUS` durch.
   a) Geben Sie die Klassifikationsregel von Fisher für dieses Beispiel an.

b) Bestimmen Sie die Klassifikationsregel der logistischen Diskriminanzanalyse.
c) Bestimmen Sie die Fehlerrate mit der Resubstitutionsmethode und der Leaving-one-out-Methode.

Exercise 34. Im Wintersemester 2000/2001 wurden Studienanfänger unter anderem danach gefragt, ob sie noch bei den Eltern wohnen. Wir bezeichnen dieses Merkmal mit `Eltern`. Außerdem wurden die Merkmale `Geschlecht`, `Studienfach` und `Berufsausbildung` erhoben. Beim Merkmal `Studienfach` wurden sie gefragt, ob sie BWL studieren. Beim Merkmal `Berufsausbildung` wurde gefragt, ob sie nach dem Abitur eine Berufsausbildung gemacht haben. In Tabelle 12.9 sind die Ergebnisse der Befragung von 20 zufällig ausgewählten Studenten zu finden.

Table 12.9. Ergebnisse einer Befragung von 20 Studenten

Geschlecht	Studienfach	Berufsausbildung	Eltern
w	j	n	n
m	n	n	j
m	j	n	j
m	j	n	n
w	j	n	n
w	j	j	n
w	n	n	n
m	j	j	n
m	j	n	j
m	n	n	n
w	j	n	n
m	n	n	n
m	j	n	j
m	j	n	n
w	n	n	j
w	j	n	j
m	j	j	j
m	j	n	j
w	j	j	j
w	j	j	j

Es soll eine Regel angegeben werden, die einen Studierenden auf Basis der Merkmale `Geschlecht`, `Studienfach` und `Berufsausbildung` einer der beiden Kategorien des Merkmals `Eltern` zuordnet.

1. Bestimmen Sie die Regel, die sich auf Grund Fishers linearer Diskriminanzanalyse ergibt. Bestimmen Sie die Resubstitutionsfehlerrate.
2. Erstellen sie den Klassifikationsbaum.

3. Führen Sie mit `S-PLUS` eine logistische Diskriminanzanalyse durch.

13 Clusteranalyse

13.1 Problemstellung

In den letzten beiden Kapiteln haben wir Gesamtheiten betrachtet, die aus Gruppen bestehen. Dabei war die Gruppenstruktur bekannt. In diesem Kapitel werden wir uns mit Verfahren beschäftigen, mit denen man in einem Datensatz Gruppen von Beobachtungen finden kann. Ausgangspunkt sind die Ausprägungen quantitativer Merkmale bei den Objekten $O = \{O_1, \ldots, O_n\}$ oder eine Distanzmatrix der Objekte. Gesucht ist eine *Partition* dieser Objekte. Unter einer Partition einer Menge $O = \{O_1, \ldots, O_n\}$ versteht man eine Zerlegung in Teilmengen $C_1, \ldots, C_k$, sodass jedes Element von O zu genau einer Teilmenge C_i für $i = 1, \ldots, k$ gehört. Diese Teilmengen bezeichnet man auch als *Klassen*.

Example 55. Sei $O = \{1, 2, 3, 4, 5, 6\}$. Dann ist

$$C_1 = \{1, 2, 4, 5\},$$
$$C_2 = \{3, 6\}$$

eine Partition von O. □

Die Objekte innerhalb einer Klasse sollen ähnlich sein, während die Klassen sich unterscheiden. Man spricht davon, dass die Klassen intern kohärent, aber extern isoliert sind. Liegt nur ein quantitatives Merkmal vor, so kann man mit Hilfe einer Abbildung leicht feststellen, ob man die Menge der Objekte in Klassen zerlegen kann.

Example 56. Das Alter von 6 Personen beträgt

```
43 38  6 47 37  9.
```

Stellen wir die Werte auf dem Zahlenstrahl dar, so können wir zwei Klassen erkennen, die isoliert und kohärent sind:

x x xx x x

1

Für die Gruppenstruktur gibt es eine einfache Erklärung. Es handelt sich um zwei Ehepaare, die jeweils ein Kind haben. □

Wurden mehrere quantitative Merkmale erhoben, oder liegt eine Distanzmatrix vor, so ist nicht so einfach zu erkennen, ob Klassen vorliegen. hmcounterend. (fortgesetzt)

Example 56. Wir bestimmen die euklidische Distanz zwischen jeweils zwei Personen und stellen die Distanzen in einer Distanzmatrix dar. Diese ist

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}. \tag{13.1}$$

□

Wir werden im Folgenden Verfahren kennenlernen, mit denen man Klassen entdecken kann.

13.2 Hierarchische Clusteranalyse

13.2.1 Theorie

Wir werden uns in diesem Abschnitt mit *hierarchischen Clusterverfahren* beschäftigen. Diese erzeugen eine Folge $P_n, P_{n-1}, \ldots, P_2, P_1$ bzw. eine Folge $P_1, P_2, \ldots, P_{n-1}, P_n$ von Partitionen der Menge $O = \{O_1, \ldots, O_n\}$, wobei die Partition P_i aus i Klassen besteht. Die Partitionen P_g und P_{g+1} haben $g-1$ Klassen gemeinsam. Beginnt man mit n Klassen, so spricht man von einem *agglomerativen* Verfahren, während es sich um ein *divisives* Verfahren handelt, wenn man mit einer Klasse beginnt.

Example 57. Sei $O = \{1, 2, 3, 4, 5, 6\}$. Dann bildet

$$\begin{aligned} P_6 &= \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}, \\ P_5 &= \{\{1\}, \{3\}, \{4\}, \{6\}, \{2, 5\}\}, \\ P_4 &= \{\{1\}, \{4\}, \{3, 6\}, \{2, 5\}\}, \\ P_3 &= \{\{1, 4\}, \{3, 6\}, \{2, 5\}\}, \\ P_2 &= \{\{1, 2, 4, 5\}, \{3, 6\}\}, \\ P_1 &= \{1, 2, 3, 4, 5, 6\} \end{aligned}$$

eine Folge von Partitionen, die durch ein agglomeratives Verfahren entstehen. Die Partitionen P_3 und P_4 haben die Klassen $\{3, 6\}$ und $\{2, 5\}$ gemeinsam. □

Zu jeder Partition gehört eine Distanz, bei der die Partition gebildet wurde. Diese Distanz hängt natürlich von dem Verfahren ab, das bei der Bildung der Partitionen verwendet wurde. hmcounterend. (fortgesetzt)

Example 57. Tabelle 13.1 gibt zu jeder Partition die Distanz an. Wir werden später sehen, wie die Distanzen gewonnen wurden.

Table 13.1. Partitionen und zugehörige Distanzen

Partition	Distanz
$\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$	0
$\{\{1\}, \{3\}, \{4\}, \{6\}, \{2, 5\}\}$	1
$\{\{1\}, \{4\}, \{3, 6\}, \{2, 5\}\}$	3
$\{\{1, 4\}, \{3, 6\}, \{2, 5\}\}$	4
$\{\{1, 2, 4, 5\}, \{3, 6\}, \}$	10
$\{1, 2, 3, 4, 5, 6\}$	41

□

Die Partitionen und zugehörigen Distanzen stellt man in einem *Dendrogramm* dar. In einem rechtwinkligen Koordinatensystem werden auf der Ordinate die Distanzen abgetragen. Zu jedem Objekt gehört eine senkrechte Linie, die von unten nach oben so weit abgetragen wird, bis das Objekt zum ersten Mal mit mindestens einem anderen Objekt in einer Klasse ist. Die Linien der Objekte werden durch eine waagerechte Linie verbunden und durch eine senkrechte Linie ersetzt. Diese wird so lange nach oben verlängert, bis die Klasse der Objekte zum ersten Mal mit einem anderen Objekt oder einer anderen Klasse verschmolzen wird. Dieser Prozess wird so lange fortgesetzt, bis alle Objekte in einer Klasse sind. hmcounterend. (fortgesetzt)

Example 57. Abbildung 13.1 zeigt das Dendrogramm.

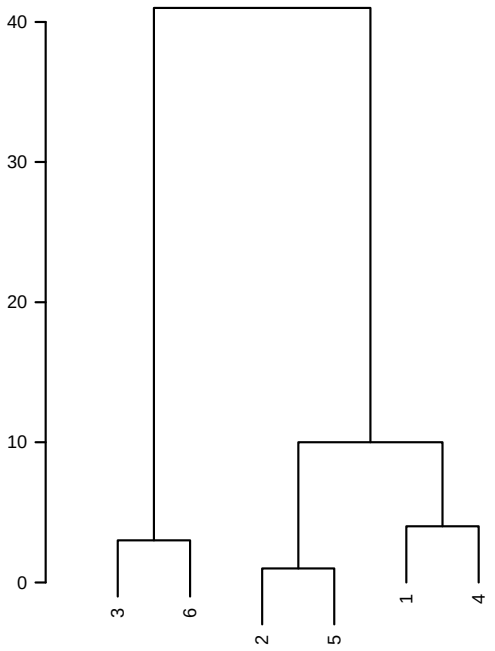


Fig. 13.1. Ein Dendrogramm

Am Dendrogramm kann man den Prozess der Klassenbildung erkennen. So werden zunächst die Objekte 2 und 5 beim Abstand 1 zu einer Klasse $\{2, 5\}$

verschmolzen. Dann werden beim Abstand 3 die Objekte 3 und 6 zur Klasse $\{3,6\}$ verschmolzen. Als nächstes werden beim Abstand 4 die Objekte 1 und 4 zur Klasse $\{1,4\}$ verschmolzen. Dann werden beim Abstand 10 die Klassen $\{1,4\}$ und $\{2,5\}$ zur Klasse $\{1,2,4,5\}$ verschmolzen. Im letzten Schritt werden beim Abstand 41 die Klassen $\{1,2,4,5\}$ und $\{3,6\}$ zur Klasse $\{1,2,3,4,5,6\}$ verschmolzen. In Abhängigkeit vom Abstand d erhalten wir somit folgende Partitionen:

$$\begin{array}{ll}
0 \leq d < 1 & \{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\}, \\
1 \leq d < 3 & \{\{1\},\{3\},\{4\},\{6\},\{2,5\}\}, \\
3 \leq d < 4 & \{\{1\},\{4\},\{3,6\},\{2,5\}\}, \\
4 \leq d < 10 & \{\{1,4\},\{3,6\},\{2,5\}\}, \\
10 \leq d < 41 & \{\{1,2,4,5\},\{3,6\},\}, \\
41 \leq d & \{1,2,3,4,5,6\}.
\end{array}$$

□

Man erhält aus dem Dendrogramm eine aus k Klassen bestehende Partition, indem man das Dendrogramm horizontal in einer Höhe durchschneidet, in der k senkrechte Linien verlaufen.

Zu jedem Dendrogramm gehört eine Abstandsmatrix D^*, bei der der Abstand zwischen zwei Objekten i und j durch die Höhe im Dendrogramm gegeben ist, auf der diese beiden Objekte zum ersten Mal in einer Klasse liegen. hmcounterend. (fortgesetzt)

Example 57. So ist der kleinste Abstand, bei dem die Objekte 1 und 2 in einer Klasse sind, gleich 10. Bestimmt man diesen Abstand für alle Paare von Objekten, so erhält man folgende Distanzmatrix:

$$\mathbf{D}^* = \begin{pmatrix}
0 & 10 & 41 & 4 & 10 & 41 \\
10 & 0 & 41 & 10 & 1 & 41 \\
41 & 41 & 0 & 41 & 41 & 3 \\
4 & 10 & 41 & 0 & 10 & 41 \\
10 & 1 & 41 & 10 & 0 & 41 \\
41 & 41 & 3 & 41 & 41 & 0
\end{pmatrix}.$$

□

Man nennt die aus dem Dendrogramm gewonnene Distanzmatrix auch *kophenetische Matrix.*

Wie kann man aus einer Distanzmatrix ein Dendrogramm gewinnen? Um diese Frage zu beantworten, betrachten wir die agglomerative Vorgehensweise. Schauen wir uns zunächst den Fall an, dass die Distanzmatrix aus dem Dendrogramm abgeleitet wurde. Wir gehen also von der Matrix D^* aus. Wir betrachten nur den Teil der Distanzmatrix, der unterhalb der Hauptdiagonalen liegt. hmcounterend. (fortgesetzt)

Example 57. Wir betrachten also

$$\mathbf{D}^* = \begin{pmatrix} & & & & \\ 10 & & & & \\ 41 & 41 & & & \\ 4 & 10 & 41 & & \\ 10 & 1 & 41 & 10 & \\ 41 & 41 & 3 & 41 & 41 \end{pmatrix}.$$

□

Bei n Objekten liegen auf der ersten Stufe n Klassen vor. Es liegt nahe, die beiden Objekte zu einer Klasse zu verschmelzen, deren Abstand am kleinsten ist. Wir wählen aus der Distanzmatrix die kleinste Zahl aus und verschmelzen die beiden zugehörigen Objekte. hmcounterend. (fortgesetzt)

Example 57. Im Beispiel ist der kleinste Abstand die 1. Zu ihm gehören die Objekte 2 und 5, sodass beim Abstand 1 die Objekte 2 und 5 verschmolzen werden. Nun haben wir es nicht mehr mit den ursprünglichen Objekten zu tun, sondern mit den Objekten $1, 3, 4, 6$ und der Klasse $\{2, 5\}$. Wir stellen diese in Tabelle 13.2 zusammen.

Table 13.2. Vorläufiges Ergebnis des ersten Schritts eines agglomerativen Verfahrens

	$\{2,5\}$	1	3	4	6
$\{2,5\}$					
1					
3		41			
4		4	41		
6		41	3	41	

In dieser Tabelle fehlen einige Zahlen, da zunächst offen ist, wie groß der Abstand zwischen den Objekten 1, 3, 4, 6 und der Klasse $\{2, 5\}$ ist. Es ist naheliegend, diesen Abstand auf Basis der Abstände der Elemente der Klasse $\{2, 5\}$ zu den restlichen Objekten zu ermitteln. Der Abstand der Klasse $\{2, 5\}$ zum Objekt 1 sollte also auf dem Abstand zwischen den Objekten 2 und 1 und dem Abstand zwischen den Objekten 5 und 1 beruhen. Dieser beträgt in beiden Fällen 10. Somit wählen wir als Abstand zwischen der Klasse $\{2, 5\}$ und dem Objekt 1 den Wert 10. Als Abstand zwischen der Klasse $\{2, 5\}$ und dem Objekt 3 erhalten wir den Wert 41. Als Abstand zwischen der Klasse $\{2, 5\}$ und dem Objekt 4 erhalten wir den Wert 10. Als Abstand zwischen der Klasse $\{2, 5\}$ und dem Objekt 6 erhalten wir den Wert 41. Wir erhalten somit Tabelle 13.3.

□

Nachdem wir die erste Klasse gebildet haben, gehen wir genauso wie oben vor und suchen das kleinste Element der Tabelle und verschmelzen die beiden

Table 13.3. Ergebnis des ersten Schritts eines agglomerativen Verfahrens

	$\{2,5\}$	1	3	4	6
$\{2,5\}$					
1	10				
3	41	41			
4	10	4	41		
6	41	41	3	41	

Klassen. Diesen Prozess führen wir so lange durch, bis alle Objekte in einer Klasse sind. hmcounterend. (fortgesetzt)

Example 57. Die kleinste Zahl in Tabelle 13.3 ist die 3. Wir verschmelzen somit die Objekte 3 und 6 zur Klasse $\{3,6\}$ und erhalten Tabelle 13.4.

Table 13.4. Ergebnis des zweiten Schritts eines agglomerativen Verfahrens

	$\{2,5\}$	1	$\{3,6\}$	4
$\{2,5\}$				
1	10			
$\{3,6\}$	41	41		
4	10	4	41	

Die kleinste Zahl in Tabelle 13.4 ist die 4. Somit verschmelzen wir die Objekte 1 und 4 zur Klasse $\{1,4\}$ und erhalten Tabelle 13.5.

Table 13.5. Ergebnis des dritten Schritts eines agglomerativen Verfahrens

	$\{2,5\}$	$\{1,4\}$	$\{3,6\}$
$\{2,5\}$			
$\{1,4\}$	10		
$\{3,6\}$	41	41	

Die kleinste Zahl in Tabelle 13.5 ist die 10. Somit verschmelzen wir die Klassen $\{1,4\}$ und $\{2,5\}$ zur Klasse $\{1,2,4,5\}$ und erhalten Tabelle 13.6.

Im letzten Schritt werden die Klassen $\{2,4\}$ und $\{1,3,5\}$ verschmolzen. Stellt man den Verschmelzungsvorgang graphisch dar, so erhält man das Dendrogramm aus Abbildung 13.1. □

Ist d_{ij} die Distanz zwischen den Objekten i und j und $D_{i.j}$ die Distanz zwischen der i-ten und j-ten Klasse, dann kann man die obige Vorgehensweise folgendermaßen beschreiben:

Table 13.6. Ergebnis des vierten Schritts eines agglomerativen Verfahrens

	$\{1,2,4,5\}$	$\{3,6\}$
$\{1,2,4,5\}$		
$\{3,6\}$	41	

1. Definiere jedes Objekt als eigene Klasse, d.h. setze $D_{i.j} = d_{ij}$.
2. Bestimme
$$min\,\{D_{i.j}|D_{i.j} > 0\}.$$
Wähle einen Wert zufällig aus, falls mehrere Werte gleich sind. Sei $D_{k.m}$ das kleinste Element. Verschmelze die Klassen k und m.
3. Bestimme den Abstand zwischen der neu gewonnenen Klasse und den restlichen Klassen. Ersetze die k-te und m-te Zeile und Spalte von D durch diese Zahlen.
4. Wiederhole die Schritte 2 und 3, bis nur noch eine Klasse vorliegt.

Der dritte Schritt ist nicht für jede Distanzmatrix eindeutig definiert. Werden zwei Klassen C_i und C_j zu einer Klasse verschmolzen, so kann die Distanz einer anderen Klasse C_k zu den beiden Klassen C_i und C_j unterschiedlich sein. In diesem Fall ist nicht klar, was die Distanz der aus C_i und C_j gebildeten Klasse zur Klasse C_k ist. hmcounterend. (fortgesetzt)

Example 56. Wir betrachten die Distanzmatrix des Alters der 6 Personen in Gleichung (13.1) auf Seite 408. Die kleinste Zahl unterhalb der Hauptdiagonalen ist die 1. Im ersten Schritt verschmelzen wir die Klassen $\{2\}$ und $\{5\}$. Nun gilt $D_{\{1\}.\{2\}} = 5$ und $D_{\{1\}.\{5\}} = 6$. □

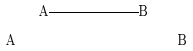
Die einzelnen hierarchischen Clusteranalyseverfahren unterscheiden sich nun dadurch, wie dieser Abstand definiert ist. Beim *Single-Linkage-Verfahren* nimmt man die kleinere, beim *Complete-Linkage-Verfahren* die größere der beiden Zahlen, während man beim *Average-Linkage-Verfahren* den Mittelwert der beiden Zahlen wählt. Formal können wir dies folgendermaßen beschreiben:

Wir betrachten die i-te, j-te und k-te Klasse. Seien $D_{i.j}$, $D_{i.k}$ und $D_{j.k}$ gegeben. Sei $D_{i.j}$ der kleinste Wert in der Distanzmatrix. Also werden die Klassen i und j verschmolzen. Gesucht ist $D_{ij.k}$. Die drei Verfahren gehen folgendermaßen vor:

1. Single-Linkage-Verfahren mit $D_{ij.k} = \min\{D_{i.k}, D_{j.k}\}$,
2. Complete-Linkage-Verfahren mit $D_{ij.k} = \max\{D_{i.k}, D_{j.k}\}$,
3. Average-Linkage-Verfahren mit dem Mittelwert der Distanzen zwischen allen Elementen der beiden Klassen.

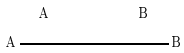
Die Verfahren unterscheiden sich dadurch, wie sie den Abstand zwischen zwei Klassen bestimmen. Beim Single-Linkage-Verfahren ist der Abstand zwischen

den Klassen A und B der kleinste Abstand zwischen Punkten der einen und Punkten der anderen Klasse. Das folgende Bild verdeutlicht dies:



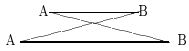
1

Beim Complete-Linkage-Verfahren ist der Abstand zwischen den Klassen A und B der größte Abstand zwischen Punkten der einen und Punkten der anderen Klasse. Das folgende Bild verdeutlicht dies:



1

Beim Average-Linkage-Verfahren ist der Abstand zwischen den Klassen A und B der Mittelwert aller Abstände zwischen Punkten der einen und Punkten der anderen Klasse. Das folgende Bild verdeutlicht dies:



1

Im nächsten Abschnitt illustrieren wir die Vorgehensweise der Verfahren am Beispiel der Distanzmatrix in Gleichung (13.1) auf Seite 408.

13.2.2 Verfahren der hierarchischen Clusterbildung

Das Single-Linkage-Verfahren Beim Single-Linkage-Verfahren wird als Distanz von zwei zu verschmelzenden Klassen die kleinste Distanz zwischen Elementen der einen Klasse und Elementen der anderen Klasse gewählt. hmcounterend. (fortgesetzt)

Example 56. Wir führen eine hierarchische Clusteranalyse mit dem Single-Linkage-Verfahren auf Basis der Distanzmatrix in (13.1) durch. Die Distanzmatrix ist

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}.$$

Das kleinste Element ist die 1. Im ersten Schritt verschmelzen wir die Klassen $\{2\}$ und $\{5\}$. Es gilt

$$\begin{aligned} D_{\{2,5\}.\{1\}} &= \min\{D_{\{2\}.\{1\}}, D_{\{5\}.\{1\}} = \min\{5,6\} = 5, \\ D_{\{2,5\}.\{3\}} &= \min\{D_{\{2\}.\{3\}}, D_{\{5\}.\{3\}}\} = \min\{32,31\} = 31, \\ D_{\{2,5\}.\{4\}} &= \min\{D_{\{2\}.\{4\}}, D_{\{5\}.\{4\}}\} = \min\{9,10\} = 9, \\ D_{\{2,5\}.\{6\}} &= \min\{D_{\{2\}.\{6\}}, D_{\{5\}.\{6\}}\} = \min\{29,28\} = 28. \end{aligned}$$

Wir erhalten somit die Tabelle 13.7.

Table 13.7. Ergebnis des ersten Schritts des Single-Linkage-Verfahrens

	$\{2,5\}$	$\{1\}$	$\{3\}$	$\{4\}$	$\{6\}$
$\{2,5\}$					
$\{1\}$	5				
$\{3\}$	31	41			
$\{4\}$	9	4	41		
$\{6\}$	28	41	3	41	

Die kleinste Zahl ist die 3, sodass die Klassen $\{3\}$ und $\{6\}$ verschmolzen werden. Es gilt

$$\begin{aligned} D_{\{3,6\}.\{2,5\}} &= \min\{D_{\{3\}.\{2,5\}}, D_{\{6\}.\{2,5\}}\} = \min\{31,28\} = 28, \\ D_{\{3,6\}.\{1\}} &= \min\{D_{\{3\}.\{1\}}, D_{\{6\}.\{1\}}\} = \min\{41,41\} = 41, \\ D_{\{3,6\}.\{4\}} &= \min\{D_{\{3\}.\{4\}}, D_{\{6\}.\{4\}}\} = \min\{41,41\} = 41. \end{aligned}$$

Somit ergibt sich Tabelle 13.8.

Table 13.8. Ergebnis des zweiten Schritts des Single-Linkage-Verfahrens

	$\{2,5\}$	$\{1\}$	$\{3,6\}$	$\{4\}$
$\{2,5\}$				
$\{1\}$	5			
$\{3,6\}$	28	41		
$\{4\}$	9	4	41	

Die kleinste Zahl ist die 4, sodass die Klassen $\{1\}$ und $\{4\}$ verschmolzen werden. Es gilt

$$D_{\{1,4\}.\{2,5\}} = \min\{D_{\{1\}.\{2,5\}}, D_{\{4\}.\{2,5\}}\} = \min\{5,9\} = 5,$$
$$D_{\{1,4\}.\{3,6\}} = \min\{D_{\{1\}.\{3,6\}}, D_{\{4\}.\{3,6\}}\} = \min\{41,41\} = 41.$$

Somit ergibt sich Tabelle 13.9.

Table 13.9. Ergebnis des dritten Schritts des Single-Linkage-Verfahrens

	$\{2,5\}$	$\{1,4\}$	$\{3,6\}$
$\{2,5\}$			
$\{1,4\}$	5		
$\{3,6\}$	28	41	

Die kleinste Zahl ist die 5, sodass die Klassen $\{1,4\}$ und $\{2,5\}$ verschmolzen werden. Es gilt

$$D_{\{1,2,4,5\}.\{3,6\}} = \min\{D_{\{1,4\}.\{3,6\}}, D_{\{2,5\}.\{3,6\}}\} = \min\{41,28\} = 28.$$

Somit ergibt sich Tabelle 13.10.

