



Clusteranalyse

Anwendungsorientierte Einführung
in Klassifikationsverfahren

von

Universitätsprofessor

Dr. Johann Bacher

Johannes-Kepler-Universität, Linz

Akademischer Rat

Dr. Andreas Pöge

Universität Bielefeld

Diplom-Sozialwirt

Knut Wenzig

Nationales Bildungspanel, Bamberg

3., ergänzte, vollständig überarbeitete und
neu gestaltete Auflage

Oldenbourg Verlag München

www.clusteranalyse.net
3.auflage@clusteranalyse.net

Bibliografische Information der Deutschen Nationalbibliothek

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

© 2010 Oldenbourg Wissenschaftsverlag GmbH
Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
oldenbourg.de

Das Werk einschließlich aller Abbildungen ist urheberrechtlich geschützt. Jede Verwertung außerhalb der Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Bearbeitung in elektronischen Systemen.

Lektorat: Rainer Berger
Herstellung: Anna Grosser
Coverentwurf: Kochan & Partner, München
Umschlagillustration: Cornelia Horn
Gedruckt auf säure- und chlorfreiem Papier
Gesamtherstellung: Grafik + Druck GmbH, München

ISBN 978-3-486-58457-8

Vorwort zur dritten Auflage

Seit der Veröffentlichung der zweiten Auflage sind fast 15 Jahre vergangen, und im Bereich der Clusteranalyse hat es viele Weiterentwicklungen gegeben – die meisten sicherlich im Bereich der probabilistischen Verfahren. Im neuen Autorenteam wurden daher – neben der Umstellung auf L^AT_EX – die beiden vorausgehenden Auflagen grundlegend überarbeitet, um den aktuellen Entwicklungen Rechnung zu tragen. Beispielhaft genannt seien: Eine Erweiterung der Einleitung um eine Sammlung von Beispielen aus der Forschungspraxis, eine taxative Nennung und Beschreibung von Kriterien für eine gute Klassifikation sowie in Teil I eine klare begriffliche Abgrenzung von Hauptkomponentenmethode und Faktorenanalyse. Ausführlich behandelt werden Datenkonstellationen, für welche die Faktorenanalyse brauchbar oder unbrauchbar ist. Daneben wurde in Teil II das K-Means-Verfahren überarbeitet: Aufgenommen wurde die Methode der multiplen zufälligen Startwerte, die eine entscheidende Weiterentwicklung darstellt und das Problem lokaler Minima weitgehend vermeidet. Dargestellt werden auch Verallgemeinerungen des K-Means-Verfahrens, die andere Distanzfunktionen und Lageparameter nutzen.

Teil III wurde um den Latent-GOLD-Ansatz, der die Modellierung von komplexen Clustermodellen ermöglicht, erweitert. Ergänzt wurde dieser Teil um eine Beschreibung von SPSS-TwoStep und AutoClass. Mit AutoClass wird in die derzeit »boomende« Bayes-Statistik eingeführt. Darüber hinaus wird die Performanz von ausgewählter Software verglichen. Im praktisch orientierten Teil IV werden häufig gestellte Anwenderfragen beantwortet und die Klassifikation von Verläufen mittels Optimal Matching, die Bildung von Konsensclustern sowie die formale Gültigkeitsprüfung dargestellt.

In allen Teilen wurden Anwendungsempfehlungen aufgenommen, wobei versucht wurde, einerseits den Praxisanforderungen Rechnung zu tragen und andererseits einfache »Kochbuchrezepte« zu vermeiden. Wie in den bisherigen Auflagen ist das Ziel, eine Darstellung dergestalt zu finden, dass die behandelten Verfahren in ihren Grundlagen von der interessierten Leserin oder dem interessierten Leser nachvollzogen werden können. Für eine wissenschaftliche Analyse erscheint es uns unumgänglich, dass zumindest eine grobe Vorstellung über die Funktionsweise eines Verfahrens vorliegt, bevor ein »Rechen-Knopf« am Computer gedrückt wird.

Über Rückmeldungen und Anregungen zu einzelnen Verfahren und Ausführungen freuen wir uns. Unser Dank gilt Arne Bethmann und Heinz Leitgöb für die Mitarbeit an einzelnen Teilen des Buches. Danken möchten wir darüber hinaus unseren Kolleginnen und Kollegen für Hinweise und Korrekturen sowie dem Oldenbourg-Verlag für seine Geduld bei der Abgabe der Druckvorlage. Die Fertigstellung des Manuskripts hat erhebliche Zeit zu Lasten von familiären Verpflichtungen gebunden. Für das Verständnis hierfür herzlichen Dank an alle.

Linz, Bielefeld, Bamberg 2010

Johann Bacher

Andreas Pöge

Knut Wenzig

Vorwort zur ersten und zweiten Auflage

Clusteranalyseverfahren gewinnen in der Forschungspraxis zunehmend an Bedeutung. Sie werden heute in zahlreichen Wissenschaftsdisziplinen zur Lösung von Klassifikationsaufgaben eingesetzt, in den Sozial- und Wirtschaftswissenschaften beispielsweise zur Identifizierung von unterschiedlichen Lebens- und Konsumstilen oder von Wertorientierungstypen.

