

Reihenherausgeber:

Prof. Dr. Holger Dette · Prof. Dr. Wolfgang Härdle

Statistik und ihre Anwendungen

Weitere Bände dieser Reihe finden Sie unter <http://www.springer.com/series/5100>

Andreas Handl

Multivariate Analysemethoden

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

2. Auflage

 Springer

Dr. Andreas Handl

in der jetzigen Form bearbeitet von

Dr. Stefan Niermann

Schneckenburgerstr. 15a

30177 Hannover

Deutschland

sniermann@ewas.de

ISBN 978-3-642-14986-3

e-ISBN 978-3-642-14987-0

DOI 10.1007/978-3-642-14987-0

Springer Heidelberg Dordrecht London New York

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

© Springer-Verlag Berlin Heidelberg 2002, 2010

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland vom 9. September 1965 in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtsgesetzes.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Einbandentwurf: WMXDesign GmbH, Heidelberg

Gedruckt auf säurefreiem Papier

Springer ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Für Claudia und Fabian

Vorwort

Vorwort zur zweiten Auflage

Bei der hiermit vorgelegten korrigierten und überarbeiteten zweiten Auflage des Buches wurden Ergänzungen und Verbesserungen des zwischenzeitlich verstorbenen Autors integriert. Seine Lehrveranstaltungen und Arbeiten zeichneten und zeichnen sich durch besonders große Verständlichkeit, Anwendungsbezug, Liebe zum Detail und das ständige Bemühen aus, die Inhalte aus Sicht seiner Hörer und Leser zu sehen. Deshalb ist es mir eine große Freude, Verantwortung und Ehre zugleich, dieses äußerst bewahrenwerte Erbe fortzuführen.

In der vorliegenden zweiten Auflage wurde im vierten Kapitel die Berechnung des Gower-Koeffizienten überarbeitet. Im fünften Kapitel wurde an Stelle der exemplarische Hauptkomponentenanalyse der PISA-Daten eine für die praktische Anwendung geeignete, auf 7 Schritten beruhende Vorgehensweise dargestellt. Das Kapitel 13 zur Clusteranalyse wurde ergänzt um das Ward-, das Median- und das Zentroid-Verfahren. Diese werden als Spezialfälle der Rekursionsbeziehung von Lance und Williams eingeführt. Darüber hinaus wurden weitere Eigenschaften hierarchischer Verfahren in das Kapitel aufgenommen.

Hannover, im April 2010

Stefan Niermann

Vorwort zur ersten Auflage

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 wie 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 Zufallsvariablen. 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 denk-

bare 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

Inhaltsverzeichnis

Teil I Grundlagen

1	Beispiele multivariater Datensätze	3
2	Elementare Behandlung der Daten	13
2.1	Beschreibung und Darstellung univariater Datensätze	13
2.1.1	Beschreibung und Darstellung qualitativer Merkmale	15
2.1.2	Beschreibung und Darstellung quantitativer Merkmale	17
2.2	Beschreibung und Darstellung multivariater Datensätze	22
2.2.1	Beschreibung und Darstellung von Datenmatrizen quantitativer Merkmale	22
2.2.2	Beschreibung und Darstellung von Datenmatrizen qualitativer Merkmale	36
2.3	Datenbehandlung in S-PLUS	41
2.3.1	Univariate Datenanalyse	41
2.3.2	Multivariate Datenanalyse	51
2.4	Ergänzungen und weiterführende Literatur	61
2.5	Übungen	61
3	Mehrdimensionale Zufallsvariablen	65
3.1	Problemstellung	65
3.2	Univariate Zufallsvariablen	65
3.3	Zufallsmatrizen und Zufallsvektoren	70
3.4	Die multivariate Normalverteilung	81
4	Ähnlichkeits- und Distanzmaße	83
4.1	Problemstellung	83
4.2	Bestimmung der Distanzen und Ähnlichkeiten aus der Datenmatrix	84
4.2.1	Quantitative Merkmale	84
4.2.2	Binäre Merkmale	94
4.2.3	Qualitative Merkmale mit mehr als zwei Merkmalsausprägungen	98
4.2.4	Qualitative Merkmale, deren Merkmalsausprägungen geordnet sind	98