Table 13.10. Ergebnis des vierten Schritts des Single-Linkage-Verfahrens

	$\{1,2,4,5\}$	$\{3,6\}$
$\{1,2,4,5\}$		
$\{3,6\}$	28	

Somit werden beim Abstand 28 die Klassen $\{1,2,4,5\}$ und $\{3,6\}$ verschmolzen.

Wir erhalten das Dendrogramm in Abbildung 13.2.

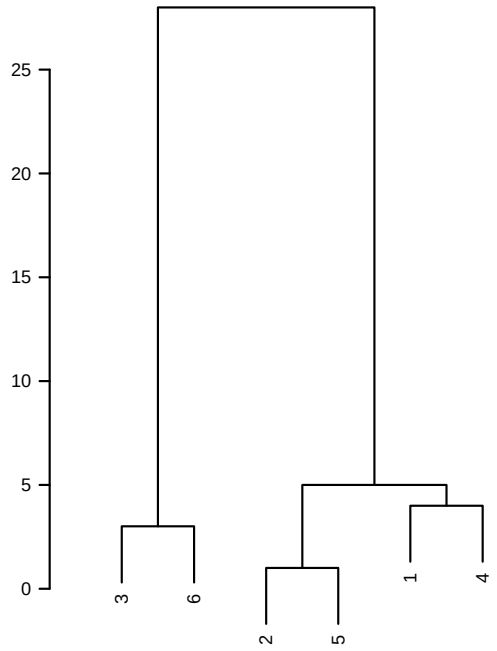


Fig. 13.2. Das Dendrogramm des Single-Linkage-Verfahrens

Die kophenetische Matrix lautet

$$\mathbf{D}^* = \begin{pmatrix} 0 & 5 & 28 & 4 & 5 & 28 \\ 5 & 0 & 28 & 5 & 1 & 28 \\ 28 & 28 & 0 & 28 & 28 & 3 \\ 4 & 5 & 28 & 0 & 5 & 28 \\ 5 & 1 & 28 & 5 & 0 & 28 \\ 28 & 28 & 3 & 28 & 28 & 0 \end{pmatrix}. \tag{13.2}$$

□

Das Complete-Linkage-Verfahren Als Distanz von zwei zu verschmelzenden Klassen wird beim Complete-Linkage-Verfahren die größte Distanz zwischen Elementen der einen Klasse und Elementen der anderen Klasse gewählt.

hmcounterend. (fortgesetzt)

Example 56. Wir führen eine hierarchische Clusteranalyse mit dem Complete-Linkage-Verfahren auf Basis der Distanzmatrix in Gleichung (13.1) auf Seite 408 durch. Das kleinste Element ist die 1.

Im ersten Schritt verschmelzen wir die Klassen $\{2\}$ und $\{5\}$. Es gilt

$$D_{\{2,5\}.\{1\}} = \max\{D_{\{2\}.\{1\}}, D_{\{5\}.\{1\}}\} = \max\{5,6\} = 6,$$

$$D_{\{2,5\}.\{3\}} = \max\{D_{\{2\}.\{3\}}, D_{\{5\}.\{3\}}\} = \max\{32,31\} = 32,$$

$$D_{\{2,5\}.\{4\}} = \max\{D_{\{2\}.\{4\}}, D_{\{5\}.\{4\}}\} = \max\{9,10\} = 10,$$

$$D_{\{2,5\}.\{6\}} = \max\{D_{\{2\}.\{6\}}, D_{\{5\}.\{6\}}\} = \max\{29,28\} = 29.$$

Wir erhalten somit Tabelle 13.11.

Table 13.11. Ergebnis des ersten Schritts des Complete-Linkage-Verfahrens

	$\{2,5\}$	$\{1\}$	$\{3\}$	$\{4\}$	$\{6\}$
$\{2,5\}$					
$\{1\}$	6				
$\{3\}$	32	41			
$\{4\}$	10	4	41		
$\{6\}$	29	41	3	41	

Die kleinste Zahl ist die 3, sodass die Klassen $\{3\}$ und $\{6\}$ verschmolzen werden. Es gilt

$$D_{\{3,6\}.\{2,5\}} = \max\{D_{\{3\}.\{2,5\}}, D_{\{6\}.\{2,5\}}\} = \max\{32, 29\} = 32,$$
$$D_{\{3,6\}.\{1\}} = \max\{D_{\{3\}.\{1\}}, D_{\{6\}.\{1\}}\} = \max\{41, 41\} = 41,$$
$$D_{\{3,6\}.\{4\}} = \max\{D_{\{3\}.\{4\}}, D_{\{6\}.\{4\}}\} = \max\{41, 41\} = 41.$$

Somit ergibt sich Tabelle 13.12.

Table 13.12. Ergebnis des zweiten Schritts des Complete-Linkage-Verfahrens

	$\{2,5\}$	$\{1\}$	$\{3,6\}$	$\{4\}$
$\{2,5\}$				
$\{1\}$	6			
$\{3,6\}$	32	41		
$\{4\}$	10	4	41	

Die kleinste Zahl ist die 4, sodass die Klassen $\{1\}$ und $\{4\}$ verschmolzen werden. Es gilt

$$D_{\{1,4\}.\{2,5\}} = \max\{D_{\{1\}.\{2,5\}}, D_{\{4\}.\{2,5\}}\} = \max\{6, 10\} = 10,$$
$$D_{\{1,4\}.\{3,6\}} = \max\{D_{\{1\}.\{3,6\}}, D_{\{4\}.\{3,6\}}\} = \max\{41, 41\} = 41.$$

Somit ergibt sich Tabelle 13.13.

Die kleinste Zahl ist die 10, sodass die Klassen $\{1,4\}$ und $\{2,5\}$ verschmolzen werden. Es gilt

$$D_{\{1,2,4,5\}.\{3,6\}} = \max\{D_{\{1,4\}.\{3,6\}}, D_{\{2,5\}.\{3,6\}}\} = \max\{41, 32\} = 41.$$

Somit ergibt sich Tabelle 13.14.

Table 13.13. Ergebnis des dritten Schritts des Complete-Linkage-Verfahrens

	$\{2,5\}$	$\{1,4\}$	$\{3,6\}$
$\{2,5\}$			
$\{1,4\}$	10		
$\{3,6\}$	32	41	

Table 13.14. Ergebnis des vierten Schritts des Complete-Linkage-Verfahrens

	$\{1,2,4,5\}$	$\{3,6\}$
$\{1,2,4,5\}$		
$\{3,6\}$	41	

Beim Abstand 41 werden die Klassen $\{1,2,4,5\}$ und $\{3,6\}$ verschmolzen. Wir erhalten das Dendrogramm in Abbildung 13.1 auf Seite 410. Die kophenetische Matrix lautet

$$\mathbf{D}^* = \begin{pmatrix} 0 & 10 & 41 & 4 & 10 & 41 \\ 10 & 0 & 41 & 10 & 1 & 41 \\ 41 & 41 & 0 & 41 & 41 & 3 \\ 4 & 10 & 41 & 0 & 10 & 41 \\ 10 & 1 & 41 & 10 & 0 & 41 \\ 41 & 41 & 3 & 41 & 41 & 0 \end{pmatrix}. \tag{13.3}$$

Wir sehen, dass das Beispiel 57 auf Seite 409 von dieser kophenetischen Matrix ausging. □

Das Average-Linkage-Verfahren Beim Average-Linkage-Verfahren wird als Distanz von zwei zu verschmelzenden Klassen der Mittelwert aller Distanzen zwischen Elementen der einen Klasse und Elementen der anderen Klasse gewählt. hmcounterend. (fortgesetzt)

Example 56. Wir gehen wieder von der Distanzmatrix in Gleichung (13.1) auf Seite 408 aus. Das kleinste Element ist die 1. Im ersten Schritt verschmelzen wir die Klassen $\{2\}$ und $\{5\}$. Es gilt

$$\begin{aligned} D_{\{2,5\}.\{1\}} &= \frac{d_{21} + d_{51}}{2} = \frac{5+6}{2} = 5.5, \\ D_{\{2,5\}.\{3\}} &= \frac{d_{23} + d_{53}}{2} = \frac{32+31}{2} = 31.5, \\ D_{\{2,5\}.\{4\}} &= \frac{d_{24} + d_{54}}{2} = \frac{9+10}{2} = 9.5, \\ D_{\{2,5\}.\{6\}} &= \frac{d_{26} + d_{56}}{2} = \frac{29+28}{2} = 28.5. \end{aligned}$$

Wir erhalten somit Tabelle 13.15.

Table 13.15. Ergebnis des ersten Schritts des Average-Linkage-Verfahrens

	$\{2,5\}$	$\{1\}$	$\{3\}$	$\{4\}$	$\{6\}$
$\{2,5\}$					
$\{1\}$	5.5				
$\{3\}$	31.5	41			
$\{4\}$	9.5	4	41		
$\{6\}$	28.5	41	3	41	

Die kleinste Zahl ist die 3, sodass die Klassen $\{3\}$ und $\{6\}$ verschmolzen werden. Es gilt

$$D_{\{3,6\}.\{2,5\}} = \frac{d_{32}+d_{35}+d_{62}+d_{65}}{4} = \frac{32+31+29+28}{4} = 30,$$

$$D_{\{3,6\}.\{1\}} = \frac{d_{31}+d_{61}}{2} = \frac{37+34}{2} = 35.5,$$

$$D_{\{3,6\}.\{4\}} = \frac{d_{34}+d_{64}}{2} = \frac{41+38}{2} = 39.5.$$

Somit ergibt sich Tabelle 13.16.

Table 13.16. Ergebnis des zweiten Schritts des Average-Linkage-Verfahrens

	$\{2,5\}$	$\{1\}$	$\{3,6\}$	$\{4\}$
$\{2,5\}$				
$\{1\}$	5.5			
$\{3,6\}$	30	35.5		
$\{4\}$	9.5	4	39.5	

Die kleinste Zahl ist die 4, sodass die Klassen $\{1\}$ und $\{4\}$ verschmolzen werden. Es gilt

$$D_{\{1,4\}.\{2,5\}} = \frac{d_{12}+d_{15}+d_{42}+d_{45}}{4} = \frac{5+6+9+10}{4} = 7.5,$$

$$D_{\{1,4\}.\{3,6\}} = \frac{d_{13}+d_{16}+d_{43}+d_{46}}{4} = \frac{37+34+41+38}{4} = 37.5.$$

Somit ergibt sich Tabelle 13.17.

Die kleinste Zahl ist die 7.5, sodass die Klassen $\{1,4\}$ und $\{2,5\}$ verschmolzen werden. Es gilt

$$D_{\{1,2,4,5\}.\{3,6\}} = \frac{d_{13}+d_{16}+d_{43}+d_{46}+d_{23}+d_{26}+d_{53}+d_{56}}{8} =$$

$$= \frac{37+34+41+38+32+29+31+28}{8} = 33.75.$$

Table 13.17. Ergebnis des dritten Schritts des Average-Linkage-Verfahrens

	$\{2,5\}$	$\{1,4\}$	$\{3,6\}$
$\{2,5\}$			
$\{1,4\}$	7.5		
$\{3,6\}$	30	37.5	

Somit ergibt sich Tabelle 13.18.

Table 13.18. Ergebnis des vierten Schritts des Average-Linkage-Verfahrens

	$\{1,2,4,5\}$	$\{3,6\}$
$\{1,2,4,5\}$		
$\{3,6\}$	33.75	

Somit werden beim Abstand 33.75 die Klassen $\{1,2,4,5\}$ und $\{3,6\}$ verschmolzen. Wir erhalten das Dendrogramm in Abbildung 13.3.
Die kophenetische Matrix lautet

$$D^* = \begin{pmatrix} 0 & 7.5 & 33.75 & 4 & 7.5 & 33.75 \\ 7.5 & 0 & 33.75 & 7.5 & 1 & 33.75 \\ 33.75 & 33.75 & 0 & 33.75 & 33.75 & 3 \\ 4 & 7.5 & 33.75 & 0 & 7.5 & 33.75 \\ 7.5 & 1 & 33.75 & 7.5 & 0 & 33.75 \\ 33.75 & 33.75 & 3 & 33.75 & 33.75 & 0 \end{pmatrix}. \tag{13.4}$$

□

13.2.3 Praktische Aspekte

Eigenschaften der Verfahren Wir schauen uns im Folgenden an, welche Dendrogramme die einzelnen Verfahren liefern, wenn sie in speziellen Situationen angewendet werden. Wir haben im ersten Abschnitt dieses Kapitels davon gesprochen, dass die Klassen kohärent und isoliert sein sollen. Wir schauen uns drei Fälle an.

Example 58. In Abbildung 13.4 ist eine Konfiguration zu sehen, in der die Klassen kohärent und isoliert sind. Die Dendrogramme der drei Verfahren zeigen, dass durch jedes der Verfahren die drei Klassen entdeckt werden.

□

Example 59. In Abbildung 13.5 ist eine Konfiguration zu sehen, in der die Klassen kohärent, aber nicht isoliert sind. Hier sind in den Dendrogrammen

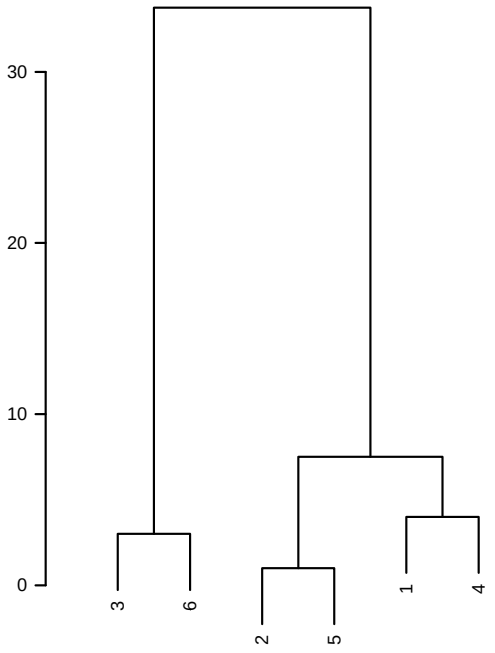


Fig. 13.3. Das Dendrogramm des Average-Linkage-Verfahrens

des Complete-Linkage-Verfahrens und des Average-Linkage-Verfahrens die zwei Klassen sehr gut zu erkennen, während das Single-Linkage-Verfahren keine Klassen erkennen läßt. Bei nicht isolierten Klassen bildet das Single-Linkage-Verfahren eine Kettenstruktur, da der Algorithmus den minimalen Abstand wählt.

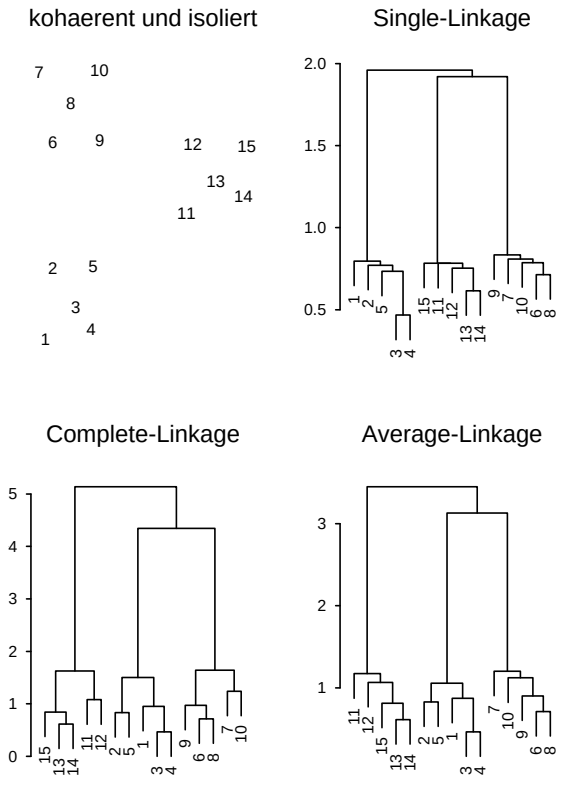


Fig. 13.4. Die drei Verfahren bei kohärenten und isolierten Klassen

□

Example 60. In Abbildung 13.6 ist eine Konfiguration zu sehen, in der die Klassen isoliert, aber nicht kohärent sind. Hier sind im Dendrogramm des Single-Linkage-Verfahrens die beiden Klassen gut zu erkennen, während die beiden anderen Verfahren einen Teil der Punkte der falschen Klasse zuordnen. In einer solchen Situation erweist es sich vorteilhaft, dass das Single-Linkage-Verfahren auf dem kleinsten Abstand beruht.

□

Güte der Lösung Um zu entscheiden, wie gut die Lösung eines Verfahrens ist, vergleicht man die Distanzen der Distanzmatrix $\mathbf{D}$ mit den Distanzen der kophenetischen Matrix $\mathbf{D}^*$. Da die Matrizen symmetrisch sind, benötigen wir nur die Elemente unterhalb der Hauptdiagonalen, also d_{ij} mit $i < j$ und d^*_{ij} mit $i < j$. Gilt $d_{ij} < d_{kl}$, so sollte auch $d^*_{ij} < d^*_{kl}$ gelten. Um dies zu überprüfen, betrachten wir das Streudiagramm der d_{ij} und d^*_{ij} und bestim-

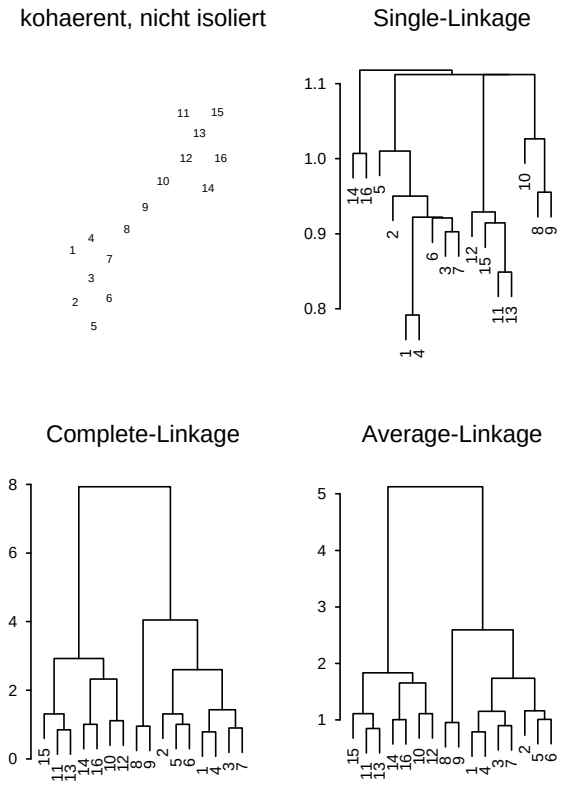


Fig. 13.5. Die drei Verfahren bei kohärenten und nicht isolierten Klassen

men den empirischen Korrelationskoeffizienten zwischen den d_{ij} und d^*_{ij}. Dies ist der *kophenetische Korrelationskoeffizient*. hmcounterend. (fortgesetzt)

Example 56. Wir betrachten die Distanzmatrix $\mathbf{D}$ in Gleichung (13.1) auf Seite 408 und die kophenetische Matrix $\mathbf{D}^*$, die sich aus dem Complete-Linkage-Verfahren ergibt. Diese ist in Gleichung (13.3) zu finden. Abbildung 13.7 zeigt das Streudiagramm von d^*_{ij} gegen d_{ij}.
Wir sehen, dass ein positiver Zusammenhang besteht. Der kophenetische Korrelationskoeffizient nimmt den Wert 0.974 an. □

Man kann mit dem kophenetischen Korrelationskoeffizienten auch unterschiedliche Verfahren der Clusteranalyse vergleichen. Man entscheidet sich für das Verfahren mit dem größten Wert des kophenetischen Korrelationskoeffizienten. hmcounterend. (fortgesetzt)

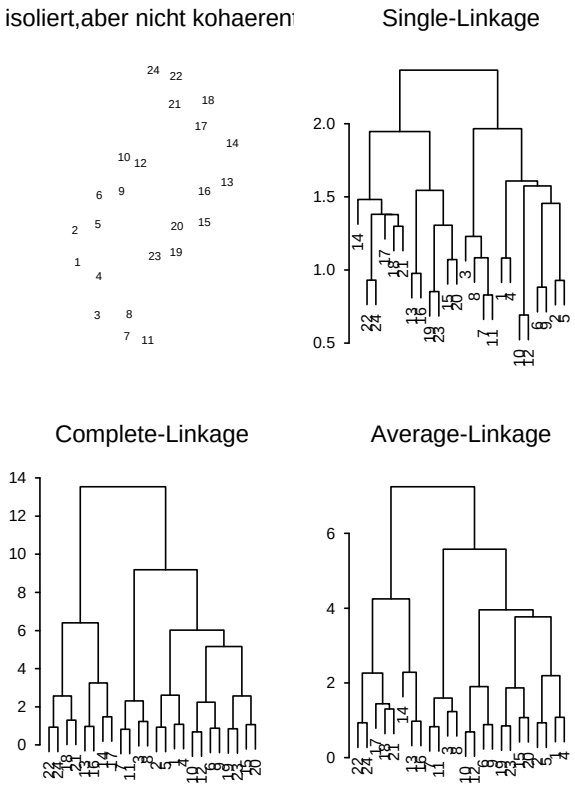


Fig. 13.6. Die drei Verfahren bei isolierten und nicht kohärenten Klassen

Example 56. Beim Single-Linkage-Verfahren beträgt der Wert des kophenetischen Korrelationskoeffizienten 0.973 und beim Average-Linkage-Verfahren 0.974. Wir sehen, dass die drei Verfahren sich nicht stark unterscheiden. □

Hubert (1974) hat vorgeschlagen, den *Gamma-Koeffizienten* zur Beurteilung einer Clusterlösung zu verwenden. Dieser wurde von Goodman & Kruskal (1954) entwickelt. Bei diesem betrachtet man alle Paare der d_{ij} für $i < j$, so zum Beispiel

$$\begin{pmatrix} d_{12} \\ d_{13} \end{pmatrix}.$$

Außerdem betrachtet man noch das entsprechende Paar bei den d^*_{ij}. Man nennt die Paare

$$\begin{pmatrix} d_{ij} \\ d_{kl} \end{pmatrix}$$

und

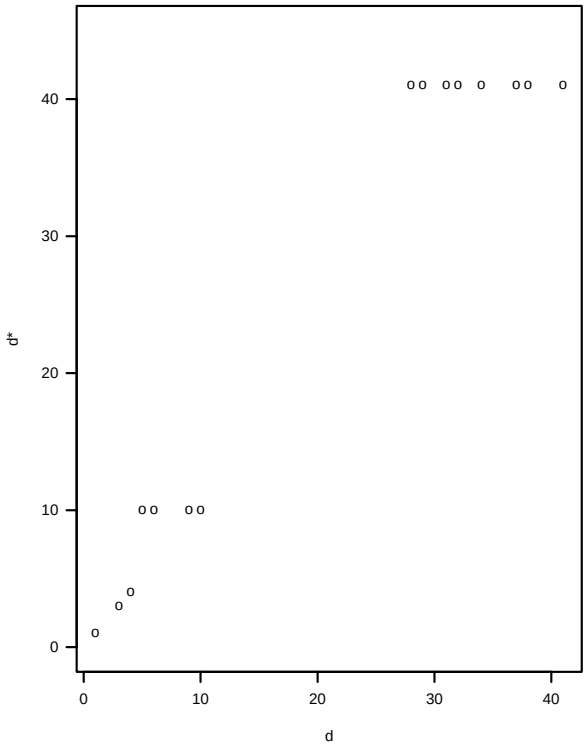


Fig. 13.7. Streudiagramm der d_{ij}^* gegen die d_{ij} für $i < j$

$$\begin{pmatrix} d_{ij}^* \\ d_{kl}^* \end{pmatrix}$$

konkordant, wenn gilt

$$d_{ij} < d_{kl}$$

und

$$d_{ij}^* < d_{kl}^*.$$

Sind alle Paare konkordant, so besteht eine streng monoton wachsende Beziehung zwischen den d_{ij} und den d_{ij}^*. Zwischen zwei Paaren kann aber auch eine gegenläufige Beziehung bestehen. Es gilt also

$$d_{ij} < d_{kl}$$

und

$$d_{ij}^* > d_{kl}^*.$$

Man nennt die Paare in diesem Fall *diskordant*. Sind alle Paare diskordant, so besteht eine streng monoton fallende Beziehung zwischen den d_{ij} und den d_{ij}^*.