Die Arbeit gibt – auf Grundlage der neueren Methodenliteratur – eine systematische Einführung, wie bei der Bestimmung von Typen (Clustern) vorzugehen ist, welche Verfahren dabei zur Verfügung stehen und wie die Ergebnisse zu interpretieren sind. Folgende Verfahren werden behandelt: bivariate und multiple Korrespondenzanalyse, nichtmetrische mehrdimensionale Skalierung, Faktorenanalyseverfahren (nominale Faktorenanalyse nach McDonald, R- und Q-Analyse), hierarchisch-agglomerative Verfahren, Repräsentanten-Verfahren, explorative und konfirmatorische K-Means-Verfahren, explorative und konfirmatorische probabilistische Clusteranalyseverfahren (latente Profilanalyse, Analyse latenter Klassen für nominale, ordinale und gemischte Variablen).

Alle Verfahren werden anhand von Beispielen aus der Forschungspraxis behandelt und durchgerechnet. Dabei werden eine Reihe von methodologischen Fragen beantwortet, wie beispielsweise die Frage, welche Konsequenzen eine empirische Standardisierung hat oder wie sich irrelevante Variablen auf die Ergebnisse einer Clusteranalyse auswirken.

Wie jede Arbeit wäre auch das hier vorliegende Buch nicht ohne die Unterstützung meiner akademischen Lehrer, des Oldenbourg-Verlages, der Teilnehmer an unterschiedlichen Kursen zur Datenanalyse sowie meiner Bekannten und Freunden und meiner Familie zustandegekommen. Ihnen allen möchte ich an dieser Stelle herzlichst danken.

Johann Bacher

Inhalt

1	Einleitung	15
1.1	Zielsetzung clusteranalytischer Verfahren	15
1.2	Homogenität als Grundprinzip der Bildung von Clustern	16
1.3	Clusteranalyseverfahren	18
1.4	Grundlage der Clusterbildung	20
1.5	Konfirmatorische und explorative Clusteranalyse	22
1.6	Anwendungsbeispiele	23
1.7	Modellprüfung und Validierung	27
1.8	Fehleranalyse	28
1.9	Datenanalyse als iterativer Prozess	30
1.10	Computerprogramme	32
I	Unvollständige Clusteranalyseverfahren	35
2	Einleitende Übersicht	37
3	Multiple Korrespondenzanalyse	43
3.1	Ein Anwendungsbeispiel	43
3.1.1	Faktorenanalytische Interpretation	46
3.1.2	Clusteranalytische Interpretation	52
3.2	Das Modell der multiplen Korrespondenzanalyse	57
3.2.1	Berechnung der empirischen Zusammenhangsmatrix G	58
3.2.2	Berechnung der Eigenwerte, Faktorladungen und Koordinatenwerte	61
3.2.3	Berechnung der Skalenwerte und Interpretation der Koordinaten	63
3.2.4	Unerwünschter Effekt der Reskalierung der Faktorladungen und Rotation der Faktoren	65
3.3	Modellprüfgrößen	67
3.3.1	Signifikanz der Zusammenhangsstruktur	67
3.3.2	Die Zahl maximal möglicher und bedeutsamer Dimensionen	68
3.3.3	Überprüfung der faktorenanalytischen Interpretation	69
3.3.4	Modellprüfgrößen für die clusteranalytische Interpretation	72

3.4	Anwendungsempfehlungen	74
4	Nichtmetrische mehrdimensionale Skalierung	77
4.1	Aufgabenstellung und Ähnlichkeitsmessung	77
4.2	Schätzalgorithmus	80
4.3	Maximale und angemessene Dimensionszahl	87
4.4	Unbekannter Metrikparameter p	90
4.5	Weitere Modellanpassungsgrößen	92
4.6	Freizeitverhalten von Kindern	94
4.6.1	Clusteranalytische Interpretation	96
4.6.2	Faktorenanalytische Interpretation	97
4.6.3	Freizeitaktivitäten und Sozialstruktur	100
4.7	Anwendungsempfehlungen	107
5	Weitere räumliche Darstellungsverfahren	109
5.1	Die bivariate Korrespondenzanalyse	109
5.2	Nominale Faktorenanalyse nach McDonald	117
5.3	Die Hauptkomponenten- und Faktorenanalyse	122
5.3.1	Hauptkomponenten- und R-Faktorenanalyse	122
5.3.2	Die Q-Faktorenanalyse	136
5.4	Anwendungsempfehlungen	143
II	Deterministische Clusteranalyseverfahren	145
6	Einleitende Übersicht	147
6.1	Überlappende und überlappungsfreie Clusterlösungen	147
6.2	Grundvorstellungen über die zu bildenden Cluster	148
6.3	Complete- und Single-Linkage als Basismodelle	150
6.4	Auswahl eines geeigneten Verfahrens	153
6.5	Lösungsschritte einer Klassifikationsaufgabe	156
6.6	Ein Anwendungsbeispiel	156
6.7	Fehlerquellen	165
7	Gewichtung und Transformation von Variablen	175
7.1	Vergleichbarkeit von Klassifikationsmerkmalen	175
7.2	Lösungsstrategien	176
7.3	Theoretische und empirische Standardisierung	177
7.4	Hierarchische Variablen	183
7.5	Gemischte Variablen	184