4.2.5 Unterschiedliche Messniveaus 98
 4.3 Distanzmaße in S-PLUS 102
 4.4 Direkte Bestimmung der Distanzen 108
 4.5 Übungen 110

Teil II Darstellung hochdimensionaler Daten in niedrigdimensionalen Räumen

5 Hauptkomponentenanalyse 115
 5.1 Problemstellung 115
 5.2 Hauptkomponentenanalyse bei bekannter Varianz-Kovarianz-Matrix 120
 5.3 Hauptkomponentenanalyse bei unbekannter Varianz-Kovarianz-Matrix 123
 5.4 Praktische Aspekte 126
 5.4.1 Anzahl der Hauptkomponenten 128
 5.4.2 Überprüfung der Güte der Anpassung 130
 5.4.3 Analyse auf Basis der Varianz-Kovarianz-Matrix oder auf Basis der Korrelationsmatrix 133
 5.5 Wie geht man bei einer Hauptkomponentenanalyse vor? 135
 5.6 Hauptkomponentenanalyse in S-PLUS 140
 5.7 Ergänzungen und weiterführende Literatur 144
 5.8 Übungen 145

6 Mehrdimensionale Skalierung 149
 6.1 Problemstellung 149
 6.2 Metrische mehrdimensionale Skalierung 150
 6.2.1 Theorie 150
 6.2.2 Praktische Aspekte 165
 6.2.3 Metrische mehrdimensionale Skalierung der Rangreihung der Politikerpaare 167
 6.2.4 Metrische mehrdimensionale Skalierung in S-PLUS ... 169
 6.3 Nichtmetrische mehrdimensionale Skalierung 171
 6.3.1 Theorie 171
 6.3.2 Nichtmetrische mehrdimensionale Skalierung in S-PLUS 179
 6.4 Ergänzungen und weiterführende Literatur 182
 6.5 Übungen 182

7 Procrustes-Analyse 185
 7.1 Problemstellung und Grundlagen 185
 7.2 Illustration der Vorgehensweise 187
 7.3 Theorie 192
 7.4 Procrustes-Analyse der Reisezeiten 194
 7.5 Procrustes-Analyse in S-PLUS 196

7.6	Ergänzungen und weiterführende Literatur	198
7.7	Übungen	198

Teil III Abhängigkeitsstrukturen

8	Lineare Regression	203
8.1	Problemstellung und Modell	203
8.2	Schätzung der Parameter	206
8.3	Praktische Aspekte	211
8.3.1	Interpretation der Parameter bei mehreren erklärenden Variablen	211
8.3.2	Die Güte der Anpassung	215
8.3.3	Tests	219
8.4	Lineare Regression in S-PLUS	222
8.5	Ergänzungen und weiterführende Literatur	224
8.6	Übungen	224
9	Explorative Faktorenanalyse	227
9.1	Problemstellung und Grundlagen	227
9.2	Theorie	235
9.2.1	Das allgemeine Modell	235
9.2.2	Nichteindeutigkeit der Lösung	238
9.2.3	Schätzung von \mathbf{L} und Ψ	240
9.3	Praktische Aspekte	246
9.3.1	Bestimmung der Anzahl der Faktoren	246
9.3.2	Rotation	247
9.4	Faktorenanalyse in S-PLUS	249
9.5	Ergänzungen und weiterführende Literatur	251
9.6	Übungen	252
10	Hierarchische loglineare Modelle	255
10.1	Problemstellung und Grundlagen	255
10.2	Zweidimensionale Kontingenztabelle	265
10.2.1	Modell 0	265
10.2.2	Modell A	267
10.2.3	Der IPF-Algorithmus	268
10.2.4	Modell B	270
10.2.5	Modell A, B	272
10.2.6	Modell AB	274
10.2.7	Modellselektion	274
10.3	Dreidimensionale Kontingenztabelle	277
10.3.1	Das Modell der totalen Unabhängigkeit	277
10.3.2	Das Modell der Unabhängigkeit einer Variablen	281
10.3.3	Das Modell der bedingten Unabhängigkeit	285