Goodman & Kruskal (1954) haben vorgeschlagen, die Anzahl C der konkordanten Paare und die Anzahl D der diskordanten Paare zu bestimmen. Der Gamma-Koeffizient als Maß für den monotonen Zusammenhang zwischen den d_{ij} und den d_{ij}^* ist definiert durch

$$\gamma = \frac{C - D}{C + D}. \tag{13.5}$$

hmcounterend. (fortgesetzt)

Example 56. Beim Complete-Linkage-Verfahren gilt $C = 71$ und $D = 0$. Also ist γ gleich 1. □

Die Tabelle 13.19 aus Bacher (1994), S. 73 gibt Anhaltspunkte zur Beurteilung einer Clusterlösung.

Table 13.19. Beurteilung einer Clusterlösung anhand des Wertes des Gamma-Koeffizienten

Wert von γ	Beurteilung
$0.9 \leq \gamma \leq 1.0$	sehr gut
$0.8 \leq \gamma < 0.9$	gut
$0.7 \leq \gamma < 0.8$	befriedigend
$0.6 \leq \gamma < 0.7$	noch ausreichend
$0.0 \leq \gamma < 0.6$	nicht ausreichend

Anzahl der Klassen Verfahren, mit denen entschieden werden kann, wie viele Klassen vorliegen, beruhen in der Regel auf den Distanzen, bei denen die einzelnen Partitionen gebildet werden, Wir bezeichnen diese Verschmelzungsniveaus im Folgenden mit $\alpha_1, \ldots, \alpha_{n-1}$. So ist α_1 der Abstand, bei dem zum ersten Mal zwei Objekte zu einer Klasse zusammengefasst werden. hmcounterend. (fortgesetzt)

Example 56. Wir betrachten das Dendrogramm des Complete-Linkage-Verfahrens in Abbildung 13.1 auf Seite 410. Es gilt

$$\alpha_1 = 1, \quad \alpha_2 = 3, \quad \alpha_3 = 4, \quad \alpha_4 = 10, \quad \alpha_5 = 41.$$

□

Jedem Verschmelzungsniveau α_i, $i = 1, \ldots, n-1$ ist die Anzahl von $n - i$ Klassen zugeordnet. Zu α_1 zum Beispiel gehören $n-1$ Klassen, da auf diesem

Niveau zum ersten Mal zwei Objekte zu einer Klasse verbunden werden. Jobson (1992) schlägt vor, die $\alpha_1, \ldots, \alpha_{n-1}$ um $\alpha_0 = 0$ zu ergänzen. Ist die Differenz $\alpha_{j+1} - \alpha_j$ groß im Verhältnis zu den Differenzen $\alpha_{i+1} - \alpha_i$ für $i < j$, so spricht dies für $n - j$ Klassen. Eine Graphik erleichtert die Auswahl. Man ordnet α_0 den Wert n zu und visualisiert die Zuwächse der α_i in Form einer Treppenfunktion. hmcounterend. (fortgesetzt)

Example 56. Abbildung 13.8 zeigt die Treppenfunktion. Diese deutet auf das Vorliegen von 2 Klassen hin.

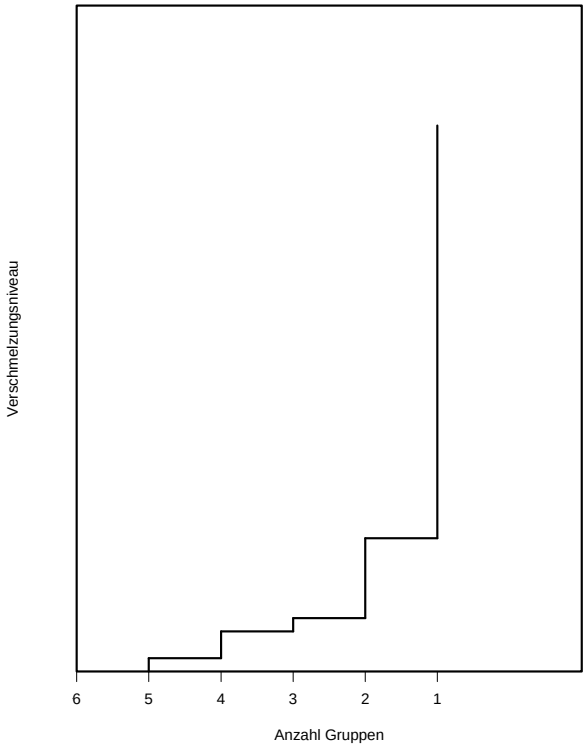


Fig. 13.8.

□

Mojena (1977) hat einen Test vorgeschlagen, der das oben beschriebene Verfahren objektiviert. Zur Durchführung des Tests berechnet man zunächst den Mittelwert und die Stichprobenvarianz der α_i:

$$\bar{\alpha} = \frac{1}{n-1} \sum_{i=1}^{n-1} \alpha_i$$

und

$$s_\alpha = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n-1} (\alpha_i - \bar{\alpha})^2}.$$

hmcounterend. (fortgesetzt)

Example 56. Es gilt $\bar{\alpha} = 11.8$ und $s_\alpha = 277.7$. □

Anschließend bestimmt man die standardisierten α_i:

$$\tilde{\alpha}_i = \frac{\alpha_i - \bar{\alpha}}{s_\alpha}.$$

hmcounterend. (fortgesetzt)

Example 56. Es gilt

$$\begin{aligned}
\tilde{\alpha}_1 &= -0.648, \\
\tilde{\alpha}_2 &= -0.528, \\
\tilde{\alpha}_3 &= -0.468, \\
\tilde{\alpha}_4 &= -0.108, \\
\tilde{\alpha}_5 &= 1.752.
\end{aligned}$$

□

Mojena (1977) hat vorgeschlagen, den Index i zu bestimmen, für den zum ersten Mal gilt

$$\tilde{\alpha}_i > 2.75,$$

und den Wert $n + 1 - i$ für die Anzahl der Klassen zu wählen. Auf Grund von Simulationsstudien empfehlen Milligan & Cooper (1985) die $\tilde{\alpha}_i$ mit 1.25 zu vergleichen. Wir folgen dieser Empfehlung. hmcounterend. (fortgesetzt)

Example 56. Es gilt $i = 5$. Also entscheiden wir uns für $6+1-5 = 2$ Klassen. □

13.2.4 Hierarchische Clusteranalyse in S-PLUS

Wir betrachten die Distanzmatrix **D** in Gleichung (13.1) auf Seite 408. Wir geben zunächst die Daten ein, bestimmen die euklidischen Distanzen und erstellen die Distanzmatrix `dm` mit der Funktion `distfull`, die auf Seite 495 zu finden ist:

```
> alter<-c(43,38,6,47,37,9)
> d<-dist(alter)
> dm<-distfull(d)
> dm
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,]    0    5   37    4    6   34
 [2,]    5    0   32    9    1   29
 [3,]   37   32    0   41   31    3
 [4,]    4    9   41    0   10   38
 [5,]    6    1   31   10    0   28
 [6,]   34   29    3   38   28    0
```

In `S-PLUS` gibt es eine Funktion `hclust`, die für eine Distanzmatrix eine hierarchische Clusteranalyse durchführt. Der Aufruf von `hclust` ist

```
hclust(dist, method = "compact")
```

Dabei ist `dist` die Distanzmatrix. Wir können dem Argument `dist` die Werte unterhalb der Hauptdiagonalen der Distanzmatrix als Vektor `d` zuweisen. In diesem Fall muss aber `attr(d,"Size")` die Anzahl der Objekte enthalten. Mit dem Argument `method` übergibt man das Verfahren, das man benutzen will. Dieses ist standardmäßig auf `"compact"` gesetzt. In diesem Fall wird das Complete-Linkage-Verfahren durchgeführt. Will man das Single-Linkage-Verfahren benutzen, so muss man `method` auf `"connected"` setzen, beim Average-Linkage-Verfahren auf `"average"`.

Wir beginnen mit dem Single-Linkage-Verfahren und interpretieren das Ergebnis der Funktion `hclust` am Beispiel:

```
> e<-hclust(dm,method="compact")
> e
 $merge:
      [,1] [,2]
 [1,]   -2   -5
 [2,]   -3   -6
 [3,]   -1   -4
 [4,]    1    3
 [5,]    2    4

 $height:
 [1]  1  3  4 10 41

 $order:
 [1] 3 6 2 5 1 4
```

Das Ergebnis der Funktion `hclust` ist eine Liste, die drei Komponenten besitzt. Die erste Komponente ist eine (n-1,2)-Matrix. In der i-ten Zeile dieser Matrix steht, welche Objekte bzw. Klassen auf der i-ten Stufe verschmolzen

wurden. Handelt es sich um Objekte, so sind diese Zahlen negativ, bei Klassen sind sie positiv. So stehen in der ersten Zeile die Zahlen -2 und -5. Dies bedeutet, dass auf der ersten Stufe die Objekte 2 und 5 verschmolzen werden. In der vierten Zeile stehen die Zahlen 1 und 3. Dies bedeutet, dass auf der vierten Stufe die auf der ersten Stufe entstandene Klasse mit der Klasse verschmolzen wird, die auf der dritten Stufe gebildet wurde. Die zweite Komponente gibt an, bei welchen Abständen die Klassen gebildet werden. Die dritte Komponente gibt die Objekte in einer Reihenfolge an, in der das Dendrogramm so gezeichnet werden kann, dass sich keine Linien schneiden.

Gezeichnet wird ein Dendrogramm mit der Funktion `plclust`, deren Argument das Ergebnis der Funktion `hclust` ist. Wir geben also ein

```
> plclust(hclust(dm,method="compact"))
```

Dies liefert das Dendrogramm in Abbildung 13.1 auf Seite 410.

Um die Güte der Clusterlösung bestimmen zu können, benötigt man die kophenetische Matrix. Im Anhang B ist auf Seite 498 eine Funktion `cophenetic` angegeben, mit der man die kophenetische Matrix bestimmen kann. Wir wenden die Funktion `cophenetic` auf das Beispiel beim Complete-Linkage-Verfahren an. Wir bestimmen zunächst die einzelnen Stufen des Verschmelzungsprozesses und die Verschmelzungsniveaus:

```
> e<-hclust(dm,method="compact")
```

Dann bestimmen wir mit der Funktion `cophenetic` die kophenetische Matrix:

```
> coph<-cophenetic(e$m,e$h)

> coph
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0   10   41    4   10   41
[2,]   10    0   41   10    1   41
[3,]   41   41    0   41   41    3
[4,]    4   10   41    0   10   41
[5,]   10    1   41   10    0   41
[6,]   41   41    3   41   41    0
```

Den Wert des kophenetischen Korrelationskoeffizienten erhalten wir dann durch

```
> cor(dm[lower.tri(dm)],coph[lower.tri(coph)])
 [1] 0.9735457
```

Im Anhang B ist auf Seite 499 eine Funktion `gammakoeffizient` zu finden, die den Gamma-Koeffizienten bestimmt. Der folgende Aufruf bestimmt den Gamma-Koeffizienten zwischen der Distanzmatrix und der kophenetischen Matrix:

```
> gammakoeffizient(dm[lower.tri(dm)], coph[lower.tri(coph)])
[1] 1
```

Wir können auch den Test von Mojena durchführen. Hierzu bestimmen wir die Verschmelzungsniveaus:

```
> e$h
[1]  1  3  4 10 41
```

Wir standardisieren diese Werte:

```
> (e$h-mean(e$h))/sqrt(d.c$h)
[1] -10.800000  -5.080682  -3.900000  -0.569210   4.560274
```

und sehen, dass zwei Klassen vorliegen.

Wir können dieses Ergebnis auch direkt erhalten:

```
> 1+sum((e$h-mean(e$h))/sqrt(e$h)>1.25)
[1] 2
```

Die von Jobson (1992) vorgeschlagene Abbildung 13.8 auf Seite 432 erhalten wir durch folgende Befehlsfolge:

```
> plot(rep(1,2),c(0,e$h[1]),xaxt="n",yaxt="n",xlim=c(0,7),
    xaxs="i",yaxs="i",ylim=c(0,50),type="l",
    xlab="Anzahl Gruppen",ylab="Verschmelzungsniveau")
> for(i in 2:5) lines(c(i,i),c(e$h[i-1],e$h[i]))
> for (i in 1:4) lines(c(i,i+1),rep(e$h[i],2))
> axis(1,at=0:5,labels=6:1)
```

Mit den Argumenten `xaxt` und `yaxt` kann man festlegen, ob Ticks und Zahlen an die Achsen geschrieben werden sollen. Setzt man diese Argumente auf `"n"`, so werden keine Ticks und Zahlen an die Achsen geschrieben. Mit der Funktion `axis` kann man eine eigene Beschriftung wählen. Mit dem ersten Argument von `axis` legt man fest, welche Achse beschriftet werden soll. Eine `1` steht für die x-Achse, eine 2 für die y-Achse. Mit dem Argument `at` legt man fest, an welchen Stellen die Achse beschriftet werden soll. Das Argument `labels` enthält die Beschriftung. Hat man sich für eine bestimmte Anzahl von Klassen entschieden, so will man natürlich wissen, welche Objekte in den einzelnen Klassen sind, und die Klassen gegebenenfalls beschreiben. Im Anhang B ist auf Seite 499 eine Funktion `welche.cluster` zu finden, die für jedes Objekt die Nummer der Klasse angibt, zu der es gehört. Wir rufen die Funktion `welche.cluster` auf:

```
> welche.cluster(e$m,e$h,2)
[1] 1 1 2 1 1 2
```

13.3 Partitionierende Verfahren

13.3.1 Theorie

Bei den hierarchischen Verfahren bleiben zwei Objekte in einer Klasse, sobald sie verschmolzen sind. Dies ist bei partitionierenden Verfahren nicht der

Fall. Von diesen wollen wir uns im Folgenden mit *K-Means* und *K-Medoids* beschäftigen. K-Means geht davon aus, dass an jedem von n Objekten p quantitative Merkmale erhoben wurden. Es liegen also $\mathbf{y}_1, \ldots, \mathbf{y}_n$ vor.

Example 61. Wir betrachten das Alter der 6 Personen:

```
43 38  6 47 37  9.
```

□

Die n Objekte sollen nun so auf K Klassen aufgeteilt werden, dass die Objekte innerhalb einer Klasse sich sehr ähnlich sind, während die Klassen sich unterscheiden. Bei K-Means muss man die Anzahl K der Klassen vorgeben. Außerdem beginnt man mit einer Startlösung, bei der man jedes Objekt genau einer Klasse zuordnet. hmcounterend. (fortgesetzt)

Example 61. Wir bilden zwei Klassen. Das Alter der Personen beträgt in der ersten Klasse $43, 38, 6$ und in der zweiten Klasse $47, 37, 9$. □

Wir bezeichnen die Anzahl der Elemente in der k-ten Klasse mit n_k, $k = 1, \ldots, K$. hmcounterend. (fortgesetzt)

Example 61. Es gilt $n_1 = 3$ und $n_2 = 3$. □

Wir bezeichnen den Mittelwert $\bar{\mathbf{y}}_k$ der k-ten Klasse auch als Zentrum der k-ten Klasse. hmcounterend. (fortgesetzt)

Example 61. Es gilt $\bar{y}_1 = 29$ und $\bar{y}_2 = 31$. Die folgende Graphik verdeutlicht die Ausgangssituation, wobei die Zentren der beiden Klassen mit **1** und **2** bezeichnet werden:

1 2 1 2 2 1 1 2

1

□

Die Beschreibung der weiteren Vorgehensweise wird vereinfacht, wenn wir die Objekte aus der Sicht ihrer Klassen betrachten. Die Klasse des i-ten Objektes bezeichnen wir mit $C(i)$. hmcounterend. (fortgesetzt)

Example 61. Es gilt

$$C(1) = 1, \quad C(2) = 1, \quad C(3) = 1, \quad C(4) = 2, \quad C(5) = 2, \quad C(6) = 2.$$

□

Das Zentrum der Klasse des i-ten Objekts bezeichnen wir mit $\bar{\mathbf{y}}_{C(i)}$. hmcounterend. (fortgesetzt)

Example 61. Es gilt

$$\bar{y}_{C(1)} = 29, \ \bar{y}_{C(2)} = 29, \ \bar{y}_{C(3)} = 29, \ \bar{y}_{C(4)} = 31, \ \bar{y}_{C(5)} = 31, \ \bar{y}_{C(6)} = 31.$$

□

Für jedes Objekt i bestimmen wir die quadrierte euklidische Distanz $d^2_{i.C(i)}$ zwischen $\mathbf{y}_i$ und $\bar{\mathbf{y}}_{C(i)}$. hmcounterend. (fortgesetzt)

Example 61. Die quadrierten euklidischen Distanzen der Objekte von den Zentren der beiden Klassen sind

$$\begin{aligned} d^2_{1.C(1)} &= (43-29)^2 = 196, \\ d^2_{2.C(2)} &= (38-29)^2 = 81, \\ d^2_{3.C(3)} &= (6-29)^2 = 529, \\ d^2_{4.C(4)} &= (47-31)^2 = 256, \\ d^2_{5.C(5)} &= (37-31)^2 = 36, \\ d^2_{6.C(6)} &= (9-31)^2 = 484\,. \end{aligned}$$

□

Als Güte der Lösung bestimmen wir

$$\sum_{i=1}^{n} d^2_{i.C(i)}. \tag{13.6}$$

hmcounterend. (fortgesetzt)

Example 61. Es gilt

$$\sum_{i=1}^{n} d^2_{i.C(i)} = 196 + 81 + 529 + 256 + 36 + 484 = 1582.$$

□

Ziel von K-Means ist es, die n Beobachtungen so auf die K Klassen zu verteilen, dass (13.6) minimal wird. Um diese Partition zu finden, wird der Reihe nach für jedes der Objekte bestimmt, wie sich (13.6) ändert, wenn das Objekt von seiner Klasse in eine andere Klasse wechselt. Ist die Veränderung negativ, so wird das Objekt verschoben. Lohnt sich keine Verschiebung mehr, so ist der Algorithmus beendet. hmcounterend. (fortgesetzt)

Example 61. Wir verschieben das erste Objekt in die zweite Klasse. Hierdurch ändern sich die Zentren der beiden Klassen zu

$$\bar{y}_1 = 22, \quad \bar{y}_2 = 34.$$

Die quadrierte euklidische Distanz jedes Objekts zum Zentrum seiner Klasse ist

$$\begin{aligned} d^2_{1.C(1)} &= (43-34)^2 = 81, \\ d^2_{2.C(2)} &= (38-22)^2 = 256, \\ d^2_{3.C(3)} &= (6-22)^2 = 256, \\ d^2_{4.C(4)} &= (47-34)^2 = 169, \\ d^2_{5.C(5)} &= (37-34)^2 = 9, \\ d^2_{6.C(6)} &= (9-34)^2 = 625\,. \end{aligned}$$

Der neue Wert von (13.6) ist gegeben durch

$$\sum_{i=1}^{n} d^2_{i.C(i)} = 81 + 256 + 256 + 169 + 9 + 625 = 1396.$$

Da sich (13.6) um 186 vermindert, lohnt es sich, das Objekt 1 in die andere Klasse zu verschieben.

Die folgende Graphik verdeutlicht die neue Situation:

□

Nun wird der Reihe nach für jedes weitere Objekt überprüft, ob es in eine andere Klasse transferiert werden soll. Der Algorithmus endet, wenn durch das Verschieben eines Objekts keine Verbesserung mehr erreicht werden kann. hmcounterend. (fortgesetzt)

Example 61. Der Algorithmus endet, wenn die Objekte 3 und 6 in der ersten Klasse und die restlichen Objekte in der zweiten Klasse sind. Die folgende Graphik veranschaulicht die Lösung:

Wir sehen, dass die gefundenen Klassen kohärent und isoliert sind. □

Ein Nachteil von K-Means ist es, dass die Beobachtungen quantitativ sein müssen, damit man die Mittelwerte in den Klassen bestimmen kann. Dieses Problem kann man dadurch umgehen, dass man Objekte als Zentren der Klassen wählt. Kaufman & Rousseeuw (1990) nennen diese *Medoide* und das Verfahren K-Medoids. Sei $d_{i.C(i)}$ die Distanz des i-ten Objekts zum Medoid seiner Klasse. Die Objekt werden so auf die K Klassen verteilt, dass

$$\sum_{i=1}^{n} d_{i.C(i)}$$

minimal ist. Dieses Verfahren hat auch den Vorteil, dass man nur die Distanzen zwischen den Objekten benötigt. Man kann K-Medoid also auch auf Basis einer Distanzmatrix durchführen. Wir wollen den Algorithmus hier nicht darstellen, sondern weiter hinten zeigen, wie man K-Medoids in `S-PLUS` anwendet.

13.3.2 Praktische Aspekte

Silhouetten Dendrogramme sind eine graphische Darstellung des Ergebnisses einer hierarchischen Clusteranalyse. Wir wollen uns in diesem Abschnitt mit einem Verfahren von Rousseeuw (1987) beschäftigen, das es uns erlaubt, das Ergebnis jeder Clusteranalyse graphisch darzustellen. Hierzu benötigt man die Distanzmatrix $\mathbf{D} = (d_{ij})$ und die Information, zu welcher Klasse das i-te Objekt gehört. Wir nehmen an, dass es K Klassen gibt, die wir mit $C_1, \ldots, C_K$ bezeichnen wollen. Die Anzahl der Objekte in der k-ten Klasse bezeichnen wir mit n_k, $k = 1, \ldots, K$.

Example 62. Wir betrachten wieder das Alter der 6 Personen:

```
43 38  6 47 37  9
```

Mit K-Means wurden die 3. und 6. Person der ersten Klasse und die anderen Personen der zweiten Klasse zugeordnet. Es gilt also

$$C_1 = \{3, 6\}, \quad C_2 = \{1, 2, 4, 5\}$$

und

$$n_1 = 2, \quad n_2 = 4.$$

Die Distanzmatrix zwischen den Objekten ist

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 37 & 4 & 6 & 34 \\ 5 & 0 & 32 & 9 & 1 & 29 \\ 37 & 32 & 0 & 41 & 31 & 3 \\ 4 & 9 & 41 & 0 & 10 & 38 \\ 6 & 1 & 31 & 10 & 0 & 28 \\ 34 & 29 & 3 & 38 & 28 & 0 \end{pmatrix}.$$

□

Jedem Objekt wird nun eine Zahl $s(i)$ zugeordnet, die angibt, wie gut das Objekt klassifiziert wurde. Dabei werden zwei Aspekte betrachtet. Einerseits wird durch eine Maßzahl beschrieben, wie nah ein Objekt an allen anderen Objekten seiner Klasse liegt, andererseits wird eine Maßzahl bestimmt, die die Nähe eines Objekts zu seiner nächsten Klasse beschreibt. Beide Maßzahlen werden zu einer Maßzahl zusammengefaßt.