7.6	Standardisierung von Objekten	188
7.7	Exkurs: Die Problematik einer automatischen Orthogonalisierung	193
8	Unähnlichkeits- und Ähnlichkeitsmaße	195
8.1	Auswahl eines (Un-)Ähnlichkeitsmaßes	196
8.2	Dichotome Variablen	197
8.3	Nomiale Variablen	207
8.4	Ordinale Variablen	211
8.5	Quantitative Variablen	219
8.6	A-priori-Prüfung auf Vorhandensein einer Clusterstruktur	224
8.7	Gewichtung von Variablen und Distanzen, Standardisierung von Objekten	226
8.8	Fehlende Werte	228
8.9	Exkurs: Quantifizierung und Konsequenzen der Kategorisierung	230
9	Nächste-Nachbarn- und Mittelwertverfahren	233
9.1	Der Complete-Linkage als Basismodell	233
9.1.1	Der hierarchisch-agglomerative Algorithmus	233
9.1.2	Hierarchische Darstellung von Ähnlichkeitsbeziehungen	237
9.1.3	Maßzahlen zur Bestimmung der Clusterzahl	241
9.1.4	Zufallstestung des Verschmelzungsschemas	245
9.1.5	Maßzahlen zur Beurteilung einer bestimmten Clusterlösung	247
9.2	Der Single-Linkage	251
9.3	Complete-Linkage für überlappende Cluster	255
9.4	Verallgemeinerte Nächste-Nachbarn-Verfahren	259
9.5	Mittelwertverfahren	264
9.6	Anwendungsempfehlungen	274
10	Repräsentanten-Verfahren	277
10.1	Modellansatz	277
10.2	Anwendungsbeispiel	279
10.3	Die Wahl der Schwellenwerte	280
10.4	Weitere Repräsentanten-Verfahren	282
10.5	Anwendungsempfehlungen	283
11	Hierarchische Verfahren zur Konstruktion von Clusterzentren	285
11.1	Modellansätze, Algorithmen und Ward-Verfahren	285
11.2	Bestimmung der Clusterzahl und Modellprüfgrößen	290
11.3	Analyse durchschnittlicher Befragter	290
11.4	Anwendungsempfehlungen	295

12	K-Means-Verfahren	299
12.1	Modellansatz und Algorithmus	299
12.2	Bestimmung der Clusterzahl	305
12.3	Zufallstestung einer bestimmten Clusterlösung	313
12.4	Beschreibung und Interpretation der Cluster	314
12.5	Formale Beschreibung der Cluster	321
12.6	Analyse von Ausreißern	325
12.7	Stabilitätsprüfung	328
12.8	Inhaltliche Validitätsprüfung	332
12.9	Alternative Startwertverfahren	335
12.10	Gemischtes Messniveau	336
12.11	Modifikation des Algorithmus	338
12.12	Verwendung der Mahalanobis-Distanz	339
12.13	Konfirmatorisches K-Means-Verfahren	341
12.14	K-Median- und K-Modus-Verfahren	345
12.15	Anwendungsempfehlungen	347
III	Probabilistische Clusteranalyseverfahren	349
13	Einleitende Übersicht	351
14	Latente Profilanalyse	355
14.1	Modellansatz und Algorithmus	355
14.2	Modellprüfgrößen	362
14.2.1	Bestimmung der Klassenzahl	362
14.2.2	Zufallstestung einer Klassenlösung	366
14.3	Beschreibung und Interpretation einer Klassenlösung	367
14.4	Überlappungsindizes	368
14.5	Überprüfung der Annahme der lokalen Unabhängigkeit	372
14.6	Konfirmatorische latente Profilanalyse	373
15	Analyse latenter Klassen für nominale, ordinale und gemischtskalierte Variablen	377
15.1	Modellansatz und Algorithmus für nominale Variablen	377
15.2	Modellprüfung und Interpretation	382
15.3	Konfirmatorische Analyse	387
15.4	Modellansatz und Algorithmus für ordinale und gemischtskalierte Variablen	390