10.3.4 Das Modell ohne Drei-Faktor-Interaktion 288
 10.3.5 Das saturierte Modell 290
 10.3.6 Modellselektion 291
 10.4 Loglineare Modelle in S-PLUS 292
 10.5 Ergänzungen und weiterführende Literatur 298
 10.6 Übungen 298

Teil IV Gruppenstruktur

11 Einfaktorielle Varianzanalyse 303
 11.1 Problemstellung 303
 11.2 Univariate einfaktorielle Varianzanalyse 303
 11.2.1 Theorie 303
 11.2.2 Praktische Aspekte 311
 11.3 Multivariate einfaktorielle Varianzanalyse 317
 11.4 Einfaktorielle Varianzanalyse in S-PLUS 319
 11.5 Ergänzungen und weiterführende Literatur 322
 11.6 Übungen 322

12 Diskriminanzanalyse 325
 12.1 Problemstellung und theoretische Grundlagen 325
 12.2 Diskriminanzanalyse bei normalverteilten Grundgesamtheiten 334
 12.2.1 Diskriminanzanalyse bei Normalverteilung mit
 bekannten Parametern 334
 12.2.2 Diskriminanzanalyse bei Normalverteilung mit
 unbekannten Parametern 340
 12.3 Fishers lineare Diskriminanzanalyse 343
 12.4 Logistische Diskriminanzanalyse 348
 12.5 Klassifikationsbäume 351
 12.6 Praktische Aspekte 358
 12.7 Diskriminanzanalyse in S-PLUS 362
 12.8 Ergänzungen und weiterführende Literatur 369
 12.9 Übungen 369

13 Clusteranalyse 373
 13.1 Problemstellung 373
 13.2 Hierarchische Clusteranalyse 374
 13.2.1 Theorie 374
 13.2.2 Verfahren der hierarchischen Clusterbildung 381
 13.2.3 Praktische Aspekte 407
 13.2.4 Hierarchische Clusteranalyse in S-PLUS 411
 13.3 Partitionierende Verfahren 414
 13.3.1 Theorie 414
 13.3.2 Praktische Aspekte 417

13.3.3 Partitionierende Verfahren in S-PLUS 422
 13.4 Clusteranalyse der Daten der Regionen 427
 13.5 Ergänzungen und weiterführende Literatur 429
 13.6 Übungen 429

Teil V Anhänge

A Mathematische Grundlagen 435
 A.1 Matrizenrechnung 435
 A.1.1 Definitionen und spezielle Matrizen 436
 A.1.2 Matrixverknüpfungen 437
 A.1.3 Die inverse Matrix 441
 A.1.4 Orthogonale Matrizen 442
 A.1.5 Spur einer Matrix 443
 A.1.6 Determinante einer Matrix 444
 A.1.7 Lineare Gleichungssysteme 445
 A.1.8 Eigenwerte und Eigenvektoren 447
 A.1.9 Die Spektralzerlegung einer symmetrischen Matrix 449
 A.1.10 Die Singulärwertzerlegung 451
 A.1.11 Quadratische Formen 452
 A.2 Extremwerte 453
 A.2.1 Der Gradient und die Hesse-Matrix 454
 A.2.2 Extremwerte ohne Nebenbedingungen 456
 A.2.3 Extremwerte unter Nebenbedingungen 457
 A.3 Matrizenrechnung in S-PLUS 459

B S-PLUS-Funktionen 465
 B.1 Quartile 465
 B.2 Distanzmatrix 465
 B.3 Monotone Regression 466
 B.4 STRESS1 467
 B.5 Bestimmung einer neuen Konfiguration 467
 B.6 Kophenetische Matrix 468
 B.7 Gamma-Koeffizient 469
 B.8 Bestimmung der Zugehörigkeit zu Klassen 469
 B.9 Silhouette 470
 B.10 Zeichnen einer Silhouette 471

C Tabellen 473
 C.1 Standardnormalverteilung 473
 C.2 χ^2 -Verteilung 475
 C.3 t -Verteilung 476
 C.4 F -Verteilung 477

XVI Inhaltsverzeichnis

Literaturverzeichnis 479

Index 485