Beginnen wir mit der Bestimmung der Distanz des i-ten Objekts zu allen anderen Objekten seiner Klasse. Nehmen wir an, das Objekt i gehört zur Klasse C_k. Wir bestimmen den mittleren Abstand $a(i)$ des Objektes i zu allen anderen Objekten, die zur Klasse C_k gehören:

$$a(i) = \frac{1}{n_k - 1} \sum_{j \in C_k, j \neq i} d_{ij}. \tag{13.7}$$

hmcounterend. (fortgesetzt)

Example 62. Sei $i = 1$. Das erste Objekt gehört zur zweiten Klasse. Außerdem sind in dieser Klasse noch die Objekte 2, 4 und 5. Somit gilt

$$a(1) = \frac{d_{12} + d_{14} + d_{15}}{3} = \frac{5 + 4 + 6}{3} = 5.$$

Für die anderen Objekte erhalten wir:

$$a(2) = 5, \quad a(3) = 3, \quad a(4) = 7.67, \quad a(5) = 5.67, \quad a(6) = 3.$$

□

Nun bestimmen wir für alle $j \neq k$ den mittleren Abstand $d(i, C_j)$ des Objektes i zu allen Objekten der Klasse C_j

$$d(i, C_j) = \frac{1}{n_j} \sum_{l \in C_j} d_{il}. \tag{13.8}$$

Sei

$$b(i) = \min_{j \neq i} d(i, C_j). \tag{13.9}$$

Wir merken uns noch diese Klasse. Es ist die Klasse, die am nächsten am Objekt i liegt. hmcounterend. (fortgesetzt)

Example 62. Betrachten wir wieder das erste Objekt. Da nur eine andere Klasse vorliegt, gilt

$$b(1) = \frac{d_{13} + d_{16}}{2} = \frac{37 + 34}{2} = 35.5.$$

Analog erhalten wir

$$b(2) = 30.5, \quad b(3) = 35.25, \quad b(4) = 39.5, \quad b(5) = 29.5, \quad b(6) = 32.25.$$

□

Aus $a(i)$ und $b(i)$ bestimmen wir nun die Zahl $s(i)$, die beschreibt, wie gut ein Objekt klassifiziert wurde:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & \text{falls} \quad a(i) < b(i) \\ 0 & \text{falls} \quad a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & \text{falls} \quad a(i) > b(i) \end{cases} \tag{13.10}$$

Man kann (13.10) auch folgendermaßen kompakt schreiben:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

Wenn eine Klasse nur ein Objekt enthält, so setzen wir $s(i)$ gleich 0. Liegt $s(i)$ in der Nähe von 1, so liegt das i-te Objekt im Mittel in der Nähe der anderen Objekte seiner Klasse, während es im Mittel weit entfernt von den Objekten der Klasse ist, die ihm am nächsten ist. Das i-te Objekt liegt also in der richtigen Klasse. Liegt $s(i)$ in der Nähe von 0, so liegt das i-te Objekt im Mittel genauso nah an den anderen Objekten seiner Klasse wie an den Objekten der Klasse, die ihm am nächsten ist. Das i-te Objekt kann also nicht eindeutig einer Klasse zugeordnet werden. Liegt $s(i)$ in der Nähe von -1, so liegt das i-te Objekt im Mittel näher an den Objekten seiner nächsten Klasse als an den Objekten seiner eigenen Klasse. hmcounterend. (fortgesetzt)

Example 62. Es gilt

$$\begin{aligned} s(1) &= 0.859, \\ s(2) &= 0.836, \\ s(3) &= 0.915, \\ s(4) &= 0.806, \\ s(5) &= 0.808, \\ s(6) &= 0.907. \end{aligned}$$

Wir sehen, dass alle Werte groß sind, was auf ein gutes Ergebnis der Klassenbildung hindeutet. □

Wir bestimmen nun noch für jede Klasse den Mittelwert der $s(i)$. Außerdem bestimmen wir den Mittelwert $\bar{s}(K)$ aller $s(i)$. Dieser dient als Kriterium bei der Entscheidung, wie viele Klassen gebildet werden sollen. hmcounterend. (fortgesetzt)

Example 62. Für die erste Klasse ist der Mittelwert 0.911 und in der zweiten Klasse 0.827. Außerdem gilt $\bar{s}(2) = 0.855$. □

Rousseeuw (1987) nennt eine graphische Darstellung von $s(i)$, $i = 1, \ldots, n$ *Silhouette*. Jedes $s(i)$ wird als Balkendiagramm abgetragen. Dabei werden die $s(i)$ einer Klasse nebeneinander der Größe nach abgetragen, wobei die größte zuerst kommt. hmcounterend. (fortgesetzt)

Example 62. Abbildung 13.9 zeigt die Silhouette des Beispiels. Das Bild zeigt, dass die Daten durch zwei Gruppen sehr gut beschrieben werden können. □

Anzahl der Klassen Kaufman & Rousseeuw (1990) schlagen vor, für $K = 2, \ldots, n-1$ ein partitionierendes Verfahren wie K-Means anzuwenden und $\bar{s}(K)$ zu bestimmen. Es soll dann die Partition gewählt werden, bei der $\bar{s}(K)$ am größten ist. hmcounterend. (fortgesetzt)

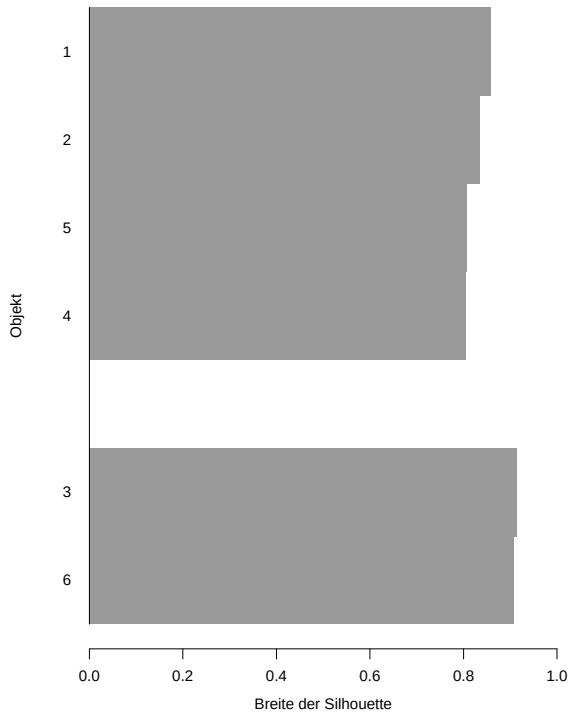


Fig. 13.9. Silhouette

Table 13.20. Werte von $\bar{s}(K)$ in Abhängigkeit von K für die Altersdaten

K	Partition	$\bar{s}(K)$
2	$\{\{1,2,4,5\},\{3,6\}\}$	0.86
3	$\{\{1,4\},\{2,5\},\{3,6\}\}$	0.73
4	$\{\{1\},\{4\},\{2,5\},\{3,6\}\}$	0.57
5	$\{\{1\},\{4\},\{2,5\},\{3\},\{6\}\}$	0.27

Example 62. Tabelle 13.20 zeigt die Partition und den Wert von $\bar{s}(K)$ in Abhängigkeit von K.
Die Tabelle zeigt, dass $K = 2$ der angemessene Wert ist. □

Es gibt noch eine Reihe weiterer Verfahren zur Bestimmung der angemessenen Anzahl der Klassen. Milligan & Cooper (1985) verglichen über 30 Verfahren zur Bestimmung der Anzahl der Klassen. Zu den besten Verfahren gehört das Verfahren von Calinski & Harabasz (1974). Dieses wollen wir im Folgenden näher betrachten.

Example 63. Wir betrachten das Alter der 6 Personen:

```
43 38  6 47 37  9.
```

Bei der Lösung mit zwei Klassen waren die Personen 3 und 6 in einer Klasse, die restlichen Personen in der anderen. □

Calinski & Harabasz (1974) betrachten die Auswahl der Verfahren aus Sicht der multivariaten Varianzanalyse, die wir in Kapitel 11.3 behandelt haben. Für eine Lösung mit K Klassen sei $\mathbf{y}_{kj}$ der Merkmalsvektor des j-ten Objekts in der k-ten Klasse, $k = 1, \ldots, K$, $j = 1, \ldots, n_k$. Calinski & Harabasz (1974) bestimmen die Zwischen-Gruppen-Streumatrix

$$\mathbf{B} = \sum_{k=1}^{K} n_k \, (\bar{\mathbf{y}}_k - \bar{\mathbf{y}})(\bar{\mathbf{y}}_k - \bar{\mathbf{y}})' \tag{13.11}$$

und die Inner-Gruppen-Streumatrix

$$\mathbf{W} = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\mathbf{y}_{kj} - \bar{\mathbf{y}}_k)(\mathbf{y}_{kj} - \bar{\mathbf{y}}_i)'. \tag{13.12}$$

Dabei ist $\bar{\mathbf{y}}_k$ der Mittelwert der Beobachtungen in der k-ten Klasse, $k = 1, \ldots, K$ und $\bar{\mathbf{y}}$ der Mittelwert aller Beobachtungen. Als Kriterium wählen Calinski & Harabasz (1974):

$$G1(K) = \frac{tr(\mathbf{B})}{tr(\mathbf{W})} \, \frac{n - K}{K - 1}.$$

Wurde nur ein Merkmal erhoben, so ist $G1(K)$ identisch mit der Teststatistik des F-Tests in Gleichung (11.9) auf Seite 334. Wächst $G1(K)$ monoton in J, so deutet dies darauf hin, dass keine Klassenstruktur vorliegt. Fällt $G1(K)$ monoton in K, so ist dies ein Indikator für eine hierarchische Struktur. Existiert ein Maximum von $G1(K)$ an der Stelle K_M, die nicht am Rand liegt, so deutet dies auf das Vorliegen von K_M Klassen hin. hmcounterend. (fortgesetzt)

Example 63. Das Zentrum der ersten Klasse ist $\bar{y}_1 = 7.5$, das Zentrum der zweiten Klasse ist $\bar{y}_2 = 41.25$ und das Zentrum aller Beobachtungen ist gleich $\bar{y}_2 = 30$. Somit gilt

$$\mathbf{B} = \sum_{k=1}^{K} n_k \left(\bar{y}_k - \bar{y}\right)^2 = 2 \cdot (7.5 - 30)^2 + 4 \cdot (41.25 - 30)^2 = 1518.75$$

und

$$\begin{aligned}\mathbf{W} = \sum_{k=1}^{K} \sum_{k=1}^{n_k} (y_{kj} - \bar{y}_k)^2 &= (6 - 7.5)^2 + (9 - 7.5)^2 + (43 - 41.25)^2 \\ &+ (38 - 41.25)^2 + (47 - 41.25)^2 + (37 - 41.25)^2 = 69.25\,.\end{aligned}$$

Es gilt

$$G1(2) = \frac{1518.75}{69.25}\,\frac{4}{1} = 87.72563.$$

Tabelle 13.21 gibt $G(K)$ in Abhängigkeit von K an.

Table 13.21. Werte von $G1(K)$ in Abhängigkeit von K für die Altersdaten

K	Partition	$G1(K)$
2	$\{\{1,2,4,5\},\{3,6\}\}$	87.72563
3	$\{\{1,4\},\{2,5\},\{3,6\}\}$	181.73077
4	$\{\{1\},\{4\},\{2,5\},\{3,6\}\}$	211.06667
5	$\{\{1\},\{2\},\{3,6\},\{4\},\{5\}\}$	87.97222

Wir sehen, dass nach dem Kriterium von Calinski und Harabasz eine Lösung mit 4 Klassen gewählt wird. □

Güte der Lösung Kaufman & Rousseeuw (1990), S. 88 schlagen vor, den sogenannten *Silhouettenkoeffizienten SC* zu bestimmen. Dieser ist definiert durch

$$SC = \max_K \bar{s}(K).$$

Anhand des Wertes von SC wird die Lösung mit Hilfe der Tabelle 13.22 begutachtet.
Im Beispiel ist SC gleich 0.86. Somit wurde eine starke Gruppenstruktur gefunden. Die nach dem Kriterium von Calinski und Harabasz gefundene Lösung ist mit einem Wert von 0.57 hingegen nur eine vernünftige Lösung.

Table 13.22. Beurteilung einer Clusterlösung anhand des Silhouettenkoeffizienten

SC	Interpretation
0.71 bis 1.00	starke Struktur
0.51 bis 0.70	vernünftige Struktur
0.26 bis 0.50	schwache Struktur
0.00 bis 0.25	keine substantielle Struktur

13.3.3 Partitionierende Verfahren in S-PLUS

Wir betrachten die Daten im Beispiel 56 auf Seite 407 und geben sie ein:

```
> alter<-c(43,38,6,47,37,9)
```

In S-PLUS gibt es für K-Means eine Funktion `kmeans`, die folgendermaßen aufgerufen wird:

```
kmeans(x, centers, iter.max=10)
```

Dabei ist `x` die Datenmatrix. In den Zeilen der Matrix `centers` stehen Startwerte für die Mittelwerte der Klassen. Die maximale Anzahl der Iterationen steht in `iter.max`. Das Ergebnis der Funktion `kmeans` ist eine Liste, die wir uns am Beispiel ansehen. Wir wählen zwei Klassen. Als Startwerte für die Klassen wählen wir die ersten beiden Beobachtungen.

```
> e<-kmeans(matrix(alter,6,1),matrix(alter[1:2],2,1))
> e
 Centers:
       [,1]
 [1,] 41.25
 [2,]  7.50

 Clustering vector:
 [1] 1 1 2 1 1 2

 Within cluster sum of squares:
 [1] 64.75  4.50

 Cluster sizes:
 [1] 4 2

 Available arguments:
 [1] "cluster"  "centers"  "withinss" "size"
```

Die erste Komponente von `e` gibt die Mittelwerte der Klasse an. Die zweite Komponente der Liste `e` ist ein Vektor, dessen i-te Komponente die Nummer der Klasse der i-ten Beobachtung ist. Wir sehen, dass die Beobachtungen 1,

2, 4 und 5 in der ersten Klasse und die beiden anderen Beobachtungen in der zweiten Klasse sind. Die dritte Komponente ist ein Vektor, dessen i-te Komponente die Spur folgender Matrix ist:

$$\sum_{k=1}^{n_k} (\mathbf{y}_{kj} - \bar{\mathbf{y}}_k)(\mathbf{y}_{kj} - \bar{\mathbf{y}}_i)'.$$

Die Summe der Komponenten dieses Vektors ist gleich der Spur von $\mathbf{W}$ in Gleichung (13.12) auf Seite 447. Die vierte Komponente von `e` gibt die Größen der Klassen an.

Um K-Medoids durchführen zu können, müssen wir zuerst die Bibliothek `cluster` laden. Dies geschieht durch

```
> library(cluster)
```

In dieser Bibliothek gibt es eine Funktion `pam`, mit der man K-Medoids durchführen kann. Dabei steht `pam` für *partitioning around medoids*. Die Funktion `pam` wird folgendermaßen aufgerufen:

```
pam(x, k, diss = F, metric = "euclidean", stand = F,
   save.x = T, save.diss = T)
```

Das Argument `x` enthält die Datenmatrix, wenn `diss` gleich `F` ist. Ist `diss` gleich `T`, so ist `x` eine Distanzmatrix. Die Anzahl der Klassen wählt man durch das Argument `k`. Wurde eine Datenmatrix übergeben, so kann man durch das Argument `metric` festlegen, ob man die euklidische Metrik oder die Manhattan-Metrik berechnen will. Sollen die Beobachtungen skaliert werden, so setzt man das Argument `stand` auf `T`.

Schauen wir uns das Ergebnis von `pam` für das Beispiel an, wobei wir wieder zwei Klassen wählen:

```
> e<-pam(alter,k=2)
> e
 Call:
 pam(x = alter, k = 2)
 Medoids:
 [1] 43  9
 Clustering vector:
 [1] 1 1 2 1 1 2
 Objective function:
    build swap
 3.333333    3
 Available arguments:
 [1] "medoids"    "clustering" "objective" "isolation"
     "clusinfo"   "silinfo"    "diss"       "data" "call"
```

Es werden identische Klassen wie bei K-Means gewählt. Die Medoide der Klassen sind 43 und 9.

Hat man mit der Funktion `pam` die Klassen bestimmt, so kann man problemlos die Silhouette zeichnen. Wir rufen die Funktion `plot` mit dem Ergebnis der Funktion `pam` auf:

```
> plot(e)
```

Dies geht so leicht, da eine Komponente von `e` die wesentlichen Informationen der Silhouette enthält:

```
> e$silinfo
 $widths:
   cluster neighbor sil_width
1        1        2 0.8591549
2        1        2 0.8360656
5        1        2 0.8079096
4        1        2 0.8059072
3        2        1 0.9148936
6        2        1 0.9069767

 $clus.avg.widths:
 [1] 0.8272593 0.9109352

 $avg.width:
 [1] 0.8551513
```

Nach dem Aufruf von `kmeans` ist diese Information nicht vorhanden. Die Funktion `silhouette` im Anhang B auf Seite 500 liefert diese. Wir rufen `silhouette` mit den Daten des Beispiels auf:

```
> e<-kmeans(matrix(alter,6,1),matrix(c(29,31),2,1))
> dm<-distfull(dist(alter))
> es<-silhouette(e$clus,dm,1:6)
> es
 [[1]]:
   wo naechstes        si
3   1         2 0.9148936
6   1         2 0.9069767
1   2         1 0.8591549
2   2         1 0.8360656
5   2         1 0.8079096
4   2         1 0.8059072

 [[2]]:
 [1] 0.9109352 0.8272593

 [[3]]:
 [1] 0.8551513
```

Das Ergebnis der Funktion `silhouette` ist eine Liste `es`. Die erste Komponente ist eine Matrix. In der ersten Spalte steht die Nummer der Klasse des Objekts, in der zweiten Spalte die Nummer der Klasse, die am nächsten liegt, und in der dritten Spalte der Wert von $s(i)$. Die Namen der Objekte sind die Namen der ersten Dimension der Matrix. Die zweite Komponente von `es` ist eine Vektor, dessen j-te Komponente gleich dem Mittelwert der $s(i)$ der j-ten Klasse ist. Die letzte Komponente von `es` ist der Mittelwert aller $s(i)$.

Die Funktion `plotsilhouette` im Anhang B auf Seite 501 zeichnet die Silhouette. Der Aufruf

```
> plotsilhouette(es)
```

liefert die Abbildung 13.9 auf Seite 446.

Um die Werte in Tabelle 13.20 auf Seite 446 zu erhalten, müssen wir für $k = 2, \ldots, n-1$ die Werte von $\bar{s}(k)$ bestimmen. Hierzu verwenden wir eine Iteration. Als Startwert für K-Means bei k Klassen wählen wir die ersten k Beobachtungen:

```
> si<-rep(0,length(alter)-2)
> dm<-distfull(dist(alter))
> for (i in 2:(length(alter)-1))
> {wo<-kmeans(matrix(alter,6,1),
              matrix(alter[1:i],i,1))$cluster
+ si[i-1]<-silhouette(wo,dm,1:6)[[3]]}
> si
[1] 0.8551513 0.7305527 0.5721387 0.2993472
```

Wie man die Maßzahl $G1(k)$ in Gleichung (13.13) auf Seite 447 bestimmen kann, schauen wir uns exemplarisch für $k = 2$ im Beispiel 61 an.

Wir setzen `k` auf 2:

```
> k<-2
```

Wir bestimmen zuerst $tr(\mathbf{B}+\mathbf{W})$, wobei $\mathbf{B}$ in Gleichung (13.11) auf Seite 447 und $\mathbf{W}$ in Gleichung (13.12) auf Seite 447 steht. Dies erhalten wir durch

```
> spT<-sum((length(alter)-1)*var(alter))
> spT
[1] 1588
```

Anschließend rufen wir die Funktion `kmeans` wie oben auf:

```
> e<-kmeans(matrix(alter,ncol=1),matrix(alter[1:k],ncol=1))
```

Die Summe der Komponenten von `e$withinss` ist gleich der Spur von $\mathbf{W}$. Den Wert von $G1(k)$ erhalten wir also durch

```
> spW<-sum(e$withinss)
> spW
[1] 69.25
```

```
> spB<-spT-spW
> spB
[1] 1518.75
> (spB/spW)*(length(alter)-k)/(k-1)
[1] 87.72563
```

Wurde mehr als ein Merkmal erhoben, so muß man die obige Befehlsfolge nur leicht variieren. Wir wollen im Beispiel 1 auf Seite 3 mit Hilfe von K-Means eine Lösung mit drei Klassen bestimmen. Die Daten mögen in der Matrix `PISA` stehen. Die nachstehende Befehlsfolge liefert den Wert von $G1(k)$:

```
> n<-dim(PISA)[1]
> spT<-sum(diag((n-1)*var(PISA)))
> k<-3
> e<-kmeans(PISA,PISA[1:3,])
> spW<-sum(e$withinss)
> spB<-spT-spW
> (spB/spW)*(n-k)/(k-1)
[1] 56.2101
```

13.4 Clusteranalyse der Daten der Regionen

Wir wollen nun noch eine Clusteranalyse der Daten aus Beispiel 12 auf Seite 11 durchführen. Wir beginnen mit der hierarchischen Clusteranalyse. Bevor wir die Distanzen bestimmen, schauen wir uns die Stichprobenvarianzen der Merkmale an. Es gilt

$$s_1^2 = 637438,\ \ s_2^2 = 29988,\quad s_3^2 = 1017,$$
$$s_4^2 = 315,\ \ s_5^2 = 37,\quad s_6^2 = 96473\,.$$

Die Varianzen unterscheiden sich sehr stark. Deshalb führen wir die Analyse auf Basis der skalierten Daten durch. Die Distanzmatrix der skalierten euklidischen Distanz lautet:

$$\mathbf{D} = \begin{pmatrix} 0 & 0.80 & 4.16 & 3.28 & 3.92 & 4.69 \\ 0.80 & 0 & 4.14 & 3.57 & 4.00 & 4.64 \\ 4.16 & 4.14 & 0 & 2.71 & 1.98 & 2.15 \\ 3.28 & 3.57 & 2.71 & 0 & 3.16 & 3.78 \\ 3.92 & 4.00 & 1.98 & 3.16 & 0 & 2.52 \\ 4.69 & 4.64 & 2.15 & 3.78 & 2.52 & 0 \end{pmatrix}. \tag{13.13}$$

Wir führen eine hierarchische Clusteranalyse mit dem Single-Linkage-Verfahren, dem Complete-Linkage-Verfahren und dem Average-Linkage-Verfahren durch. Der Wert des kophenetischen Korrelationskoeffizienten beträgt beim Single-Linkage-Verfahren 0.925, beim Complete-Linkage-Verfahren 0.883 und beim Average-Linkage-Verfahren 0.929. Der Wert des Gamma-Koeffizienten beträgt beim Single-Linkage-Verfahren 0.945, beim Complete-Linkage-Verfahren 0.881 und beim Average-Linkage-Verfahren 0.945. Wir entscheiden uns für das Average-Linkage-Verfahren. Abbildung 13.10 zeigt das Dendrogramm.

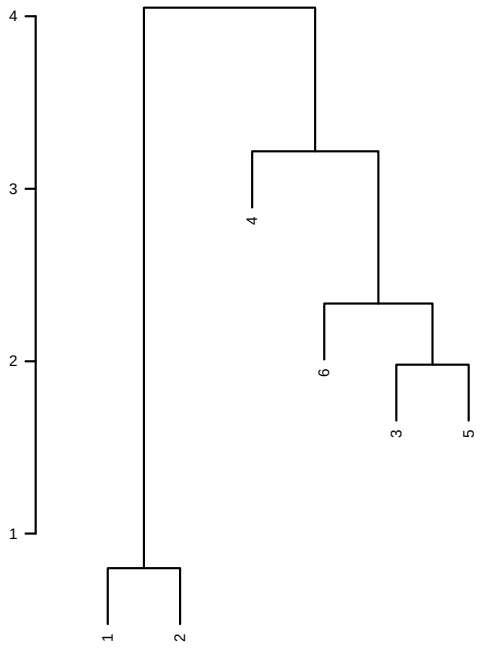


Fig. 13.10. Dendrogramm des Average-Linkage-Verfahrens

Die standardisierten Verschmelzungsniveaus sind:

$$\begin{aligned}\tilde{\alpha}_1 &= -1.357,\\ \tilde{\alpha}_2 &= -0.402,\\ \tilde{\alpha}_3 &= -0.114,\\ \tilde{\alpha}_4 &= 0.599,\\ \tilde{\alpha}_5 &= 1.274\,.\end{aligned}$$

Wir entscheiden uns für zwei Klassen. Die erste Klasse besteht aus Bielefeld und Münster, die zweite Klasse aus den restlichen Regionen.

Wir wollen uns noch anschauen, zu welchem Ergebnis wir mit K-Means gelangen. Wir betrachten auch hier die skalierten Merkmale.

Tabelle 13.23 zeigt die Partition und den Wert von $\bar{s}(k)$ in Abhängigkeit von k.

Table 13.23. Werte von $\bar{s}(k)$ in Abhängigkeit von k für die Daten der Regionen

k	Partition	$\bar{s}(k)$
2	$\{\{1,2\},\{3,4,5,6\}\}$	0.477
3	$\{\{1,2\},\{3,5,6\},\{4\}\}$	0.407
4	$\{\{1\},\{2\},\{3,5,6\},\{4\}\}$	0.151
5	$\{\{1\},\{2\},\{3,5\},\{4\},\{6\}\}$	0.049

Die Werte in Tabelle 13.23 sprechen für die gleiche Lösung wie beim Average-Linkage-Verfahren. Abbildung 13.11 zeigt die Silhouette.