16 Latent-GOLD-Ansatz	395
16.1 Allgemeiner Ansatz und Überblick	395
16.2 Modellansatz der Latent-Class-Clusteranalyse	398
16.2.1 Der klassische Ansatz	399
16.2.2 Erweiterung mit Kovariaten	404
16.2.3 Ordinale Indikatorvariablen	407
16.2.4 Kontinuierliche Indikatorvariablen	409
16.2.5 Zählvariablen	409
16.3 Parameterschätzung	411
16.4 Statistiken zur Modellanpassung	416
16.4.1 χ^2 -Statistiken	417
16.4.2 Log-Likelihood-Statistiken	420
16.4.3 Klassifikations-Statistiken	420
16.4.4 Signifikanztests mit parametrischem Bootstrap	422
16.4.5 Bivariate Residuen	424
16.4.6 Beurteilung und Auswahl von Modellen	425
16.5 Ein Anwendungsbeispiel	426
16.5.1 Kontinuierliche Daten (latente Profilanalyse)	426
16.5.2 Hinzunahme von Kovariaten	431
17 Weiterentwicklungen und Modifikationen	439
17.1 AutoClass (<i>gemeinsam mit Arne Bethmann</i>)	439
17.1.1 Modell	441
17.1.2 Schätzverfahren und Bestimmung der Clusterzahl	443
17.1.3 Vergleichsrechnung mit den Denz-Daten	444
17.2 TwoStep-Cluster	446
17.2.1 Allgemeiner Ansatz	446
17.2.2 Ergebnis für die metrischen Denz-Daten und für gemischte Skalenniveaus	450
17.3 Vergleich ausgewählter Software (<i>gemeinsam mit Arne Bethmann</i>)	451
17.3.1 Simulationsmodelle	451
17.3.2 Ergebnisse	452
IV Spezielle Anwendungsfragen	457
18 Häufig gestellte Anwendungsfragen	459
18.1 Welches Verfahren?	459
18.1.1 Bildung abgeleiteter Variablen	459
18.1.2 Räumliche Darstellung von Objekten oder Variablen	460
18.1.3 Auffinden einer hierarchischen Ähnlichkeitsstruktur	461

18.1.4	Räumliche oder hierarchische Darstellung?	462
18.1.5	Klassifikation von Variablen	463
18.1.6	Klassifikation von Objekten	464
18.2	Verwendung aller Variablen?	466
18.3	Welches Un- bzw. Ähnlichkeitsmaß?	469
18.4	Wie viele Cluster?	470
18.5	Modale Klassenzugehörigkeit oder Zuordnungswahrscheinlichkeiten?	472
19	Klassifikation von Verläufen mittels Optimal Matching	
	<i>Heinz Leitgöb</i>	475
19.1	Einführung	475
19.2	Methodische Grundlagen	476
19.3	Die Hamming-Distanz	479
19.4	Die Levenshtein-Distanz	481
19.5	Ein theoretisches Beispiel	482
19.6	Die Festsetzung der Transformationskosten	485
19.7	Der Analyseablauf	488
19.8	Fazit und Anwendungsempfehlungen	491
20	Formale Gültigkeitsprüfung und Konsensuslösungen	493
20.1	Formale Gültigkeitsprüfung	493
20.2	Konsensuslösungen	497
20.2.1	Konsensus für Clustermittelwerte	498
20.2.2	Konsensus auf der Basis der Clusterzuordnungen	499
Literatur		503
Register		523

1 Einleitung

1.1 Zielsetzung clusteranalytischer Verfahren

Primäres Ziel clusteranalytischer Auswertungsverfahren ist, eine Menge von Klassifikationsobjekten in homogene Gruppen (Klassen, Cluster, Typen) zusammenzufassen – oder kurz ausgedrückt – das *Auffinden einer empirischen Klassifikation* (Gruppeneinteilung, Typologie). Von einer empirischen Klassifikation soll dann gesprochen werden, wenn die Klassifikation auf empirischen Beobachtungen, zum Beispiel auf der Grundlage einer Befragung, basiert. Klassifikationsobjekte (siehe Übersichtstabelle 1.1) können sein: Individuen (Personen, Befragte), Aggregate (Organisationen, Nationen, Berufsgruppen usw.) oder Variablen (Merkmale).

Tab. 1.1: Beispiele für clusteranalytische Auswertungen

Klassifikationsobjekte	Beispiele für eine clusteranalytische Auswertung
Personen	A) Befragte werden aufgrund ihrer sozialstrukturellen Merkmale in (homogene) soziale Schichten zusammengefasst. B) Befragte werden aufgrund ihrer Lebensstile (Freizeitpräferenzen, Musikgeschmack, Wertorientierungen) in homogene Lebensstilgruppen zusammengefasst.
Aggregate	C) Nationen werden aufgrund ihrer demographischen, wirtschaftlichen und/oder sozialen Entwicklung in homogene Gruppen zusammengefasst. D) Berufe werden aufgrund ihrer Tätigkeitsprofile in homogene Gruppen zusammengefasst.
Variablen	E) Freizeitaktivitäten (Variablen) werden aufgrund ihres gemeinsamen Auftretens bei Personen in homogene Gruppen von Variablen zusammengefasst. F) Indikatoren der demographischen, wirtschaftlichen und sozialen Entwicklung werden aufgrund ihrer Korrelationen bei Nationen in homogene (Variablen-)Gruppen zusammengefasst.

