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.

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 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 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