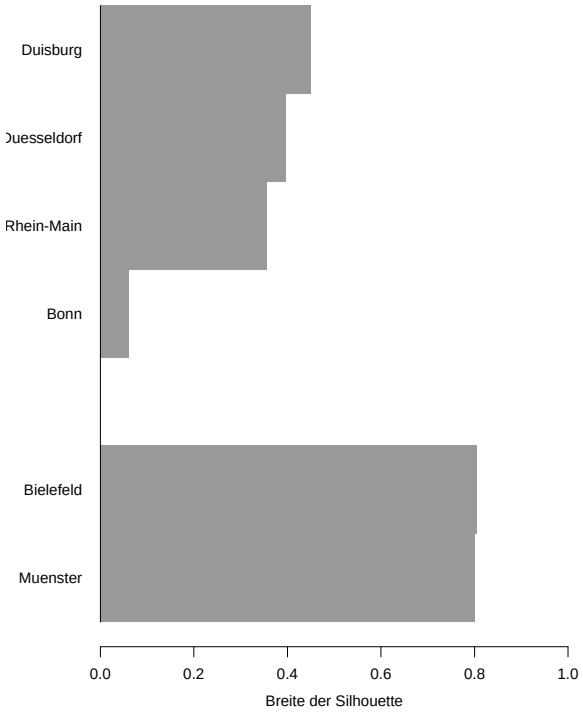


Fig. 13.11. Silhouette der Daten der Regionen

Die Silhouette zeigt, dass die Region Bonn nicht gut angepasst ist. Dies führt auch dazu, dass der Silhouettenkoeffizient mit einem Wert von 0.477 nach Tabelle 13.22 auf Seite 449 auf eine nur schwache Struktur hinweist. Ist man

trotz dieser Bedenken mit der Lösung zufrieden, so bestände der nächste Schritt der Analyse in einer Beschreibung der Klassen. Diesen möge der Leser selber vollziehen.

13.5 Ergänzungen und weiterführende Literatur

Wir haben in diesem Kapitel einige Verfahren der hierarchischen und partitionierenden Clusteranalyse beschrieben. Neben diesen gibt es noch viele andere, die ausführlich bei Everitt (2001), Gordon (1999), Bacher (1994) und im Kapitel 9 in Fahrmeir et al. (1996) dargestellt werden. Diese Bücher enthalten auch viele praktische Aspekte. Das Buch von Kaufman & Rousseeuw (1990) enthält eine Vielzahl von Verfahren, die in `S-PLUS` ab Version 4.0 verfügbar sind. Diese kann man mit dem Befehl `library(cluster)` aktivieren.

13.6 Übungen

Exercise 35. Betrachten Sie die Übung 14 auf Seite 195.

1. Erstellen Sie die Distanzmatrix.
2. Führen Sie für die Distanzmatrix das Complete-Linkage-Verfahren durch.
3. Erstellen Sie ein Dendrogramm mit Hilfe des Complete-Linkage-Verfahrens.
4. Erstellen Sie die kophenetische Matrix für das Ergebnis des Complete-Linkage-Verfahrens.
5. Bestimmen Sie den Wert des Gamma-Koeffizienten.
6. Führen Sie den Test von Mojena durch.

Exercise 36. Im Wintersemester 2000/2001 wurden 299 Studenten in der Veranstaltung Statistik I befragt. Neben dem Merkmal `Geschlecht` mit den Merkmalsausprägungen `1` für weiblich und `0` für männlich wurden noch die Merkmale `Alter` und `Größe`, `Abiturnote in Mathematik` und `Durchschnittsnote im Abitur` erhoben. Für die Merkmale `Abiturnote in Mathematik` und `Durchschnittsnote im Abitur` wählen wir die Abürzungen `MatheNote` beziehungweise `AbiNote`. Außerdem wurden die Studierenden gefragt, ob sie ein eigenes Handy oder einen eigenen PC besitzen, ob sie in Bielefeld studieren wollten, ob sie nach dem Abitur eine Berufsausbildung gemacht haben und ob sie den Leistungskurs Mathematik besucht haben. Wir bezeichnen diese Merkmale mit `Handy`, `PC`, `Biele`, `Ausb` und `MatheLK`. Ihre Ausprägungsmöglichkeiten sind `0` und `1`. Die Ergebnisse der Befragung von 4 Studenten sind in Tabelle 13.24 zu finden.
Die folgende Matrix gibt die Werte der Distanzen zwischen den Studenten an, die mit Hilfe des Gower-Koeffizienten bestimmt wurden. Dabei wurde davon ausgegangen, dass alle binären Merkmale symmetrisch sind. Die Werte sind auf zwei Stellen nach dem Komma gerundet.

Table 13.24. Ergebnis der Befragung von 4 Studenten

Geschlecht	Alter	Größe	Handy	PC	Biele	Ausb	MatheLK	MatheNote	AbiNote
0	20	183	0	0	1	0	1	4	3.1
0	22	185	1	1	0	0	0	5	3.4
1	28	160	1	1	1	1	1	1	1.7
1	23	168	1	1	1	1	0	2	2.5

$$\mathbf{D} = \begin{pmatrix} 0 & 0.48 & 0.75 & 0.68 \\ 0.48 & 0 & 0.78 & 0.51 \\ 0.75 & 0.78 & 0 & 0.27 \\ 0.68 & 0.51 & 0.27 & 0 \end{pmatrix}.$$

1. Verifizieren Sie den Wert des Gower-Koeffizienten zwischen den ersten beiden Studenten.
2. Erstellen Sie das Dendrogramm mit Hilfe des Single-Linkage-Verfahrens.
3. Erstellen Sie die kophenetische Matrix.
4. Bestimmen Sie den Wert des kophenetischen Korrelationskoeffizienten.
5. Bestimmen Sie den Wert des Gamma-Koeffizienten.

Exercise 37. Wir betrachten die Daten im Beispiel 1 auf Seite 3. Verwenden Sie im Folgenden `S-PLUS`.

1. Führen Sie zunächst eine hierarchische Clusteranalyse durch, die auf euklidischen Distanzen beruhen soll.
   a) Welches der drei hierarchischen Verfahren ist auf Grund des Gamma-Koeffizienten am besten geeignet?
   b) Erstellen Sie das Dendrogramm.
   c) Wie viele Klassen sollte man bilden?
   d) Beschreiben Sie die Charakteristika der Klassen.
2. Wenden Sie nun K-Means an.
   a) Wie viele Klassen sollte man bilden?
   b) Erstellen Sie die Silhouette.

Exercise 38. Eine Population besteht aus 6 Studenten. Jeder Student wurde gefragt, wie viel Geld er monatlich zur Verfügung hat. Es ergaben sich folgende Beträge in EUR:

```
334 412 772 374 688 382
```

1. Bilden Sie zwei Klassen mit K-Means.
2. Erstellen Sie die Silhouette.

Exercise 39. Zeigen Sie, dass die kophenetische Matrix folgende Eigenschaft besitzt:

$$d_{ij} \leq max\{d_{ik}, d_{jk}\} \tag{13.14}$$

für alle Tripel (i, j, k) von Objekten.

Skizzieren Sie die Lage der Objekte i, j und k im $\mathbb{R}^2$, wenn sie die Bedingung (13.14) erfüllen.

Exercise 40. Vollziehen Sie die Analyse der Daten der Regionen in Kapitel 13.4 auf Seite 454 in `S-PLUS` nach.

Part V

Anhänge

A Mathematische Grundlagen

A.1 Matrizenrechnung

Das Erlernen und die Anwendung multivariater Verfahren setzt insbesondere Grundkenntnisse der Matrizenrechnung voraus. So sind zum Beispiel auf Seite 223 folgende Umformungen zu finden:

$$
\begin{aligned}
S(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) && \text{(A.1)}\\
&= (\mathbf{y}' - (\mathbf{X}\boldsymbol{\beta})') (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) && \text{(A.2)}\\
&= (\mathbf{y}' - \boldsymbol{\beta}'\mathbf{X}') (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) && \text{(A.3)}\\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} && \text{(A.4)}\\
&= \mathbf{y}'\mathbf{y} - (\mathbf{y}'\mathbf{X}\boldsymbol{\beta})' - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} && \text{(A.5)}\\
&= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} && \text{(A.6)}\\
&= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,. && \text{(A.7)}
\end{aligned}
$$

Dabei ist $\mathbf{y}$ ein n-dimensionaler Vektor, $\mathbf{X}$ eine $(n, k+1)$-Matrix und $\boldsymbol{\beta}$ ein $k+1$-dimensionaler Vektor. Der Übergang von einer zur nächsten Zeile erfolgt jeweils nach einer bestimmten Regel. Diese Regeln werden im weiteren Verlauf dieses Kapitels nach und nach dargestellt. Vorausgeschickt sei dabei, dass folgende Gleichungen benutzt werden, um von einer Zeile zur nächsten zu gelangen:

von (A.1) zu (A.2) : (A.14)
von (A.2) zu (A.3) : (A.24)
von (A.3) zu (A.4) : (A.23)
von (A.4) zu (A.5) : (A.10)
von (A.5) zu (A.6) : (A.9) und (A.24)

Die Regel beim Übergang von (A.6) zu (A.7) kennt man aus der elementaren Mathematik.

A.1.1 Definitionen und spezielle Matrizen

Definition 24. *Eine (n,p)-Matrix $\mathbf{A}$ ist ein rechteckiges Schema reeller Zahlen, das aus n Zeilen und p Spalten besteht:*

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix}. \tag{A.8}$$

Dabei heißt (n,p) die Ordnung der Matrix, i Zeilenindex, j Spaltenindex und a_{ij} Element der Matrix, das in der i-ten Zeile und j-ten Spalte steht. Wir schreiben kurz $\mathbf{A} = (a_{ij})$. Eine $(1,1)$-Matrix ist ein Skalar a.

Example 64. Wir betrachten im Folgenden die Matrizen

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}.$$

□

Definition 25. *Vertauscht man bei einer (n,p)-Matrix $\mathbf{A}$ Zeilen und Spalten, so erhält man die transponierte Matrix $\mathbf{A}'$ von $\mathbf{A}$.*

hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$\mathbf{C}' = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 2 \end{pmatrix}.$$

□

Es gilt

$$(\mathbf{A}')' = \mathbf{A}. \tag{A.9}$$

Für einen Skalar a gilt

$$a' = a\,. \tag{A.10}$$

Definition 26. *Eine Matrix $\mathbf{A}$, für die $\mathbf{A}' = \mathbf{A}$ gilt, heißt symmetrisch.*

hmcounterend. (fortgesetzt)

Example 64. Die Matrizen $\mathbf{A}$ und $\mathbf{D}$ sind symmetrisch. □

Definition 27. *Eine $(n,1)$-Matrix heißt n-dimensionaler Spaltenvektor $\mathbf{a}$ und eine $(1,n)$-Matrix n-dimensionaler Zeilenvektor $\mathbf{b}'$.*

hmcounterend. (fortgesetzt)

Example 64. Im Folgenden betrachten wir die beiden Spaltenvektoren

$$\mathbf{a}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{a}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

□

Die beiden Vektoren des Beispiels sind die Spalten der Matrix $\mathbf{A}$. Eine (n, p)-Matrix $\mathbf{A}$ besteht aus p n-dimensionalen Spaltenvektoren $\mathbf{a}_1, \ldots, \mathbf{a}_p$. Wir schreiben hierfür auch

$$\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p).$$

Entsprechend können wir die Matrix aus Zeilenvektoren aufbauen.

Eine Matrix, die aus lauter Nullen besteht, heißt *Nullmatrix* $\mathbf{0}$. Wir bezeichnen den n-dimensionalen Spaltenvektor, der aus lauter Nullen besteht, ebenfalls mit $\mathbf{0}$ und nennen ihn den *Nullvektor*. Eine Matrix, die aus lauter Einsen besteht, bezeichnen wir mit $\mathbf{E}$. Der n-dimensionale Spaltenvektor $\mathbf{1}$, der aus lauter Einsen besteht, heißt *Einservektor* oder *summierender Vektor*. Wir werden später eine Begründung für die letzte Bezeichnungsweise liefern. Ein Vektor, bei dem die i-te Komponente gleich 1 und alle anderen Komponenten gleich 0 sind, heißt i-ter *Einheitsvektor* $\mathbf{e}_i$.

Definition 28. *Eine (n, p)-Matrix heißt quadratisch, wenn gilt $n = p$.*

hmcounterend. (fortgesetzt)

Example 64. Die Matrizen $\mathbf{A}$, $\mathbf{B}$ und $\mathbf{D}$ sind quadratisch. □

Definition 29. *Eine quadratische Matrix, bei der alle Elemente außerhalb der Hauptdiagonalen gleich Null sind, heißt Diagonalmatrix.*

hmcounterend. (fortgesetzt)

Example 64. Die Matrix $\mathbf{D}$ ist eine Diagonalmatrix. □

Definition 30. *Sind bei einer (n, n)-Diagonalmatrix alle Hauptdiagonalelemente gleich 1, so spricht man von der Einheitsmatrix $\mathbf{I}_n$.*

A.1.2 Matrixverknüpfungen

Definition 31. *Sind $\mathbf{A}$ und $\mathbf{B}$ (n, p)-Matrizen, dann ist die Summe $\mathbf{A} + \mathbf{B}$ definiert durch*

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & \ldots & a_{1p} + b_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & \ldots & a_{np} + b_{np} \end{pmatrix}.$$

Die Differenz $\mathbf{A} - \mathbf{B}$ *ist definiert durch*

$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} a_{11} - b_{11} & \dots & a_{1p} - b_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} - b_{n1} & \dots & a_{np} - b_{np} \end{pmatrix}.$$

hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 2 & 0 \\ 2 & 2 \end{pmatrix}$$

und

$$\mathbf{A} - \mathbf{B} = \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}.$$

□

Schauen wir uns einige Rechenregeln für die Summe und Differenz von Matrizen an. Dabei sind $\mathbf{A}$, $\mathbf{B}$ und $\mathbf{C}$ (n, p)-Matrizen. Es gilt

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}, \tag{A.11}$$

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}), \tag{A.12}$$

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}' \tag{A.13}$$

und

$$(\mathbf{A} - \mathbf{B})' = \mathbf{A}' - \mathbf{B}'. \tag{A.14}$$

Eine Matrix kann mit einem Skalar multipliziert werden.

Definition 32. *Ist* $\mathbf{A}$ *eine* (n, p)*-Matrix und* $k \in \mathbb{R}$ *ein Skalar, dann ist das Produkt* $k\mathbf{A}$ *definiert durch*

$$k\mathbf{A} = \begin{pmatrix} k\,a_{11} & \dots & k\,a_{1p} \\ \vdots & \ddots & \vdots \\ k\,a_{n1} & \dots & k\,a_{np} \end{pmatrix}.$$

hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$2\mathbf{A} = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}.$$

□

Sind $\mathbf{A}$ und $\mathbf{B}$ (n,p)-Matrizen und k und l Skalare, dann gilt

$$k(\mathbf{A}+\mathbf{B}) = k\mathbf{A} + k\mathbf{B} \tag{A.15}$$

und

$$(k+l)\mathbf{A} = k\mathbf{A} + l\mathbf{A}\,. \tag{A.16}$$

Geeignet gewählte Matrizen kann man miteinander multiplizieren. Das Produkt von Matrizen beruht auf dem inneren Produkt von Vektoren.

Definition 33. *Seien*

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

und

$$\mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

n-dimensionale Spaltenvektoren. Das innere Produkt von $\mathbf{a}$ *und* $\mathbf{b}$ *ist definiert durch*

$$\mathbf{a}'\mathbf{b} = \sum_{i=1}^{n} a_i\, b_i.$$

hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$\mathbf{a}_1'\mathbf{a}_2 = 2 \cdot 1 + 1 \cdot 2 = 4.$$

□

Offensichtlich gilt

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}. \tag{A.17}$$

Bildet man $\mathbf{a}'\mathbf{a}$, so erhält man gerade die Summe der quadrierten Komponenten von $\mathbf{a}$:

$$\mathbf{a}'\mathbf{a} = \sum_{i=1}^{n} a_i\, a_i = \sum_{i=1}^{n} a_i^2\,. \tag{A.18}$$

hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$\mathbf{a}_1'\mathbf{a}_1 = 2^2 + 1^2 = 5\,.$$

□

Die Länge $||\mathbf{a}||$ eines n-dimensionalen Vektors $\mathbf{a}$ ist definiert durch

$$||\mathbf{a}|| = \sqrt{\mathbf{a}'\mathbf{a}}\,. \tag{A.19}$$

Dividiert man einen Vektor durch seine Länge, so spricht man von einem normierten Vektor. Die Länge eines normierten Vektors ist gleich 1. hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$\frac{1}{\sqrt{\mathbf{a}_1'\,\mathbf{a}_1}}\,\mathbf{a}_1 = \frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix}.$$

□

Das innere Produkt des n-dimensionalen Einservektors $\mathbf{1}$ mit einem n-dimensionalen Vektor $\mathbf{a}$ liefert die Summe der Komponenten von $\mathbf{a}$. Deshalb heißt $\mathbf{1}$ auch summierender Vektor.

Nun können wir uns dem Produkt zweier Matrizen zuwenden.

Definition 34. *Das Produkt* $\mathbf{AB}$ *einer* (n,p)*-Matrix* $\mathbf{A}$ *und einer* (p,q)*-Matrix* $\mathbf{B}$ *ist definiert durch*

$$\mathbf{AB} = \begin{pmatrix} \sum_{k=1}^{p} a_{1k}b_{k1} & \dots & \sum_{k=1}^{p} a_{1k}b_{kq} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{p} a_{nk}b_{k1} & \dots & \sum_{k=1}^{p} a_{nk}b_{kq} \end{pmatrix}.$$

Das Element in der i-ten Zeile und j-ten Spalte von $\mathbf{AB}$ erhält man, indem man das innere Produkt aus dem i-ten Zeilenvektor von $\mathbf{A}$ und dem j-ten Spaltenvktor von $\mathbf{B}$ bildet. Das Produkt $\mathbf{AB}$ einer (n,p)-Matrix $\mathbf{A}$ und einer (p,q)-Matrix $\mathbf{B}$ ist eine (n,q)-Matrix. hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$\mathbf{AB} = \begin{pmatrix}1 & -2\\ 2 & -1\end{pmatrix}$$

und

$$\mathbf{BA} = \begin{pmatrix}-1 & -2\\ 2 & 1\end{pmatrix}.$$

□

Das Beispiel zeigt, dass $\mathbf{AB}$ nicht notwendigerweise gleich $\mathbf{BA}$ ist.

Ist $\mathbf{A}$ eine (n,p)-Matrix, $\mathbf{B}$ eine (p,q)-Matrix und $\mathbf{C}$ eine (q,r)-Matrix, dann gilt

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})\,. \tag{A.20}$$

Ist $\mathbf{A}$ eine (n,p)-Matrix, $\mathbf{B}$ eine (p,q)-Matrix und k ein Skalar, dann gilt

$$k\mathbf{AB} = \mathbf{A}k\mathbf{B} = \mathbf{AB}k. \tag{A.21}$$

Sind $\mathbf{A}$ und $\mathbf{B}$ (n,p)-Matrizen und $\mathbf{C}$ und $\mathbf{D}$ (p,q)-Matrizen, dann gilt

$$(\mathbf{A}+\mathbf{B})(\mathbf{C}+\mathbf{D}) = \mathbf{AC}+\mathbf{AD}+\mathbf{BC}+\mathbf{BD} \tag{A.22}$$

und

$$(\mathbf{A}-\mathbf{B})(\mathbf{C}-\mathbf{D}) = \mathbf{AC}-\mathbf{AD}-\mathbf{BC}+\mathbf{BD}. \tag{A.23}$$

Ist $\mathbf{A}$ eine (n,p)-Matrix und $\mathbf{B}$ eine (p,q)-Matrix, dann gilt

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'. \tag{A.24}$$

Der Beweis ist bei Zurmühl & Falk (1997), S.21-22 zu finden.

Definition 35. *Das äußere Produkt des n-dimensionalen Spaltenvektors* $\mathbf{a}$ *mit dem p-dimensionalen Spaltenvektor* $\mathbf{b}$ *ist definiert durch*

$$\mathbf{ab}' = \begin{pmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_p \\ a_2b_1 & a_2b_2 & \dots & a_2b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \dots & a_nb_p \end{pmatrix}. \tag{A.25}$$

Man nennt das äußere Produkt auch das *dyadische Produkt*. Ist $\mathbf{a}$ ein n-dimensionaler Spaltenvektor und $\mathbf{b}$ ein p-dimensionaler Spaltenvektor, so ist $\mathbf{ab}'$ eine (n,p)-Matrix und $\mathbf{ba}'$ eine (p,n)-Matrix. In der Regel ist $\mathbf{ab}'$ ungleich $\mathbf{ba}'$. hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$\mathbf{a}_1\mathbf{a}_2' = \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix}$$

und

$$\mathbf{a}_2\mathbf{a}_1' = \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}.$$

□

A.1.3 Die inverse Matrix

Definition 36. *Die (n,n)-Matrix* $\mathbf{A}$ *heißt invertierbar, wenn eine (n,n)-Matrix* $\mathbf{A}^{-1}$ *existiert, sodass gilt*

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n. \tag{A.26}$$

Man nennt $\mathbf{A}^{-1}$ *auch die inverse Matrix von* $\mathbf{A}$.

hmcounterend. (fortgesetzt)

Example 64. Die inverse Matrix von

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

ist

$$\mathbf{A}^{-1} = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Es gilt nämlich

$$\mathbf{A}^{-1}\mathbf{A} = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} = \mathbf{I}_2.$$

□

Für uns sind folgende Eigenschaften wichtig:

1. Die inverse Matrix $\mathbf{A}^{-1}$ der Matrix $\mathbf{A}$ ist eindeutig. Der Beweis ist bei Strang (1988), S.42 zu finden.
2. Sei $\mathbf{A}$ eine invertierbare (n, n)-Matrix. Dann gilt:

$$\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}. \tag{A.27}$$

3. Sei $\mathbf{A}$ eine invertierbare (n, n)-Matrix. Dann gilt:

$$\left(\mathbf{A}^{-1}\right)' = (\mathbf{A}')^{-1}. \tag{A.28}$$

   Der Beweis ist bei Zurmühl & Falk (1997), S.38 zu finden.
4. Sind die (n, n)-Matrizen $\mathbf{A}$ und $\mathbf{B}$ invertierbar, so gilt

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \tag{A.29}$$

   Der Beweis ist bei Zurmühl & Falk (1997), S. 38 zu finden.

A.1.4 Orthogonale Matrizen

Definition 37. *Die n-dimensionalen Spaltenvektoren* $\mathbf{a}$ *und* $\mathbf{b}$ *heißen orthogonal, wenn gilt* $\mathbf{a}'\mathbf{b} = 0$.

Definition 38. *Eine (n, n)-Matrix* $\mathbf{A}$ *heißt orthogonal, wenn gilt*

$$\mathbf{A}\mathbf{A}' = \mathbf{A}'\mathbf{A} = \mathbf{I}_n. \tag{A.30}$$

hmcounterend. (fortgesetzt)

Example 64. Die Matrix $\mathbf{B}$ ist orthogonal. □

In einer orthogonalen Matrix haben alle Spaltenvektoren die Länge 1, und die Spaltenvektoren sind paarweise orthogonal.

Man kann einen n-dimensionalen Spaltenvektor als Punkt in einem kartesischen Koordinatensystem einzeichnen. Multipliziert man eine orthogonale (n, n)-Matrix $\mathbf{T}$ mit einem Spaltenvektor $\mathbf{x}$, so wird der Vektor bezüglich des Nullpunkts gedreht. Schauen wir uns dies in einem zweidimensionalen kartesischen Koordinatensystem an. Multipliziert man die orthogonale Matrix

$$\mathbf{T} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

mit dem Vektor

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

so wird der Vektor $\mathbf{x}$ um α Grad im Gegenzeigersinn gedreht. Eine Begründung hierfür ist bei Zurmühl & Falk (1997), S.6 zu finden. hmcounterend. (fortgesetzt)

Example 64. Es gilt $cos(0.5\pi) = 0$ und $sin(0.5\pi) = 1$. Somit erhalten wir für $\alpha = 0.5\pi$ die Matrix $\mathbf{B}$. Wir bilden

$$\mathbf{B}\mathbf{a}_1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

Abbildung A.1 verdeutlicht den Zusammenhang.

□

A.1.5 Spur einer Matrix

Definition 39. *Sei* $\mathbf{A}$ *eine* (n, n)*-Matrix. Die Spur* $tr(\mathbf{A})$ *ist gleich der Summe der Hauptdiagonalelemente:*

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}. \tag{A.31}$$

Dabei steht tr für trace.

hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$tr(\mathbf{A}) = 4.$$

□

Sind $\mathbf{A}$ und $\mathbf{B}$ (n, n)-Matrizen, so gilt

$$tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B}). \tag{A.32}$$

Ist $\mathbf{A}$ eine (n, p)-Matrix und $\mathbf{B}$ eine (p, n)-Matrix, so gilt

$$tr(\mathbf{AB}) = tr(\mathbf{BA}). \tag{A.33}$$

Der Beweis ist bei Zurmühl & Falk (1997), S.22 zu finden.