Formal unterscheiden sich die sechs Beispiele der Übersicht darin, dass in den Beispielen A bis D die Zeilen einer Datenmatrix die Klassifikationsobjekte bilden. In den Beispielen E bis F sind dies dagegen die Spalten einer Datenmatrix. Im ersten Fall soll von einer *objektorientierten* Datenanalyse gesprochen werden, im letzten Fall von einer *variablenorientierten* Datenanalyse. Die Bezeichnung Klassifikationsobjekte kann sich auf die Spalten oder Zeilen einer Datenmatrix beziehen und bezeichnet jene Einheiten (Personen, Aggregate oder Variablen), die geclustert werden. Die untersuchten Objekte (Zeilen) müssen dabei keinesfalls mit den Erhebungseinheiten identisch sein. So zum Beispiel können durch Aggregierung über eine oder mehrere Variablen (zum Beispiel Beruf, regionale oder nationale Zugehörigkeit usw.) »neue« Objekte (Aggregate) erzeugt werden, die dann geclustert werden.

1.2 Homogenität als Grundprinzip der Bildung von Clustern

Jeder Clusterbildung liegt – unabhängig von Unterschieden im Detail – die Grundvorstellung der *Homogenität* bzw. von »homogenen« Gruppen zugrunde (Kozelka 1982, S. 6; Sodeur 1974, S. 118–124; u. a.). Mit dem Begriff *homogene Gruppe* sind folgende Vorstellungen verbunden:

1. Die Klassifikationsobjekte, die einer homogenen Gruppe angehören, sollen untereinander ähnlich sein. Es soll *Homogenität innerhalb der Cluster* vorliegen.
2. Die Klassifikationsobjekte, die unterschiedlichen homogenen Gruppen angehören, sollen verschieden sein. Es soll *Heterogenität zwischen den Clustern* vorliegen.

Diese beiden Grundvorstellungen werden in der Literatur unterschiedlich bezeichnet. Cormack (1971) spricht von *externer Isolierung* (Heterogenität zwischen den Clustern) und *interner Kohäsion* (Homogenität innerhalb der Cluster). Everitt (1980, S. 60) und andere sprechen von »natürlichen« Clustern, wenn beide Forderungen erfüllt sind, und beschreiben Cluster als dichte Punktwolken in einem p-dimensionalen Raum, die durch Regionen mit einer geringen Dichte voneinander getrennt sind.¹ Sind die beiden Grundvorstellungen nicht erfüllt, ist es wenig sinnvoll, eine Klassifikation durchzuführen. Abbildung 1.1a verdeutlicht diesen Sachverhalt. Die untersuchten Klassifikationsobjekte bilden in den beiden Variablen X und Y eine große, relativ geschlossene Punktwolke. Eine Aussage der Art, den Daten liegen K Cluster (zum Beispiel K = 3) zugrunde, ist nicht sinnvoll. In der Abbildung 1.1b dagegen sind drei Cluster zu erkennen. Sie sind

¹ Zur Vorstellung von »natürlichen« Clustern siehe auch Aldenderfer und Blashfield (1984, S. 33–34) oder Vogel (1975, S. 16–17).