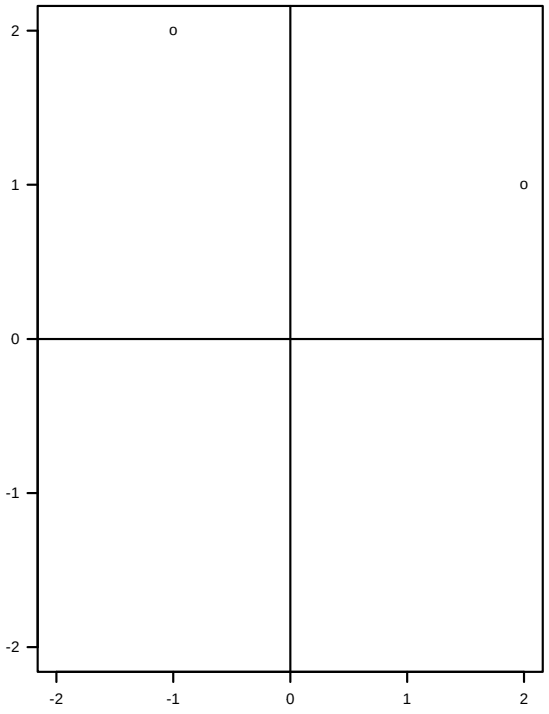


Fig. A.1. Drehung eines Punktes um 90 Grad im Gegenzeigersinn

A.1.6 Determinante einer Matrix

Man kann einer (n,n)-Matrix $\mathbf{A}$ eine reelle Zahl zuordnen, die $\mathbf{A}$ charakterisiert. Dies ist die Determinante $|\mathbf{A}|$.

Definition 40. *Seien* $\mathbf{A}$ *eine* (n,n)*-Matrix und* $\mathbf{A}_{ij}$ *die* $(n-1,n-1)$*-Matrix, die man dadurch erhält, dass man die* i*-te Zeile und* j*-te Spalte von* $\mathbf{A}$ *streicht. Die Determinante* $|\mathbf{A}|$ *von* $\mathbf{A}$ *ist definiert durch*

$$|\mathbf{A}| = \begin{cases} a_{11} & \text{für } n = 1 \\ \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |\mathbf{A}_{ij}| & \text{für } n \geq 2, \text{ } i \text{ fest, aber beliebig, } 1 \leq i \leq n\,. \end{cases}$$

Sei

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

eine $(2,2)$-Matrix. Wir bestimmen $|\mathbf{A}|$ von $\mathbf{A}$ für $i = 1$:

$$|\mathbf{A}| = (-1)^{1+1} a_{11} |\mathbf{A}_{11}| + (-1)^{1+2} a_{12} |\mathbf{A}_{12}| = a_{11} a_{22} - a_{12} a_{21} \,.$$

hmcounterend. (fortgesetzt)

Example 64. Es gilt

$$|\mathbf{A}| = 2 \cdot 2 - 1 \cdot 1 = 3.$$

□

Für uns sind drei Eigenschaften der Determinante wichtig:

1. Für eine (n,n)-Diagonalmatrix

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{pmatrix}$$

   gilt

$$|\mathbf{D}| = d_1 \cdot d_2 \cdot \ldots \cdot d_n. \tag{A.34}$$

   Der Beweis ist bei Wetzel et al. (1981), S. 112 zu finden.
2. Sind $\mathbf{A}$ und $\mathbf{B}$ (n,n)-Matrizen, so gilt

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|. \tag{A.35}$$

   Der Beweis ist bei Jänich (2000), S. 148 zu finden.
3. Die Invertierbarkeit einer (n,n)-Matrix $\mathbf{A}$ kann über die Determinante von $\mathbf{A}$ charakterisiert werden. Die (n,n)-Matrix $\mathbf{A}$ ist genau dann invertierbar, wenn die Determinante von $\mathbf{A}$ ungleich Null ist. Der Beweis ist bei Jänich (2000), S.147-148 zu finden.

A.1.7 Lineare Gleichungssysteme

Wir betrachten ein in den p Unbekannten $x_1, \ldots, x_p$ lineares Gleichungssystem :

$$\begin{aligned} a_{11}\, x_1 + a_{12}\, x_2 + \ldots + a_{1p}\, x_p &= b_1 \\ a_{21}\, x_1 + a_{22}\, x_2 + \ldots + a_{2p}\, x_p &= b_2 \\ \vdots \quad & \quad \vdots \\ a_{n1}\, x_1 + a_{n2}\, x_2 + \ldots + a_{np}\, x_p &= b_n \end{aligned} \tag{A.36}$$

Gilt $b_i = 0$ für $i = 1, \ldots, n$, so spricht man von einem *linear homogenen Gleichungssystem*, ansonsten von einem *linear inhomogenen Gleichungssystem*. Gesucht sind Werte von $x_1, \ldots, x_p$, die das Gleichungssystem erfüllen.

Mit

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{np} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

können wir (A.36) auch folgendermaßen schreiben:

$$\mathbf{Ax} = \mathbf{b}. \tag{A.37}$$

Um die Lösbarkeit von (A.37) diskutieren zu können, benötigen wir den Begriff der linearen Unabhängigkeit.

Definition 41. *Die n-dimensionalen Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_p$ heißen linear unabhängig, wenn aus*

$$x_1\mathbf{a}_1 + \ldots + x_p\mathbf{a}_p = \mathbf{0}$$

folgt

$$x_1 = x_2 = \ldots = 0.$$

Matrizen sind aus Spaltenvektoren beziehungsweise Zeilenvektoren aufgebaut. Die Maximalzahl linear unabhängiger Spaltenvektoren einer Matrix $\mathbf{A}$ nennt man den *Spaltenrang* von $\mathbf{A}$. Die Maximalzahl linear unabhängiger Zeilenvektoren einer Matrix $\mathbf{A}$ nennt man den *Zeilenrang* von $\mathbf{A}$. Die Maximalzahl linear unabhängiger Spaltenvektoren ist gleich der Maximalzahl linear unabhängiger Zeilenvektoren. Der Beweis ist bei Jänich (2000), S. 116-117 zu finden. Diese Zahl bezeichnet man als *Rang* $rg(\mathbf{A})$ von $\mathbf{A}$.

Mit Hilfe des Rangs kann man die Lösbarkeit von (A.37) diskutieren. Wir beginnen mit dem linear homogenen Gleichungssystem.

Das linear homogene Gleichungssystem

$$\mathbf{Ax} = \mathbf{0} \tag{A.38}$$

besitzt für $n \geq p$ genau eine Lösung, wenn der Rang von $\mathbf{A}$ gleich p ist. Dies folgt aus der Definition der linearen Unabhängigkeit. Ist der Rang von $\mathbf{A}$ kleiner als p, so besitzt das linear inhomogene Gleichungssystem mehr als eine Lösung. Die Struktur des Lösungsraums ist bei Wetzel et al. (1981), S.72-74 beschrieben. Gilt $p = n$, so kann man die Lösbarkeit von (A.38) auch über die Determinante der Matrix $\mathbf{A}$ charakterisieren. Die Determinante $|\mathbf{A}|$ von $\mathbf{A}$ ist genau dann ungleich 0, wenn der Rang von $\mathbf{A}$ gleich n ist (siehe dazu Wetzel et al. (1981), S. 114). Ist die Determinante der (n, n)-Matrix $\mathbf{A}$

also gleich 0, so ist der Rang von $\mathbf{A}$ kleiner als n und das linear homogene Gleichungssystem (A.38) hat mehr als eine Lösung.

Wir betrachten nun das linear inhomogene Gleichungssystem

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{A.39}$$

wobei $\mathbf{A}$ eine (n, n)-Matrix ist. Ist der Rang von $\mathbf{A}$ gleich n, so existiert die inverse Matrix $\mathbf{A}^{-1}$ von $\mathbf{A}$ (siehe dazu Wetzel et al. (1981), S. 97). Wir multiplizieren (A.39) von links mit $\mathbf{A}^{-1}$ und erhalten die Lösung

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{A.40}$$

Da die inverse Matrix $\mathbf{A}^{-1}$ eindeutig ist, ist diese Lösung eindeutig.

Wir betrachten an einigen Stellen in diesem Buch eine (n, p)-Matrix $\mathbf{X}$, wobei gilt $rg(\mathbf{X}) = p$, und bilden die Matrix $\mathbf{X}'\mathbf{X}$. Schauen wir uns diese (p, p)-Matrix genauer an. Sie ist symmetrisch. Dies sieht man folgendermaßen:

$$(\mathbf{X}'\mathbf{X})' = \mathbf{X}'(\mathbf{X}')' = \mathbf{X}'\mathbf{X}. \tag{A.41}$$

Der Rang von $\mathbf{X}'\mathbf{X}$ ist gleich p. Da $\mathbf{X}$ den Spaltenrang p besitzt, gilt

$$\mathbf{X}\mathbf{y} = \mathbf{0} \Rightarrow \mathbf{y} = \mathbf{0}.$$

Wir haben zu zeigen

$$\mathbf{X}'\mathbf{X}\mathbf{y} = \mathbf{0} \Rightarrow \mathbf{y} = \mathbf{0}.$$

Es gilt

$$\begin{aligned}
\mathbf{X}'\mathbf{X}\mathbf{y} = \mathbf{0} &\Rightarrow \mathbf{y}'\mathbf{X}'\mathbf{X}\mathbf{y} = \mathbf{0} \\
&\Rightarrow (\mathbf{X}\mathbf{y})'\mathbf{X}\mathbf{y} = \mathbf{0} \\
&\Rightarrow \mathbf{X}\mathbf{y} = \mathbf{0} \\
&\Rightarrow \mathbf{y} = \mathbf{0}.
\end{aligned}$$

Da $\mathbf{X}'\mathbf{X}$ eine (p, p)-Matrix mit Rang p ist, existiert $(\mathbf{X}'\mathbf{X})^{-1}$. Die Matrix $(\mathbf{X}'\mathbf{X})^{-1}$ ist symmetrisch. Mit (A.28) und (A.41) gilt nämlich

$$\left((\mathbf{X}'\mathbf{X})^{-1}\right)' = \left((\mathbf{X}'\mathbf{X})'\right)^{-1} = (\mathbf{X}'\mathbf{X})^{-1}. \tag{A.42}$$

A.1.8 Eigenwerte und Eigenvektoren

Bei einer Reihe multivariater Verfahren benötigt man die *Eigenwerte* und *Eigenvektoren* einer symmetrischen (n, n)-Matrix $\mathbf{A}$.

Definition 42. *Sei* $\mathbf{A}$ *eine* (n,n)*-Matrix. Erfüllen ein Skalar* λ *und ein* n*-dimensionaler Spaltenvektor* $\mathbf{u}$ *mit* $\mathbf{u} \neq \mathbf{0}$ *das Gleichungssystem*

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \tag{A.43}$$

so heißt λ *Eigenwert von* $\mathbf{A}$ *und* $\mathbf{u}$ *zugehöriger Eigenvektor von* $\mathbf{A}$*.*

Erfüllt ein Vektor $\mathbf{u}$ die Gleichung (A.43), so erfüllt auch jedes Vielfache von $\mathbf{u}$ die Gleichung (A.43).

Um eine Lösung von (A.43) zu erhalten, formen wir (A.43) um zu

$$(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{u} = \mathbf{0}. \tag{A.44}$$

Für festes λ ist (A.44) ein linear homogenes Gleichungssystem. Dieses besitzt genau dann Lösungen, die ungleich dem Nullvektor sind, wenn die Spalten von $\mathbf{A} - \lambda\mathbf{I}_n$ linear abhängig sind. Dies ist genau dann der Fall, wenn gilt

$$|\mathbf{A} - \lambda\mathbf{I}_n| = 0. \tag{A.45}$$

Gleichung (A.45) ist ein Polynom n-ten Grades in λ. Dieses besitzt genau n Nullstellen. Die Nullstellen $\lambda_1, \ldots, \lambda_n$ des Polynoms sind also die Eigenwerte der Matrix $\mathbf{A}$. Diese Eigenwerte müssen nicht notwendigerweise verschieden sein. Im Folgenden seien die Eigenwerte der Größe nach durchnummeriert, wobei der erste Eigenwert der größte ist.

Example 65. Wir bestimmen die Eigenwerte der Matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Es gilt

$$\mathbf{A} - \lambda\mathbf{I}_2 = \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix}.$$

Somit gilt

$$|\mathbf{A} - \lambda\mathbf{I}_2| = (2-\lambda)^2 - 1\,.$$

Ein Eigenwert λ erfüllt also die Gleichung

$$(2-\lambda)^2 - 1 = 0. \tag{A.46}$$

Die Nullstellen von (A.46) und somit die Eigenwerte von $\mathbf{A}$ sind $\lambda_1 = 3$ und $\lambda_2 = 1$. □

Die Eigenvektoren zum Eigenwert λ_i, $i = 1, \ldots, n$ erhalten wir dadurch, dass wir λ_i in Gleichung (A.44) für λ einsetzen und die Lösungsmenge des dadurch entstandenen linear homogenen Gleichungssystems bestimmen. hmcounterend. (fortgesetzt)

Example 65. Beginnen wir mit $\lambda_1 = 3$. Der zu $\lambda_1 = 3$ gehörende Eigenvektor

$$\mathbf{u}_1 = \begin{pmatrix} u_{11} \\ u_{21} \end{pmatrix}$$

erfüllt also das Gleichungssystem

$$(\mathbf{A} - 3\mathbf{I}_2)\mathbf{u} = \mathbf{0}.$$

Wegen

$$\mathbf{A} - 3\mathbf{I}_2 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

ergibt sich

$$\begin{aligned} -u_{11} + u_{21} &= 0, \\ u_{11} - u_{21} &= 0\,. \end{aligned}$$

Für die Komponenten des Eigenvektors $\mathbf{u}_1$ zum Eigenwert $\lambda_1 = 3$ muss also gelten $u_{11} = u_{21}$. Der Vektor

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

und alle Vielfachen dieses Vektors sind Eigenvektoren zum Eigenwert $\lambda_1 = 3$. Analoge Berechnungen zum Eigenwert $\lambda_2 = 1$ ergeben, dass für die Komponenten u_{12} und u_{22} des zu $\lambda_2 = 1$ gehörenden Eigenvektors $\mathbf{u}_2$ die Beziehung $u_{12} = -u_{22}$ gelten muss. Der Vektor

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

und alle Vielfachen dieses Vektors sind Eigenvektoren zum Eigenwert $\lambda_2 = 1$. □

In der multivariaten Analyse sollen die Eigenvektoren normiert sein. hmcounterend. (fortgesetzt)

Example 65. Die normierten Eigenvektoren von $\mathbf{A}$ sind

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

und

$$\mathbf{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

□

Folgende Eigenschaften von Eigenwerten und Eigenvektoren sind wichtig:

1. Ist λ ein Eigenwert von $\mathbf{A}$ und k eine reelle Zahl, so ist $k\lambda$ ein Eigenwert von $k\mathbf{A}$. Dies ist offensichtlich.
2. Der Rang der symmetrischen (n,n)-Matrix $\mathbf{A}$ ist gleich der Anzahl der von 0 verschiedenen Eigenwerte von $\mathbf{A}$. Der Beweis ist bei Basilevsky (1983), S. 201 zu finden.
3. Die Eigenwerte einer symmetrischen Matrix sind alle reell. Der Beweis ist bei Basilevsky (1983), S. 199 zu finden.
4. Die Eigenvektoren einer symmetrischen Matrix, die zu unterschiedlichen Eigenwerten gehören, sind orthogonal. Der Beweis ist bei Basilevsky (1983), S. 200 zu finden.
5. $$tr(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i \tag{A.47}$$
6. $$|\mathbf{A}| = \lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_n \tag{A.48}$$

Wir beweisen die Eigenschaften 5. am Ende des nächsten Abschnitts für eine symmetrische Matrix $\mathbf{A}$. Der Beweis von 6. ist bei Basilevsky (1983), S.200 zu finden.

A.1.9 Die Spektralzerlegung einer symmetrischen Matrix

Wir gehen zunächst davon aus, dass die Eigenwerte λ_i, $i = 1, \ldots, n$ der symmetrischen (n,n)-Matrix $\mathbf{A}$ alle unterschiedlich sind. Sei $\mathbf{u}_i$ der normierte Eigenvektor zum Eigenwert λ_i, $i = 1, \ldots, n$. Die Eigenwerte λ_i und Eigenvektoren $\mathbf{u}_i$ erfüllen für $i = 1, \ldots, n$ die Gleichungen

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i. \tag{A.49}$$

Mit $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ und

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix}$$

können wir diese Gleichungen auch folgendermaßen kompakt schreiben:

$$\mathbf{A}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}. \tag{A.50}$$

Da bei einer symmetrischen Matrix Eigenvektoren zu unterschiedlichen Eigenwerten orthogonal sind, ist die Matrix $\mathbf{U}$ eine Orthogonalmatrix. Es gilt also

$$\mathbf{U}\mathbf{U}' = \mathbf{U}'\mathbf{U} = \mathbf{I}_n.$$

Multiplizieren wir (A.50) von rechts mit $\mathbf{U}'$, so erhalten wir:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'. \tag{A.51}$$

Gleichung (A.51) nennt man die *Spektralzerlegung* der symmetrischen (n, n)-Matrix $\mathbf{A}$. Diese Zerlegung ist auch möglich, wenn nicht alle Eigenwerte unterschiedlich sind. Ein Beweis des allgemeinen Falles ist bei Jänich (2000), S.218 ff. zu finden. hmcounterend. (fortgesetzt)

Example 65. Es gilt

$$\boldsymbol{\Lambda} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

und

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Somit gilt

$$\begin{aligned} \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}' &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 3 \\ 1 & -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}. \end{aligned}$$

□

Wir können die Gleichung (A.51) in Abhängigkeit von den Eigenwerten $\lambda_1, \ldots, \lambda_n$ und den Eigenvektoren $\mathbf{u}_1, \ldots, \mathbf{u}_n$ auch folgendermaßen schreiben:

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \, \mathbf{u}_i \mathbf{u}_i'. \tag{A.52}$$

Dies sieht man folgendermaßen:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}' = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1' \\ \mathbf{u}_2' \\ \vdots \\ \mathbf{u}_n' \end{pmatrix}$$

$$= (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n) \begin{pmatrix} \lambda_1 \mathbf{u}_1' \\ \lambda_2 \mathbf{u}_2' \\ \vdots \\ \lambda_n \mathbf{u}_n' \end{pmatrix} = \sum_{i=1}^{n} \lambda_i \, \mathbf{u}_i \mathbf{u}_i'.$$

Wir können die Matrix $\mathbf{A}$ also als Summe von Matrizen darstellen. Sind die ersten beiden Eigenwerte groß im Verhältnis zu den restlichen Eigenwerten, so reichen vielleicht schon die ersten beiden Summmanden und somit die ersten beiden Eigenvektoren zur Approximation von $\mathbf{A}$.

Wir wollen nun noch (A.47) beweisen für den Fall, dass die Matrix $\mathbf{A}$ symmetrisch ist. Es gilt

$$\begin{aligned} tr(\mathbf{A}) &= tr(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}') && \text{(A.53)} \\ &= tr(\mathbf{U}'\mathbf{U}\boldsymbol{\Lambda}) && \text{(A.54)} \\ &= tr(\boldsymbol{\Lambda}) = \sum_{i=1}^{n} \lambda_i. \end{aligned}$$

Beim Übergang von (A.53) zu (A.54) wird (A.33) benutzt. Beim Übergang von (A.54) zu (A.55) wird (A.30) benutzt.

A.1.10 Die Singulärwertzerlegung

Wir haben im letzten Abschnitt gesehen, dass man eine symmetrische Matrix so in das Produkt von drei Matrizen zerlegen kann, dass man die Matrix durch eine Summe von einfachen Matrizen schreiben kann. Eine ähnlich nützliche Zerlegung ist für jede (n,p)-Matrix $\mathbf{A}$ mit $rg(\mathbf{A}) = r$ möglich. Zu jeder (n,p)-Matrix $\mathbf{A}$ mit $rg(\mathbf{A}) = r$ existiert eine orthogonale (n,n)-Matrix $\mathbf{U}$, eine orthogonale (p,p)-Matrix $\mathbf{V}$ und eine (n,p)-Matrix $\mathbf{D}$ mit

$$\mathbf{D} = \begin{pmatrix} d_1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_r & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix},$$

sodass gilt

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'. \qquad \text{(A.55)}$$

Dabei sind die Spalten der Matrix $\mathbf{U}$ die Eigenvektoren der Matrix $\mathbf{A}\mathbf{A}'$ und die Spalten der Matrix $\mathbf{V}$ die Eigenvektoren der Matrix $\mathbf{A}'\mathbf{A}$. $d_1, \dots, d_r$ sind die positiven Quadratwurzeln aus den positiven Eigenwerten der Matrix $\mathbf{A}\mathbf{A}'$ beziehungsweise $\mathbf{A}'\mathbf{A}$. Man nennt (A.55) auch die *Singulärwertzerlegung* der Matrix $\mathbf{A}$. Eine ausführliche Darstellung der Singulärwertzerlegung unter Berücksichtigung von Anwendungen ist bei Watkins (1991), S. 390-430 zu finden.

A.1.11 Quadratische Formen

Wir benötigen in diesem Buch an einigen Stellen quadratische Formen.

Definition 43. *Sei* $\mathbf{x}$ *ein* n*-dimensionaler Vektor und* $\mathbf{A}$ *eine symmetrische* (n, n)*-Matrix. Dann heißt*

$$Q = \mathbf{x}'\mathbf{A}\mathbf{x} \tag{A.56}$$

quadratische Form in den Variablen $x_1, \ldots, x_n$*.*

Example 66. Sei

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Es gilt

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2x_1 + x_2 \\ x_1 + 2x_2 \end{pmatrix}$$

$$= 2x_1^2 + 2x_1x_2 + 2x_2^2 .$$

□

In einer Reihe von Situationen benötigen wir die Definitheit einer Matrix.

Definition 44. *Die Matrix* $\mathbf{A}$ *heißt positiv definit, wenn* $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ *für alle* $\mathbf{x} \neq \mathbf{0}$ *gilt.*
Die Matrix $\mathbf{A}$ *heißt positiv semidefinit, wenn* $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ *für alle* $\mathbf{x} \neq \mathbf{0}$ *gilt, wobei* $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ *für mindestens ein* $\mathbf{x} \neq \mathbf{0}$ *gilt.*
Die Matrix $\mathbf{A}$ *heißt negativ definit, wenn* $\mathbf{x}'\mathbf{A}\mathbf{x} < 0$ *für alle* $\mathbf{x} \neq \mathbf{0}$ *gilt.*
Die Matrix $\mathbf{A}$ *heißt negativ semidefinit, wenn* $\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0$ *für alle* $\mathbf{x} \neq \mathbf{0}$ *gilt, wobei* $\mathbf{x}'\mathbf{A}\mathbf{x} = 0$ *für mindestens ein* $\mathbf{x} \neq \mathbf{0}$ *gilt.*

hmcounterend. (fortgesetzt)

Example 66. Es gilt

$$2x_1^2 + 2x_1x_2 + 2x_2^2 = x_1^2 + 2x_1x_2 + x_2^2 + x_1^2 + x_2^2 = (x_1 + x_2)^2 + x_1^2 + x_2^2.$$

Dieser Ausdruck ist nichtnegativ. Er wird nur Null, wenn gilt $x_1 = x_2 = 0$. Somit ist $\mathbf{A}$ positiv definit. □

Die Definitheit einer Matrix lässt sich auch über die Eigenwerte charakterisieren. Es gilt

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}',$$

wobei die Spalten von $\mathbf{U}$ die normierten Eigenvektoren enthalten und die Hauptdiagonalelemente der Diagonalmatrix $\boldsymbol{\Lambda}$ die Eigenwerte $\lambda_1, \ldots, \lambda_n$ von $\mathbf{A}$ sind. Somit gilt

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'\mathbf{x} = (\mathbf{U}'\mathbf{x})'\boldsymbol{\Lambda}\mathbf{U}'\mathbf{x} = \mathbf{z}'\boldsymbol{\Lambda}\mathbf{z} = \sum_{i=1}^{n} \lambda_i z_i^2$$

mit

$$\mathbf{z} = \mathbf{U}'\mathbf{x}.$$

Somit ist die symmetrische (n, n)-Matrix

- positiv definit, wenn alle Eigenwerte größer als Null sind,
- positiv semidefinit, wenn alle Eigenwerte größer gleich Null sind und mindestens ein Eigenwert gleich Null ist,
- negativ definit, wenn alle Eigenwerte kleiner als Null sind,
- negativ semidefinit, wenn alle Eigenwerte kleiner gleich Null sind und mindestens ein Eigenwert gleich Null ist.

hmcounterend. (fortgesetzt)

Example 66. Wir haben in Beispiel 65 die Eigenwerte von $\mathbf{A}$ bestimmt. Da diese positiv sind, ist $\mathbf{A}$ positiv definit. □

A.2 Extremwerte

Wir müssen an einigen Stellen in diesem Buch Extremwerte von Funktionen mehrerer Veränderlicher bestimmen. Dabei betrachten wir Funktionen

$$f : D \subset \mathbb{R}^n \to \mathbb{R}\,.$$

Example 67. Wir betrachten die Funktion $f : \mathbb{R}^2 \to \mathbb{R}$ mit

$$\mathbf{x} \mapsto f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$$

mit

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Wir können $f(\mathbf{x})$ auch explizit in Abhängigkeit von den Komponenten x_1 und x_2 von $\mathbf{x}$ schreiben. Es gilt

$$f(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = 2x_1^2 + 2x_1x_2 + 2x_2^2.$$

□

Die ϵ-Umgebung $U_\epsilon(\mathbf{x}_0)$ eines Punktes $\mathbf{x}_0 \in \mathbb{R}^n$ ist definiert durch

$$U_\epsilon(\mathbf{x}_0) = \{\mathbf{x} \,|\, \|\mathbf{x} - \mathbf{x}_0\| < \epsilon\}\ .$$

Definition 45. *Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ besitzt in $\mathbf{x}_0$ ein lokales Minimum, wenn eine ϵ-Umgebung $U_\epsilon(\mathbf{x}_0)$ von $\mathbf{x}_0$ existiert, sodass für alle $\mathbf{x} \in U_\epsilon(\mathbf{x}_0)$ mit $\mathbf{x} \neq \mathbf{x}_0$ gilt*

$$f(\mathbf{x}_0) < f(\mathbf{x}).$$

Die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ besitzt in $\mathbf{x}_0$ ein lokales Maximum, wenn eine ϵ-Umgebung $U_\epsilon(\mathbf{x}_0)$ von $\mathbf{x}_0$ existiert, sodass für alle $\mathbf{x} \in U_\epsilon(\mathbf{x}_0)$ mit $\mathbf{x} \neq \mathbf{x}_0$ gilt

$$f(\mathbf{x}_0) > f(\mathbf{x}).$$

A.2.1 Der Gradient und die Hesse-Matrix

Eine notwendige Bedingung für einen Extremwert in x_0 einer in x_0 differenzierbaren Funktion $f : D \subset \mathbb{R} \to \mathbb{R}$ ist $f'(x_0) = 0$. Dabei ist $f'(x_0)$ die erste Ableitung von f an der Stelle x_0. Diese ist folgendermaßen definiert:

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

Dieses Konzept kann auf eine Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ übertragen werden. Die partielle Ableitung von f nach x_i an der Stelle $\mathbf{x}_0$ ist definiert durch

$$\frac{\partial f(\mathbf{x}_0)}{\partial x_i} = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h}.$$

Dabei ist $\mathbf{e}_i$ der i-te Einheitsvektor. Wir sagen, dass die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ in $\mathbf{x}_0$ nach der i-ten Komponente partiell differenzierbar ist, wenn $\frac{\partial f(\mathbf{x}_0)}{\partial x_i}$ existiert. hmcounterend. (fortgesetzt)

Example 67. Es gilt

$$\frac{\partial f(\mathbf{x})}{\partial x_1} = 4x_1 + 2x_2$$

und

$$\frac{\partial f(\mathbf{x})}{\partial x_2} = 2\,x_1 + 4\,x_2\,.$$

□

Ist die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ nach jeder Komponente von $\mathbf{x}$ partiell differenzierbar, dann heißt der Vektor

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

Gradient der Funktion. hmcounterend. (fortgesetzt)

Example 67. Es gilt

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} 4\,x_1 + 2\,x_2 \\ 2\,x_1 + 4\,x_2 \end{pmatrix}.$$

□

Schauen wir uns den Gradienten spezieller Funktionen an. Sei

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}.$$

Dann gilt

$$\frac{\partial\, \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}\,. \tag{A.57}$$

Es gilt nämlich

$$\frac{\partial\, \mathbf{a}'\mathbf{x}}{\partial x_i} = \frac{\partial\, a_1 x_1 + \ldots + a_n x_n}{\partial x_i} = a_i\,.$$

Ist $\mathbf{A}$ eine symmetrische Matrix, so gilt

$$\frac{\partial\, \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\,\mathbf{A}\mathbf{x}\,. \tag{A.58}$$

Der Beweis ist bei Büning et al. (2000), S. 144-145 zu finden.

Man kann bei einer Funktion $f : D \subset I\!R^n \to I\!R$ auch partielle Ableitungen höherer Ordnung betrachten. Existieren die partiellen Ableitungen zweiter Ordnung und sind sie stetig, so heißt die Matrix der partiellen Ableitungen zweiter Ordnung *Hesse-Matrix*:

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix}. \tag{A.59}$$

hmcounterend. (fortgesetzt)

Example 67. Es gilt

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} = \frac{\partial\, 4x_1 + 2x_2}{\partial x_1} = 4,$$

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_1\, \partial x_2} = \frac{\partial\, 4x_1 + 2x_2}{\partial x_2} = 2,$$

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_2\, \partial x_1} = \frac{\partial\, 2x_1 + 4x_2}{\partial x_1} = 2,$$

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} = \frac{\partial\, 2x_1 + 4x_2}{\partial x_2} = 4\,.$$

Es gilt also

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}.$$

□

A.2.2 Extremwerte ohne Nebenbedingungen

Wir suchen in diesem Buch Extremwerte von Funktionen $f : D \subset \mathbb{R}^n \to \mathbb{R}$, deren erste und zweite partielle Ableitungen existieren und stetig sind. Eine notwendige Bedingung dafür, dass die Funktion $f : D \subset \mathbb{R}^n \to \mathbb{R}$ einen Extremwert an der Stelle $\mathbf{x}_0$ hat, ist

$$\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{x}} = \mathbf{0}.$$

Der Beweis ist bei Khuri (1993), S. 283 zu finden. hmcounterend. (fortgesetzt)

Example 67. Notwendige Bedingungen für Extremwerte von

$$f(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = 2x_1^2 + 2x_1x_2 + 2x_2^2$$

sind

$$\begin{aligned} 4\,x_1 + 2\,x_2 &= 0, \\ 2\,x_1 + 4\,x_2 &= 0. \end{aligned}$$

Dieses linear homogene Gleichungssystem hat die Lösung $x_1 = x_2 = 0$. □

Um zu überprüfen, ob in $\mathbf{x}_0$ ein Extremwert vorliegt, bestimmt man $\mathbf{H}(\mathbf{x}_0)$. Ist $\mathbf{H}(\mathbf{x}_0)$ negativ definit, so liegt ein lokales Maximum vor. Ist $\mathbf{H}(\mathbf{x}_0)$ positiv definit, so liegt ein lokales Minimum vor. Der Beweis ist bei Khuri (1993), S. 283-284 zu finden. hmcounterend. (fortgesetzt)

Example 67. Es gilt

$$\mathbf{H}(x_1, x_2) = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}.$$

Es gilt $\mathbf{H}(\mathbf{x}) = 2\mathbf{A}$ mit

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Wir haben die Eigenwerte von $\mathbf{A}$ bereits im Beispiel 65 bestimmt. Da die Eigenwerte von $2\mathbf{A}$ doppelt so groß wie die Eigenwerte von $\mathbf{A}$ sind, sind auch beide Eigenwerte von $2\mathbf{A}$ positiv. Also liegt ein lokales Minimum vor. □

A.2.3 Extremwerte unter Nebenbedingungen

Bei der Optimierung von $f(\mathbf{x})$ müssen oft Nebenbedingungen der Form $g(\mathbf{x}) = 0$ berücksichtigt werden. Zur Bestimmung der Extremwerte stellen wir die Lagrange-Funktion

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda\, g(\mathbf{x})$$

auf.

Eine notwendige Bedingung eines Extremwerts von $f(\mathbf{x})$ in $\mathbf{x}_0$ unter der Nebenbedingung $g(\mathbf{x}) = 0$ ist

$$\frac{\partial L(\mathbf{x}_0, \lambda_0)}{\partial \mathbf{x}} = \mathbf{0}$$

und

$$\frac{\partial L(\mathbf{x}_0, \lambda_0)}{\partial \lambda} = \mathbf{0}.$$

Der Beweis ist bei Khuri (1993), S. 287-290 zu finden.

Example 68. Wir suchen den Extremwert von

$$f(x_1, x_2) = 2x_1^2 + 2x_1x_2 + 2x_2^2$$

unter der Nebenbedingung

$$x_1^2 + x_2^2 = 1.$$

Wir stellen die Lagrange-Funktion auf:

$$L(x_1, x_2, \lambda) = 2x_1^2 + 2x_1x_2 + 2x_2^2 - \lambda(x_1^2 + x_2^2 - 1).$$

Die partiellen Ableitungen lauten:

$$\frac{\partial L(x_1, x_2, \lambda)}{\partial x_1} = 4x_1 + 2x_2 - 2\lambda x_1,$$

$$\frac{\partial L(x_1, x_2, \lambda)}{\partial x_2} = 2x_1 + 4x_2 - 2\lambda x_2$$

und

$$\frac{\partial L(x_1, x_2, \lambda)}{\partial \lambda} = -x_1^2 - x_2^2 + 1.$$

Ein Extremwert

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

muss also die folgenden Gleichungen erfüllen:

$$4x_1 + 2x_2 - 2\lambda x_1 = 0, \tag{A.60}$$

$$2x_1 + 4x_2 - 2\lambda x_2 = 0 \tag{A.61}$$

und

$$x_1^2 + x_2^2 = 1. \tag{A.62}$$

Wir können die Gleichungen (A.60) und (A.61) auch schreiben als:

$$2x_1 + x_2 = \lambda x_1, \tag{A.63}$$

$$x_1 + 2x_2 = \lambda x_2. \tag{A.64}$$

Mit

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

lauten diese Gleichungen in Matrixform

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}.$$

Dies ist aber ein Eigenwertproblem. Ein Eigenvektor von $\mathbf{A}$ erfüllt also die notwendigen Bedingungen für einen Extremwert. Da dieser auch die Nebenbedingung erfüllen muss, müssen wir ihn normieren. Die notwendigen Bedingungen erfüllen also die Punkte

$$\mathbf{x}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

und

$$\mathbf{x}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

□

Auf die hinreichenden Bedingungen für einen Extremwert unter Nebenbedingungen wollen wir hier nicht eingehen. Sie sind bei Wetzel et al. (1981) zu finden.

A.3 Matrizenrechnung in S-PLUS

In S-PLUS sind alle beschriebenen Konzepte der Matrizenrechnung implementiert. Wir schauen uns die Beispiele aus Kapitel A.1 in S-PLUS an und beginnen mit Vektoren. Wir geben zunächst die Vektoren

$$\mathbf{a}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{a}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

aus Beispiel 64 ein. Dazu verwenden wir die Funktion c:

```
> a1<-c(2,1)
> a2<-c(1,2)
```

Schauen wir uns zunächst Verknüpfungen von Vektoren an. In S-PLUS kann man die Vektoren a1 und a2 mit einem Operator verknüpfen, wenn sie die gleiche Länge besitzen. Das Ergebnis ist ein Vektor a3, der die gleiche Länge besitzt wie die Vektoren a1 und a2. Jedes Element des Vektors a3 erhält man dadurch, dass man die entsprechenden Elemente der Vektoren a1 und a2 mit dem Operator verknüpft. Die Operatoren sind dabei nicht wie in der Matrizenrechnung beschränkt auf + und -. Man kann also auch den Multiplikationsoperator * verwenden. Dieser liefert dann aber nicht das aus der Matrizenrechnung bekannte innere Produkt der Vektoren. Zuerst überprüfen wir aber, ob die beiden Vektoren die gleiche Länge besitzen:

```
> length(a1)
[1] 2
> length(a2)
[1] 2
```

Zum Vergleich der Längen der beiden Vektoren verwenden wir den Vergleichsoperator ==. Beim Vergleich von zwei Skalaren liefert dieser den Wert T, wenn beide identisch sind, ansonsten den Wert F:

```
> 3==(2+1)
[1] T
> 3==(3+1)
[1] F
```

Wir geben also ein

```
> length(a1)==length(a2)
[1] T
```

Schauen wir uns einige Beispiele für Verknüpfungen von Vektoren an:

```
> a1+a2
[1] 3 3
> a1-a2
[1]  1 -1
```

```
> a1*a2
[1] 2 2
> a1==a2
[1] F F
```

Beim letzten Beispiel war das Ergebnis ein logischer Vektor. Um zu überprüfen, ob irgendeine Komponente eines Vektors mit der entsprechenden Komponente eines anderen Vektors übereinstimmt, verwenden wir die Funktion `any`:

```
> any(a1==a2)
[1] F
```

Mit der Funktion `all` können wir überprüfen, ob alle Komponenten übereinstimmen:

```
> all(a1==a2)
[1] F
```

Das innere Produkt der gleich langen Vektoren `a1` und `a2` liefert der Operator `%*%`:

```
> a1%*%a2
     [,1]
[1,]    4
```

Das Ergebnis ist eine Matrix und kein Vektor. Man kann einen Vektor mit einem Skalar verknüpfen. Dabei wird jedes Element jeder Komponente des Vektors mit dem Skalar verknüpft:

```
> 1+a1
[1] 3 2
> 2*a1
[1] 4 2
```

Das äußere Produkt der Vektoren `a1` und `a2` gewinnt man mit der Funktion `outer`:

```
> outer(a1,a2)
      [,1] [,2]
 [1,]    2    4
 [2,]    1    2
> outer(a2,a1)
      [,1] [,2]
 [1,]    2    1
 [2,]    4    2
```

Schauen wir uns Matrizen an. Wir geben die Matrizen

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

aus Beispiel 64 mit der Funktion `matrix` ein. Dabei beachten wir, dass Matrizen spaltenweise aufgefüllt werden:

```
> A<-matrix(c(2,1,1,2),2,2)
> B<-matrix(c(0,1,-1,0),2,2)
> C<-matrix(c(1,1,1,2,1,2),3,2)
> D<-matrix(c(3,0,0,1),2,2)
```

Man kann zwei Matrizen mit einem Operator verknüpfen, wenn sie die gleiche Dimension besitzen. Die Dimension einer Matrix erhält man mit der Funktion `dim`:

```
> dim(C)
[1] 3 2
```

Das Ergebnis der Verknüpfung der Matrizen `A` und `B` ist eine Matrix `C` mit der Dimension der Objekte, die durch den Operator verknüpft werden. Jedes Element der Matrix `C` erhält man dadurch, dass man die entsprechenden Elemente der Matrizen `A` und `B` mit dem Operator verknüpft:

```
> A+B
     [,1] [,2]
[1,]    2    0
[2,]    2    2
> A-B
     [,1] [,2]
[1,]    2    2
[2,]    0    2
> A*B
     [,1] [,2]
[1,]    0   -1
[2,]    1    0
```

Um das Produkt der Matrizen $\mathbf{A}$ und $\mathbf{B}$ bilden zu können, muss die Anzahl der Spalten von $\mathbf{A}$ gleich der Anzahl der Zeilen von $\mathbf{B}$ sein. Ist dies der Fall, so liefert der Operator `%*%` das Produkt:

```
> A%*%B
     [,1] [,2]
[1,]    1   -2
[2,]    2   -1
```

Die Transponierte $\mathbf{A}'$ einer Matrix $\mathbf{A}$ erhält man mit der Funktion `t`:

```
> t(C)
     [,1] [,2] [,3]
[1,]    1    1    1
[2,]    2    1    2
```

Um zu überprüfen, ob eine quadratische Matrix symmetrisch ist, geben wir ein

```
> all(A==t(A))
[1] T
```

Mit der Funktion `diag` kann man eine Diagonalmatrix erzeugen. Der Aufruf

```
> diag(c(3,1))
```

liefert als Ergebnis

```
     [,1] [,2]
[1,]    3    0
[2,]    0    1
```

Die Einheitsmatrix $\mathbf{I}_3$ erhält man durch

```
> diag(3)
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Außerdem kann man mit der Funktion `diag` die Hauptdiagonalelemente einer Matrix extrahieren.

```
> diag(A)
[1] 2 2
```

Die inverse Matrix $\mathbf{A}^{-1}$ erhält man mit der Funktion `solve`:

```
> solve(A)
           [,1]       [,2]
[1,]  0.6666667 -0.3333333
[2,] -0.3333333  0.6666667
```

Der Aufruf

```
> solve(A)%*%A
```

liefert im Rahmen der Rechengenauigkeit die Einheitsmatrix. Mit der Funktion `solve` kann man auch lineare Gleichungssysteme lösen. Hierauf wollen wir aber nicht eingehen.

Die Spur einer quadratischen Matrix erhält man durch

```
> sum(diag(A))
[1] 4
```

Die Funktion `eigen` liefert die Eigenwerte und Eigenvektoren einer quadratischen Matrix. Das Ergebnis von `eigen` ist eine Liste. Die erste Komponente der Liste enthält die Eigenwerte und die zweite die Eigenvektoren.

```
> e<-eigen(A)
> e[[1]]
[1] 3 1
> e[[2]]
           [,1]       [,2]
 [1,] 0.7071068  0.7071068
 [2,] 0.7071068 -0.7071068
```

Wir bilden die orthogonale Matrix `U` mit den Eigenvektoren von `A` in den Spalten:

```
> U<-e[[2]]
```

und eine Diagonalmatrix `L` mit den Eigenwerten von `A`:

```
> L<-diag(e[[1]])
```

Der Aufruf

```
> U%*%L%*%t(U)
```

liefert die Matrix `A`

```
     [,1] [,2]
[1,]    2    1
[2,]    1    2
```

Der Aufruf

```
> e<-svd(C)
```

liefert die Singulärwertzerlegung `UDV'`der Matrix `C`. Das Ergebnis ist eine Liste. Die erste Komponente enthält die positiven Elemente der Matrix `D`, die zweite Komponente die Matrix `V` und die dritte Komponente die Matrix `U`. Wir bilden die Matrizen `U`, `D` und `V`:

```
> U<-e[[3]]
> D<-diag(e[[1]])
> V<-e[[2]]
```

Der Aufruf

```
> U%*%D%*%t(V)
```

liefert die Matrix `C`

```
     [,1] [,2]
[1,]    1    2
[2,]    1    1
[3,]    1    2
```

B S-PLUS-Funktionen

B.1 Quartile

Die Funktion berechnet für den Vektor `x` die Quartile nach der Methode auf Seite 20.

```
quartile<-function(x) {
  # berechnet Quartile
  # x ist Datensatz
  x <- sort(x)
  uq <- median(x[1:ceiling(length(x)/2)])
  x <- rev(x)
  oq <- median(x[1:ceiling(length(x)/2)])
  return(c(uq, oq))
}
```

B.2 Distanzmatrix

Die Funktion `distfull` bildet aus dem Vektor `dis` mit den Distanzen die Distanzmatrix. Dem Attribut `"Size"` von `dis` muss die Anzahl der Beobachtungen zugewiesen sein.

```
> distfull<-function(dis) {
    n <- attr(dis, "Size")
    full <- matrix(0, n, n)
    full[lower.tri(full)] <- dis
    full + t(full)
}
```

B.3 Monotone Regression

Die Funktion `monreg` führt eine monotone Regression für die Elemente des Vektors `p` durch.

```
monreg<-function(p)
{
  # Monotone Regression der Elemente des Vektors p
   g <- rep(1, length(p))
   while(!all(p[ - length(p)] <= p[-1])) {
       i <- pooladjacent(p)
       p <- miblock(p, i, g)
       g <- gew(i, g)
   }
   rep(p, g)
}
```

Sie ruft folgende Funktionen auf:

```
pooladjacent<-function(p)
{
   i <- numeric(0)
   j <- 1
   h <- p
   while(length(h) > 0) {
       a <- sum(cumprod(h[ - length(h)] > h[-1])) + 1
       i <- c(i, rep(j, a))
       h <- h[ - (1:a)]
       j <- j + 1
   }
   i
}
```

```
miblock<-function(p, i, g)
{
   m <- numeric(0)
   for(j in 1:max(i))
       m <- c(m, sum(g[i == j] * p[i == j])/sum(g[i == j]))
   m
}
```

```
> gew<-function(ind, ag)
{
   m <- numeric(0)
   for(i in 1:max(ind))
       m <- c(m, sum(ag[ind == i]))
   m
}
```

B.4 STRESS1

Die Funktion `stress1` berechnet STRESS1. Ihre Argumente sind der Vektor `d` der Distanzen und der Vektor `disp` der Disparitäten.

```
stress1<-function(d,disp)
{sqrt(sum((d-disp)^2)/sum(d^2)) }
```

B.5 Bestimmung einer neuen Konfiguration

Die Funktion `Neuekon` bestimmt bei nichtmetrischer mehrdimensionaler Skalierung eine neue Konfiguration. Das Argument `X` ist die alte Konfiguration und das Argument `delta` ist die Matrix $\boldsymbol{\Delta}$. Das Ergebnis ist eine Liste. Die erste Komponente ist die neue Konfiguration `xneu` und die zweite Komponente der Wert `stress` von STRESS1.

```
> Neuekon<-function(X, delta)
 { # Neue Konfiguration bei einer nichtmetrischen MDS.
   # X: Startkonfiguration
   # delta: Matrix Delta
   # Ergebnis ist Liste mit
   # xneu: neue Konfiguration # stress:
           Wert von STRESS1 fuer diese Konfiguration
    n <- dim(X)[1]
    delta <- delta[lower.tri(delta)]
    d <- dist(X)
    disp <- monreg(d[order(delta)])
    dm <- matrix(0, n, n)
    dm[lower.tri(dm)] <- d
    dm <- dm + t(dm)
    dispm <- matrix(0, n, n)
    dispm[lower.tri(dispm)] <- disp[rank(delta)]
    dispm <- dispm + t(dispm)
    xneu <- matrix(0, n, 2)
    for(i in 1:n) {
        h <- matrix((dm[i,-i] - dispm[i,-i])/dm[i,-i],n-1,2)
                    * (X[-i,]-matrix(X[i,],n-1,2,b = T))
        xneu[i,] <- X[i,] + apply(h,2,mean)
                 }
    d <- dist(xneu)
    disp <- monreg(d[order(delta)])
    stress <- stress1(delta, d, disp)
    list(xneu, stress)
}
```

B.6 Kophenetische Matrix

Die Funktion `cophenetic` bestimmt die kophenetische Matrix. Ihre Argumente sind Ergebnisse der Funktion `hclust`. Das erste Argument ist die Komponente `merge` und das zweite die Komponente `height`.

```
cophenetic<-function(m, h) {
    k <- length(h) + 1
    co <- matrix(0, k, k)
    obj <- abs(m[1,  ])
    grp <- rep(1, 2)
    co[ - m[1, 1],  - m[1, 2]] <- h[1]
    co[ - m[1, 2],  - m[1, 1]] <- h[1]
    for(i in 2:(k - 1)) {
        if(all(m[i,  ] < 0)) {
            obj <- c(obj, abs(m[i,  ]))
            grp <- c(grp, rep(i, 2))
            co[ - m[i, 1],  - m[i, 2]] <- h[i]
            co[ - m[i, 2],  - m[i, 1]] <- h[i]
        }
        else if(all(m[i,  ] > 0)) {
            z <- abs(obj[grp == abs(m[i, 1])])
            s <- abs(obj[grp == abs(m[i, 2])])
            obj <- c(obj, z, s)
            grp <- c(grp, rep(i, length(z) + length(s)))
            co[z, s] <- h[i]
            co[s, z] <- h[i]
        }
        else {
            z <- abs(m[i,  ][m[i,  ] < 0])
            obj <- c(obj, z)
            grp <- c(grp, i)
            pos <- abs(m[i,  ][m[i,  ] > 0])
            s <- abs(obj[grp == pos])
            obj <- c(obj, s)
            grp <- c(grp, rep(i, length(s)))
            co[z, s] <- h[i]
            co[s, z] <- h[i]
        }
    }
    co
}
```

B.7 Gamma-Koeffizient

Die Funktion `gammakoeffizient` bestimmt den Gamma-Koeffizienten zwischen den Vektoren `v1` und `v2`.