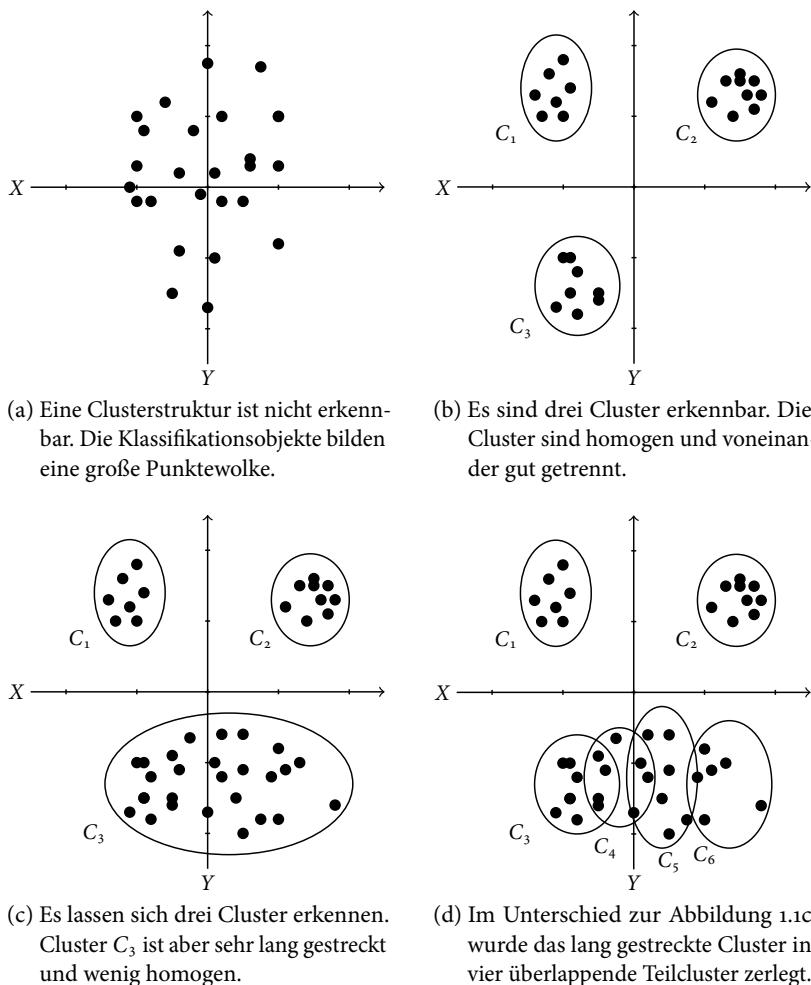


Abb. 1.1: Datenkonstellationen mit erkennbarer und nicht erkennbarer Clusterstruktur

in sich homogen und voneinander verschieden. In der Abbildung 1.1c ist zwar die Vorstellung der Heterogenität zwischen den Clustern erfüllt, Cluster 3 erfüllt aber in einem geringeren Ausmaß die Homogenitätsbedingung innerhalb der Cluster, da das Cluster sehr lang gestreckt ist. Abhängig von dem Gewicht beider Grundvorstellungen bei der Clusterbildung wird man sich entweder für drei Cluster entscheiden oder aus dem lang gestreckten Cluster mehrere überlappende Cluster bilden (siehe Abbildung 1.1d).

Neben diesen beiden Grundvorstellungen werden aus forschungspraktischen, aber auch aus inhaltlichen Gründen weitere Anforderungen an die gesuchte Klassifikation gestellt

(Schlosser 1976, S. 186; Sodeur 1974, S. 125–129), so zum Beispiel, dass die Zahl der Cluster möglichst klein sein sollte (Schlosser 1976, S. 186). Zusammenfassend lassen sich folgende Anforderungen bzw. Kriterien auflisten, die häufig von einer guten Clusterlösung verlangt werden:

1. Die Cluster sollen in sich homogen sein. Objekte, die einem Cluster angehören, sollen zueinander ähnlich sein.
2. Die Cluster sollen voneinander isoliert sein. Objekte, die verschiedenen Clustern angehören, sollen sich voneinander unterscheiden.
3. Die Cluster sollen den Daten gut angepasst sein. Die Klassifikation soll in der Lage sein, die Variation in den Daten zu erklären.
4. Die Cluster sollen stabil sein. Geringfügige Änderungen in den Daten oder im Verfahren sollen in keinen gravierenden Änderungen der Ergebnisse resultieren.
5. Die Cluster sollen inhaltlich gut interpretierbar sein. Den Clustern sollen inhaltlich sinnvolle Namen gegeben werden können. Im Idealfall sollen die Namen aus einer Theorie abgeleitet werden.
6. Die Cluster sollen (inhaltlich) valide sein. Die Cluster sollen mit externen Variablen korrelieren, von denen bekannt ist, dass sie im Zusammenhang mit den Typen stehen, die aber nicht in die Bildung der Cluster eingehen.

Gefordert wird mitunter ferner:

7. Die Zahl der Cluster soll klein und damit überschaubar sein. Angenommen wird, dass dies die inhaltliche Interpretierbarkeit (Kriterium 5) erleichtert und die Stabilität erhöht (Kriterium 4).
8. Die Cluster selbst sollen eine gewisse Mindestgröße haben. Dies soll zur Stabilität (Kriterium 4) beitragen.

1.3 Clusteranalyseverfahren

Zum Auffinden von Clustern wurde eine Vielzahl von Verfahren entwickelt, für die unterschiedliche Einteilungen und Zusammenfassungen vorgeschlagen wurden. In der vorliegenden Arbeit werden drei große Verfahrensgruppen unterschieden: *unvollständige Clusteranalyseverfahren*, *deterministische Clusteranalyseverfahren* und *probabilistische Clusteranalyseverfahren*. Grundlage der Differenzierung ist die Zuordnung der Klassifikationsobjekte zu den Clustern:

- *Unvollständige Clusteranalyseverfahren* (siehe Teil I): Diese Verfahren werden in der Literatur auch als geometrische Methoden (Gordon 1981, S. 80–120), als Repräsentations-

oder Projektionsverfahren (Jain und Dubes 1988; Opitz 1980) bezeichnet. In dieser Arbeit wurde die Bezeichnung »unvollständige Clusteranalyseverfahren« gewählt, da die Bildung von Clustern und die Zuordnung der Klassifikationsobjekte zu den Clustern von der Anwenderin bei der Interpretation der räumlichen Darstellung vorgenommen werden muss. Die unvollständigen Clusteranalyseverfahren selbst führen nur zu einer räumlichen Darstellung. Die Verfahren können auch dazu verwendet werden, abgeleitete Variablen (zum Beispiel Faktorwerte) zu bilden oder eine metrische Ähnlichkeits- oder Unähnlichkeitsmatrix (zum Beispiel mittels mehrdimensionaler Skalierung) zu schätzen, die anschließend mittels eines hierarchischen Clusteranalyseverfahrens untersucht wird.

- *Deterministische Clusteranalyseverfahren* (siehe Teil II): Die Klassifikationsobjekte werden mit einer Wahrscheinlichkeit von 1 einem oder mehreren Clustern zugeordnet. Es lassen sich zwei Verfahrensgruppen unterscheiden: *hierarchische Verfahren* und *partitionierende Verfahren*. Bei den hierarchischen Verfahren erfolgt die Clusterbildung schrittweise: Bei den sogenannten *hierarchisch-agglomerativen Verfahren* werden aus n Objekten zunächst n Cluster gebildet, aus den n Clustern durch Zusammenfassung der beiden ähnlichsten Cluster dann $n - 1$ Cluster, aus diesen wiederum $n - 2$ Cluster usw. Bei zehn Objekten werden also zunächst zehn Cluster gebildet, aus diesen zehn dann neun, aus den neun dann acht usw. Bei den *divisiven Verfahren* wird umgekehrt vorgegangen: Die n Objekte bilden ein einziges großes Cluster, dieses wird dann in zwei Cluster aufgespaltet, die zwei Cluster dann in drei usw. Bei den *partitionierenden Verfahren* muss dagegen eine Clusterzahl vorgegeben werden. Die Objekte werden dann den Clustern so zugeordnet, dass ein bestimmtes Kriterium maximiert bzw. minimiert wird. Das bekannteste partitionierende Verfahren ist das K-Means-Verfahren. Es ordnet die n Objekte K Clustern so zu, dass die Varianz in den Clustern minimiert wird.
- *Probabilistische Clusteranalyseverfahren* (siehe Teil III): Die Klassifikationsobjekte werden den Clustern nicht deterministisch mit einer Wahrscheinlichkeit von 1 oder 0, sondern mit einer dazwischen liegenden Wahrscheinlichkeit zugeordnet. Beispielsweise gehört ein Objekt g mit einer Wahrscheinlichkeit von 0,6 dem Cluster 1, mit einer Wahrscheinlichkeit von 0,2 dem Cluster 2 und mit einer Wahrscheinlichkeit von 0,1 dem Cluster 3 usw. an.

Die Beziehung zwischen unvollständigen, deterministischen und probabilistischen Clusteranalyseverfahren können wir uns wie folgt vorstellen: Eine graphische Bildung von Clustern, wie sie bei den unvollständigen Clusteranalyseverfahren durch den Anwender vorgenommen wird, ist nur in einem ein-, zwei- oder maximal dreidimensionalen Raum bei einer kleinen Klassifikationsobjektmenge möglich. Liegt ein höherdimensionaler Raum oder eine größere Klassifikationsobjektmenge vor, sind formale Verfahren zur Clusterbildung erforderlich. Hinsichtlich der Zuordnung der Klassifikationsobjek-

te verläuft die Formalisierung in zwei Richtungen: Die Klassifikationsobjekte werden deterministisch oder probabilistisch den Clustern zugeordnet. Die probabilistischen Verfahren kann man sich wiederum als Verallgemeinerung der deterministischen Clusteranalyseverfahren vorstellen: Die Annahme, dass jedes Klassifikationsobjekt mit einer Wahrscheinlichkeit von 1 einem oder mehreren Clustern angehört, wird fallengelassen. Eine Zwischenstellung in der vorgenommenen Einteilung der Clusteranalyseverfahren nehmen *überlappende Clusteranalyseverfahren* mit einer deterministischen Zuordnung ein. Bei diesen Verfahren kann ein Klassifikationsobjekt mit einer »Wahrscheinlichkeit« von 1 zwei oder mehreren Clustern angehören. Überlappende Clusteranalyseverfahren werden bei den deterministischen Verfahren behandelt, obwohl auch eine Behandlung im Rahmen der probabilistischen Ansätze möglich ist.

Die hier vorgeschlagene Terminologie zielt auf die Zuordnung der Objekte zu Clustern ab. Dies ist ein mögliches Abgrenzungskriterium. In den letzten Jahren ist auch die Unterscheidung in *modellbasierte* und *heuristische Verfahren* vorzufinden (Fraley und Raftery 1999; Frühwirth-Schnatter 2006). Merkmal modellbasierter Verfahren ist die Annahme eines zugrunde liegenden Wahrscheinlichkeitsmodells, das die untersuchten Daten generiert. Die hier behandelten probabilistischen Clusteranalyseverfahren gehören der Gruppe der modellbasierten Verfahren an. Allerdings gibt es auch modellbasierte Verfahren mit einer deterministischen Zuordnung der Objekte, wie zum Beispiel das in SPSS implementierte *TwoStep-Clusterverfahren*. TwoStep-Cluster nimmt eine deterministische Zuordnung vor, basiert aber auf einem Wahrscheinlichkeitsmodell. Bei der Clusterbildung selbst wird hierarchisch vorgegangen – das Verfahren ließe sich also mehrfach zuordnen. Es wird in dieser Arbeit bei den probabilistischen Verfahren behandelt, da es die Grundkenntnisse der *Maximum-Likelihood-Methode* voraussetzt, die erst bei den probabilistischen Verfahren eingeführt wird. Der Vorteil von modellbasierten Verfahren ist, dass sie formal besser abgesicherte Teststatistiken zur Bestimmung der Clusterzahl bereitstellen. Die Beziehung zwischen der Typologie (»heuristisch« und »modellbasiert«) und der in diesem Buch vorgenommenen Unterscheidung zeigt Abbildung 1.2.