```
gammakoeffizient<-function(v1, v2) {
    m1 <- outer(v1, v1, FUN = "<")
    m1 <- m1[lower.tri(m1)]
    m2 <- outer(v2, v2, FUN = "<")
    m2 <- m2[lower.tri(m2)]
    m3 <- outer(v1, v1, FUN = ">")
    m3 <- m3[lower.tri(m3)]
    m4 <- outer(v2, v2, FUN = ">")
    m4 <- m4[lower.tri(m4)]
    C <- sum((m1 + m2) == 2)
    C <- C + sum((m3 + m4) == 2)
    D <- sum((m1 + m4) == 2)
    D <- D + sum((m2 + m3) == 2)
    (C - D)/(C + D)
}
```

B.8 Bestimmung der Zugehörigkeit zu Klassen

Die Funktion `welche.cluster` gibt für jedes Objekt an, zu welcher Klasse es gehört. Ihre Argumente sind Ergebnisse der Funktion `hclust` und die Anzahl `anz` der Klassen. Das erste Argument ist die Komponente `merge` und das zweite die Komponente `height`. Das dritte Argument ist die Anzahl `anz` der Klassen.

```
welche.cluster<-function(mer,hei, anz) {
    co <- cophenetic(mer, hei)
    h <- hei[length(hei) + 1 - anz]
    n <- ncol(co)
    cl <- rep(0, n)
    k <- 1
    for(i in 1:n) {
        if(cl[i] == 0) {
            ind <- (1:n)[co[i,  ] <= h]
            cl[ind] <- k
            k <- k + 1
        }
    }
    cl
}
```

B.9 Silhouette

Die Funktion `silhouette` liefert die Informationen zum Zeichnen einer Silhouette. Das Argument `wo` ist ein Vektor, der für jedes Objekt die Nummer der Klasse enthält, zu der es gehört. Das Argument `d` ist die Distanzmatrix und das Argument `namen` ist ein Vektor mit den Namen der Objekte. Das Ergebnis der Funktion `silhouette` ist eine Liste. Die erste Komponente ist eine Matrix. In der ersten Spalte steht die Nummer der Klasse des Objekts, in der zweiten Spalte die Nummer der Klasse, die am nächsten liegt, und in der dritten Spalte der Wert von $s(i)$. Die Namen der Objekte sind die Namen der ersten Dimension der Matrix. Die zweite Komponente enthält den Vektor der Mittelwerte der $s(i)$ der Klasse, die dritte Komponente den Silhouettenkoeffizienten.

```
silhouette<-function(wo, d, namen) {
   if(is.numeric(namen))
       namen <- as.character(namen)
   anzgr <- max(wo)
   n <- length(wo)
   indgr <- matrix(0, anzgr, 2)
   gruppen <- numeric(0)
   mi <- 1
   for(k in 1:anzgr) {
       g <- (1:n)[wo == k]
       indgr[k,  ] <- c(mi, mi + length(g) - 1)
       mi <- mi + length(g)
       gruppen <- c(gruppen, g)
   }
   b <- rep(0, n)
   a <- rep(0, n)
   naechstes <- rep(0, n)
   for(i in 1:n) {
       andere <- (1:anzgr)[ - wo[i]]
       bgr <- rep(0, length(andere))
       for(j in 1:length(andere))
           bgr[j] <- mean(d[i, gruppen[indgr[andere[j], 1]:
                                       indgr[andere[j], 2]]])
       b[i] <- min(bgr)
       naechstes[i] <- andere[bgr == b[i]]
       eigene <- gruppen[indgr[wo[i], 1]:indgr[wo[i], 2]]
       if(length(eigene) == 1)
           a[i] <- b[i]
       else a[i] <- mean(d[i, eigene[eigene != i]])
   }
   si <- (b - a)/pmax(a, b)
```

```
    siclu <- rep(0, anzgr)
    for(l in 1:anzgr)
        siclu[l] <- mean(si[wo == l])
    m <- cbind(wo, naechstes, si)
    namen <- namen[order(m[, 1])]
    m <- m[order(m[, 1]),  ]
    ms <- numeric(0)
    namens <- numeric(0)
    clusmittel <- rep(0, anzgr)
    for(i in 1:anzgr) {
        h <- matrix(m[m[, 1] == i,  ], ncol = 3)
        clusmittel[i] <- mean(h[, 3])
        n <- namen[m[, 1] == i]
        ms <- rbind(ms, h[rev(order(h[, 3])),  ])
        namens <- c(namens, n[rev(order(h[, 3]))])
    }
    dimnames(ms) <- list(namens, dimnames(ms)[[2]])
    list(ms, clusmittel, mean(ms[, 3]))
}
```

B.10 Zeichnen einer Silhouette

Die Funktion `plotsilhouette` zeichnet eine Silhouette. Ihr Argument ist das Ergebnis der Funktion `silhouette`. Sie beruht auf der Funktion `plot.partition` aus der Library `cluster` und wurde für unsere Zwecke angepaßt.

```
 plotsilhouette<-function(silinfo) {
    S <- rev(silinfo[[1]][, 3])
    space <- c(0, rev(diff(silinfo[[1]][, 1])))
    names <- rev(dimnames(silinfo[[1]])[[1]])
    if(!is.character(names))
        names <- as.character(names)
    barplot(S, space = space, names = names,
    xlab = "Breite der Silhouette", ylab = "",
    xlim = c(min(0, min(S)), 1), horiz = T,
    mgp = c(2.5, 1, 0))
    invisible()
}
```

C Tabellen

C.1 Standardnormalverteilung

Table C.1. Quantil z_p der Standardnormalverteilung

p	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
0.50	0.000	0.002	0.005	0.008	0.010	0.012	0.015	0.018	0.020	0.023
0.51	0.025	0.028	0.030	0.033	0.035	0.038	0.040	0.043	0.045	0.048
0.52	0.050	0.053	0.055	0.058	0.060	0.063	0.065	0.068	0.070	0.073
0.53	0.075	0.078	0.080	0.083	0.085	0.088	0.090	0.093	0.095	0.098
0.54	0.100	0.103	0.106	0.108	0.110	0.113	0.116	0.118	0.121	0.123
0.55	0.126	0.128	0.131	0.133	0.136	0.138	0.141	0.143	0.146	0.148
0.56	0.151	0.154	0.156	0.159	0.161	0.164	0.166	0.169	0.171	0.174
0.57	0.176	0.179	0.182	0.184	0.187	0.189	0.192	0.194	0.197	0.199
0.58	0.202	0.204	0.207	0.210	0.212	0.215	0.217	0.220	0.222	0.225
0.59	0.228	0.230	0.233	0.235	0.238	0.240	0.243	0.246	0.248	0.251
0.60	0.253	0.256	0.258	0.261	0.264	0.266	0.269	0.272	0.274	0.277
0.61	0.279	0.282	0.284	0.287	0.290	0.292	0.295	0.298	0.300	0.303
0.62	0.306	0.308	0.311	0.313	0.316	0.319	0.321	0.324	0.327	0.329
0.63	0.332	0.334	0.337	0.340	0.342	0.345	0.348	0.350	0.353	0.356
0.64	0.358	0.361	0.364	0.366	0.369	0.372	0.374	0.377	0.380	0.383
0.65	0.385	0.388	0.391	0.393	0.396	0.399	0.402	0.404	0.407	0.410
0.66	0.412	0.415	0.418	0.421	0.423	0.426	0.429	0.432	0.434	0.437
0.67	0.440	0.443	0.445	0.448	0.451	0.454	0.456	0.459	0.462	0.465
0.68	0.468	0.470	0.473	0.476	0.479	0.482	0.484	0.487	0.490	0.493
0.69	0.496	0.499	0.501	0.504	0.507	0.510	0.513	0.516	0.519	0.522
0.70	0.524	0.527	0.530	0.533	0.536	0.539	0.542	0.545	0.548	0.550
0.71	0.553	0.556	0.559	0.562	0.565	0.568	0.571	0.574	0.577	0.580
0.72	0.583	0.586	0.589	0.592	0.595	0.598	0.601	0.604	0.607	0.610
0.73	0.613	0.616	0.619	0.622	0.625	0.628	0.631	0.634	0.637	0.640
0.74	0.643	0.646	0.650	0.653	0.656	0.659	0.662	0.665	0.668	0.671

Table C.2. Quantil z_p der Standardnormalverteilung

p	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
0.75	0.674	0.678	0.681	0.684	0.687	0.690	0.694	0.697	0.700	0.703
0.76	0.706	0.710	0.713	0.716	0.719	0.722	0.726	0.729	0.732	0.736
0.77	0.739	0.742	0.745	0.749	0.752	0.755	0.759	0.762	0.766	0.769
0.78	0.772	0.776	0.779	0.782	0.786	0.789	0.793	0.796	0.800	0.803
0.79	0.806	0.810	0.813	0.817	0.820	0.824	0.827	0.831	0.834	0.838
0.80	0.842	0.845	0.849	0.852	0.856	0.860	0.863	0.867	0.870	0.874
0.81	0.878	0.882	0.885	0.889	0.893	0.896	0.900	0.904	0.908	0.912
0.82	0.915	0.919	0.923	0.927	0.931	0.935	0.938	0.942	0.946	0.950
0.83	0.954	0.958	0.962	0.966	0.970	0.974	0.978	0.982	0.986	0.990
0.84	0.994	0.999	1.003	1.007	1.011	1.015	1.019	1.024	1.028	1.032
0.85	1.036	1.041	1.045	1.049	1.054	1.058	1.062	1.067	1.071	1.076
0.86	1.080	1.085	1.089	1.094	1.098	1.103	1.108	1.112	1.117	1.122
0.87	1.126	1.131	1.136	1.141	1.146	1.150	1.155	1.160	1.165	1.170
0.88	1.175	1.180	1.185	1.190	1.195	1.200	1.206	1.211	1.216	1.221
0.89	1.226	1.232	1.237	1.243	1.248	1.254	1.259	1.265	1.270	1.276
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.316	1.322	1.328	1.335
0.91	1.341	1.347	1.353	1.360	1.366	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.432	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.498	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290
0.99	2.326	2.366	2.409	2.457	2.512	2.576	2.652	2.748	2.878	3.090

C.2 χ^2-Verteilung

Table C.3. Quantile der χ^2-Verteilung mit k Freiheitsgraden

k	$\chi^2_{k;0.95}$	$\chi^2_{k;0.975}$	$\chi^2_{k;0.9833}$	$\chi^2_{k;0.9875}$	$\chi^2_{k;0.99}$
1	3.84	5.02	5.73	6.24	6.63
2	5.99	7.38	8.19	8.76	9.21
3	7.81	9.35	10.24	10.86	11.34
4	9.49	11.14	12.09	12.76	13.28
5	11.07	12.83	13.84	14.54	15.09
6	12.59	14.45	15.51	16.24	16.81
7	14.07	16.01	17.12	17.88	18.48
8	15.51	17.53	18.68	19.48	20.09
9	16.92	19.02	20.21	21.03	21.67
10	18.31	20.48	21.71	22.56	23.21
11	19.68	21.92	23.18	24.06	24.72
12	21.03	23.34	24.63	25.53	26.22
13	22.36	24.74	26.06	26.98	27.69
14	23.68	26.12	27.48	28.42	29.14
15	25.00	27.49	28.88	29.84	30.58
16	26.30	28.85	30.27	31.25	32.00
17	27.59	30.19	31.64	32.64	33.41
18	28.87	31.53	33.01	34.03	34.81
19	30.14	32.85	34.36	35.40	36.19
20	31.41	34.17	35.70	36.76	37.57
21	32.67	35.48	37.04	38.11	38.93
22	33.92	36.78	38.37	39.46	40.29
23	35.17	38.08	39.68	40.79	41.64
24	36.42	39.36	41.00	42.12	42.98
25	37.65	40.65	42.30	43.45	44.31

C.3 *t*-Verteilung

Table C.4. Quantile der t-Verteilung mit k Freiheitsgraden

k	$t_{k;0.90}$	$t_{k;0.95}$	$t_{k;0.975}$	$t_{k;0.99}$	$t_{k;0.995}$
1	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.3114	1.6991	2.0452	2.4620	2.7564
30	1.3104	1.6973	2.0423	2.4573	2.7500

C.4 *F*-Verteilung

Table C.5. Das 0.95-Quantil $F_{m,n;0.95}$ der F-Verteilung mit m und n Freiheitsgraden

n \ m	1	2	3	4	5	6	7	8	9	10
1	161.45	199.5	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16

Table C.6. Das 0.95-Quantil $F_{m,n;0.95}$ der F-Verteilung mit m und n Freiheitsgraden

n \ m	1	2	3	4	5	6	7	8	9	10
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11
37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09
39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
41	4.08	3.23	2.83	2.60	2.44	2.33	2.24	2.17	2.12	2.07
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06
43	4.07	3.21	2.82	2.59	2.43	2.32	2.23	2.16	2.11	2.06
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04
47	4.05	3.20	2.80	2.57	2.41	2.30	2.21	2.14	2.09	2.04
48	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03
49	4.04	3.19	2.79	2.56	2.40	2.29	2.20	2.13	2.08	2.03
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
51	4.03	3.18	2.79	2.55	2.40	2.28	2.20	2.13	2.07	2.02
52	4.03	3.18	2.78	2.55	2.39	2.28	2.19	2.12	2.07	2.02
53	4.02	3.17	2.78	2.55	2.39	2.28	2.19	2.12	2.06	2.01
54	4.02	3.17	2.78	2.54	2.39	2.27	2.18	2.12	2.06	2.01
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01
56	4.01	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.05	2.00
57	4.01	3.16	2.77	2.53	2.38	2.26	2.18	2.11	2.05	2.00
58	4.01	3.16	2.76	2.53	2.37	2.26	2.17	2.10	2.05	2.00
59	4.00	3.15	2.76	2.53	2.37	2.26	2.17	2.10	2.04	2.00
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99

References

Agresti, A. (1990): Categorical data analysis. Wiley, New York

Andersen, E. B. (1991): The statistical analysis of categorical data. Springer, Berlin, 2nd edition

Bacher, J. (1994): Clusteranalyse: anwendungsorientierte Einführung. Oldenbourg, München

Bankhofer, U. (1995): Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse. Eul, Bergisch Gladbach

Basilevsky, A. (1983): Applied matrix algebra in the statistical sciences. North-Holland, New York

Basilevsky, A. (1994): Statistical factor analysis and related methods: theory and applications. Wiley, New York

Birkes, D., Dodge, Y. (1993): Alternative methods of regression. Wiley, New York

Bödeker, M., Franke, K. (2001): Analyse der Potenziale und Grenzen von Virtual Reality Technologien auf industriellen Anwendermärkten. Diplomarbeit, Universität Bielefeld

Bollen, K. A. (1989): Structural equations with latent variables. Wiley, New York

Borg, I., Groenen, P. (1997): Modern multidimensional scaling: theory and applications. Springer, New York

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984): Classification and regression trees. Wadsworth, Belmont

Brühl, O., Kahn, T. (2001): Analyse der Standortqualität zur Beurteilung der wirtschaftlichen Leistungsfähigkeit im interregionalen Vergleich. Diplomarbeit, Universität Bielefeld

Büning, H. (1991): Robuste und adaptive Tests. de Gruyter, Berlin

Büning, H. (1996): Adaptive tests for the c-sample location problem - the case of two-sided alternatives. Communications in Statistics - Theory and Methods, **25**, 1569–1582

Büning, H., Naeve, P., Trenkler, G., Waldmann, K.-H. (2000): Mathematik für Ökonomen im Hauptstudium. Oldenbourg, München

Büning, H., Trenkler, G. (1994): Nichtparametrische statistische Methoden. de Gruyter, Berlin, 2nd edition

Calinski, T., Harabasz, J. (1974): A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods A, **3**, 1–27

Carroll, J. D., Chang, J. J. (1970): Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika, **35**, 283–320

Carroll, R. J., Ruppert, D. (1988): Transformation and weighting in regression. Chapman & Hall, London

Cattell, R. B. (1966): The scree test for the number of factors. Multivariate Behavioral Research, **1**, 245–276

Christensen, R. (1997): Log-linear models and logistic regression. Springer, New York, 2nd edition

Clark, L. A., Pregibon, D. (1992): Tree-based models. In Chambers, J. M., Hastie, T. J. (eds.) Statistical models in S, Pacific Grove

Cook, R. D., Weisberg, S. (1982): Residuals and influence in regression. Chapman & Hall, New York

Cox, T. F., Cox, M. A. A. (1994): Multidimensional scaling. Chapman & Hall, London

Davison, M. L. (1983): Multidimensional scaling. Wiley, New York

Deutsches PISA-Konsortium (Hrsg.) (2001): PISA 2000. Leske + Budrich, Opladen

Draper, N. R., Smith, H. (1998): Applied regression analysis. Wiley, New York, 3rd edition

Everitt, B. (2001): Cluster analysis. Arnold, London, 4th edition

Fahrmeir, L., Hamerle, A., Tutz, G. (1996): Multivariate statistische Verfahren. de Gruyter, Berlin, 2nd edition

Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G. (2001): Statistik : der Weg zur Datenanalyse. Springer, Berlin, 3rd edition

Friedman, J. H., Tukey, J. W. (1974): A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers, **23**, 881–890

Goodman, L. A. (1971): The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics, **13**, 33–61

Goodman, L. A., Kruskal, W. H. (1954): Measures of association for cross-classification. Journal of the American Statistical Association, **49**, 732–764

Gordon, A. D. (1999): Classification. Chapman & Hall, Boca Raton, 2nd edition

Gower, J. C. (1971): A general coefficient of similarity and some of its properties. Biometrics, **27**, 857–872

Gower, J. C. (1975): Generalized procrustes analysis. Psychometrika, **40**, 33–51

Gower, J. C., Legendre, P. (1986): Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, **3**, 5–48

Guttman, L. (1954): Some necessary conditions for common factor analysis. Psychometrika, **19**, 149–161

Hand, D. J. (1997): Construction and assessment of classification rules. Wiley, Chichester

Härdle, W. (1990a): Applied nonparametric regression. Cambridge Univ. Press, Cambridge

Härdle, W. (1990b): Smoothing techniques: with implementation in S. Springer, Berlin

Hastie, T. J., Tibshirani, R. J. (1991): Generalized additive models. Chapman & Hall, London

Hastie, T. J., Tibshirani, R. J., Friedman, J. H. (2001): The elements of statistical learning : data mining, inference, and prediction. Springer, New York

Heiler, S., Michels, P. (1994): Deskriptive und explorative Datenanalyse. Oldenbourg, München

Hosmer, D. W., Lemeshow, S. (1989): Applied logistic regression. Wiley, New York

Huber, P. J. (1985): Projection pursuit (with discussion). Ann. Statist., **13**, 435–535

Hubert, L. (1974): Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. Journal of the American Statistical Association, **69**, 698–704

Huberty, C. J. (1994): Applied discriminant analysis. Wiley, New York

Hyndman, R. J., Fan, Y. (1996): Sample quantiles in statistical packages. The American Statistician, **50**, 361–365

Jaccard, P. (1908): Nouvelles recherches sur la distribution florale. Bulletin de la Societe Vaudoise de Sciences Naturelles, **44**, 223–370

Jackson, J. E. (1991): A user's guide to principal components. Wiley, New York

Jänich, K. (2000): Lineare Algebra. Springer, Berlin, 8th edition

Jobson, J. D. (1992): Applied multivariate data analysis. Volume II. Categorical and multivariate methods. Springer, New York

Johnson, R. A., Wichern, D. W. (1998): Applied multivariate statistical analysis. Prentice Hall, New Jersey, 4th edition

Jolliffe, I. T. (1972): Discarding variables in principal component analysis, I: Artificial data. Applied Statistics, **21**, 160–173

Jolliffe, I. T. (1986): Principal component analysis. Springer, New York

Jones, M. C., Sibson, R. (1987): What is projection pursuit? Journal of the Royal Statistical Society, Series A, **150**, 1–36

Kaiser, H. F. (1958): The varimax criterion for analytic rotation in factor analysis. Psychometrika, **23**, 187–200

Kaiser, H. F. (1960): The application of electronic computers to factor analysis. Educ. Psychol. Meas., **20**, 141–151

Kaufman, L., Rousseeuw, P. J. (1990): Finding groups in data. Wiley, New York

Kearsley, A. J., Tapia, R. A., Trosset, M. W. (1998): The solution of the metric STRESS and SSTRESS problems in multidimensional scaling using Newton's method. Computational Statistics, **13**, 369–396

Khuri, A. I. (1993): Advanced calculus with applications in statistics. Wiley, New York

Kleinbaum, D. G. (1994): Logistic regression: a self-learning text. Springer, New York

Krause, A., Olson, M. (2000): The basics of S and S-PLUS. Springer, New York

Kruskal, J. B. (1956): On the shortest spanning subtree of a graph and the travelling salesman problem. Poceedings AMS, **7**, 48–50

Kruskal, J. B. (1964): Nonmetric multidimensional scaling: a numerical method. Psychometrika, **29**, 115–129

Krzanowski, W. J. (2000): Principles of multivariate analysis: a user's perspective. Oxford University Press, Oxford, rev. edition

Lachenbruch, P. A., Mickey, M. R. (1968): Estimation of error rates in discriminant analysis. Technometrics, **10**, 1–11

Lasch, R., Edel, R. (1994): Einsatz multivariater Verfahren zur Analyse von Geschäftsstellen eines Kreditinstituts. Diskussionsarbeit am Institut für Statistik und mathematische Wirtschaftstheorie der Universität Augsburg

Loh, W. Y., Shih, Y. S. (1997): Split selection methods for classification trees. Statistica Sinica, **7**, 815–840

Mardia, K. V. (1978): Some properties of classical multidimensionale scaling. Communications in Statistics - Theory and Methods A, **7**, 1233–1241

Mardia, K. V., Kent, J. T., Bibby, J. M. (1979): Multivariate analysis. Academic Press, London

McLachlan, G. J. (1992): Discriminant analysis and statistical pattern recognition. Wiley, New York

Miller, R. G. (1981): Simultaneous statistical inference. Springer, New York, 2nd edition

Milligan, G. W., Cooper, M. C. (1985): An examination of procedures for determining the number of clusters in a data set. Psychometrika, **50**, 159–179

Mojena, R. (1977): Hierarchical grouping methods and stopping rules: an evaluation. Computer Journal, **20**, 359–363

Mood, A. M., Graybill, F. A., Boes, D. C. (1974): Introduction to the theory of statistics. McGraw-Hill, New York, 3rd edition

Neuhaus, J., Wrigley, C. (1954): The quartimax method: an analytical approach to orthogonal simple structure. British Journal of Mathematical and Statistical Psychology, **7**, 81–91

Rice, J. A. (1988): Mathematical statistics and data analysis. Wadsworth, Pacific Grove

Ripley, B. D. (1996): Pattern recognition and neural networks. Cambridge Univ. Press, Cambridge

Rogers, D. J., Tanimoto, T. T. (1960): A computer program for classifying plants. Science, **132**, 1115–1118

Rousseeuw, P. J. (1984): Least median of squares regression. Journal of the American Statistical Association, **79**, 871–880

Rousseeuw, P. J. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, **20**, 53–65

Rousseeuw, P. J., Ruts, I., Tukey, J. W. (1999): The bagplot: a bivariate boxplot. The American Statistician, **53**, 382–387

Rousseeuw, P. J., van Driessen, K. (1999): A fast algorithm for the minimum covariance determinant estimator. Technometrics, **41**, 212–223

Schafer, J. L. (1997): Analysis of incomplete multivariate data. Chapman & Hall, London

Schlittgen, R. (1996): Statistische Inferenz. Oldenbourg, München

Schlittgen, R. (2000): Einführung in die Statistik. Oldenbourg, München, 9th edition

Seber, G. A. F. (1977): Linear regression analysis. Wiley, New York

Seber, G. A. F. (1984): Multivariate observations. Wiley, New York

Small, C. G. (1990): A survey of multidimensional medians. International Statistical Review, **58**, 263–277

Smith, M. (1993): Neural networks for statistical modelling. Van Nostrand Reinhold, New York

Sneath, P. H., Sokal, R. R. (1973): Principles of numerical taxinomy. Freeman, San Francisco

Strang, G. (1988): Linear algebra and its applications. Harcourt Brace Jovanovich, San Diego, 3rd edition

Süselbeck, B. (1993): S und S-PLUS : Eine Einführung in Programmierung und Anwendung. Fischer, Stuttgart

Trippel, A. (2001): Ein Vergleich numerischer Methoden zur Lösung von nichtmetrischen MDS-Problemen. Diplomarbeit, Universität Bielefeld

Tukey, J. W. (1977): Exploratory data analysis. Addison-Wesley, Reading, Mass.

Venables, W. N., Ripley, B. D. (1999): Modern applied statistics with S-PLUS. Springer, Berlin, 3rd edition

Watkins, D. S. (1991): Fundamentals of matrix computations. Wiley, New York

Wetzel, W., Skarabis, H., Naeve, P., Büning, H. (1981): Mathematische Propädeutik für Wirtschaftswissenschaftler. de Gruyter, Berlin, 4th edition

Zurmühl, R., Falk, S. (1997): Matrizen 1: Grundlagen. Springer, Berlin, 7th edition

Index