1.4 Grundlage der Clusterbildung

Für die Ähnlichkeit von Objekten innerhalb eines Clusters bestehen unterschiedliche Möglichkeiten. Zwei grundlegende Zugänge können unterschieden werden:

1. Ähnlichkeit hinsichtlich der Merkmalsausprägungen: Die Objekte eines Clusters haben ähnliche Werte in den Merkmalen X_1 , X_2 , usw. und unterscheiden sich in den Merkmalsausprägungen in X_1 , X_2 usw. von den anderen Clustern. An Stelle der UrsprungsvARIABLEN können auch abgeleitete Variablen eingesetzt werden.

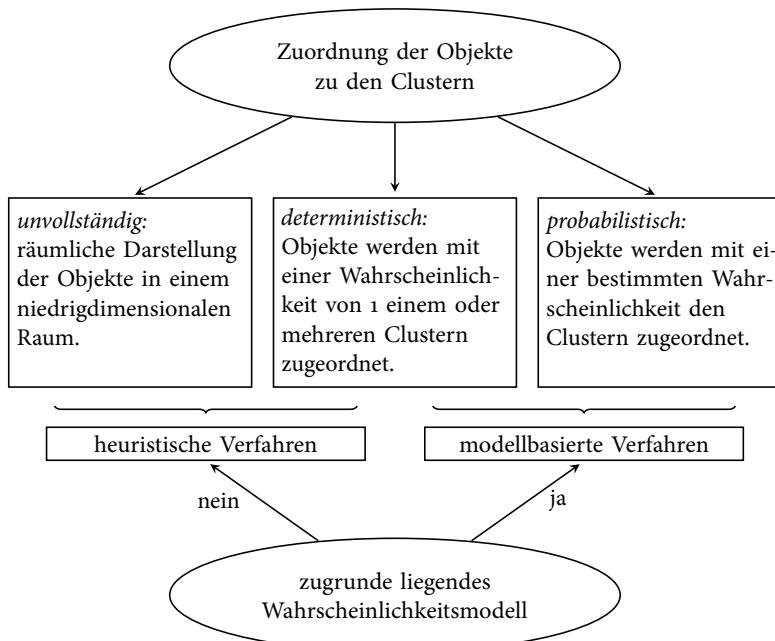


Abb. 1.2: Beziehung unterschiedlicher Typologisierungen von Clusteranalyseverfahren

2. Ähnlichkeit hinsichtlich der Zusammenhangsmuster von Variablen: Für die Objekte eines Clusters ist charakteristisch, dass für sie dieselben Zusammenhänge zwischen den Variablen X_1, X_2 usw. gelten. Objekte unterschiedlicher Cluster sind durch unterschiedliche Zusammenhänge gekennzeichnet.

Übersichtstabelle 1.2 auf der nächsten Seite verdeutlicht die Unterschiede zwischen beiden Zugängen. Die hier behandelten Verfahren basieren auf der ersten Zugangsweise.

Der in Kapitel 16 behandelte *Latent-GOLD-Ansatz* enthält ein latentes Regressionsmodell, das die zweite Zugangsweise ermöglicht, bei der die Cluster auf der Basis von Merkmalszusammenhängen und nicht von Merkmalsausprägungen gebildet werden. Eine Spezialanwendung der Klassifikation auf der Basis von Merkmalszusammenhängen stellt die Analyse von zeitbezogenen Daten dar. Auf diesen Aspekt wird in Kapitel 19 eingegangen.

Formal betrachtet können Klassifikationsobjekte die Zeilen oder Spalten einer Datenmatrix sein. Bei einer *objektorientierten Clusteranalyse* werden – einfach ausgedrückt – Zeilen, die sich ähnlich sind, zu Clustern zusammengefasst. Die Zeilen können entweder Personen oder Aggregate sein. Sie sollen hier unter dem Oberbegriff »Objekte« zusammengefasst werden. Aufgabe einer objektorientierten Clusteranalyse ist somit, Personen

Tab. 1.2: Ähnlichkeit auf der Basis von Merkmalsausprägungen oder Merkmalszusammenhängen

Ähnlichkeit auf der Basis von Merkmalsausprägungen	Ähnlichkeit auf der Basis von Merkmalszusammenhängen
Die Objekte g eines Clusters k haben ähnliche Werte (Merkmalsausprägungen) in den Klassifikationsmerkmalen. Es soll also gelten:	Für die Objekte g aus dem Cluster k gelten dieselben Zusammenhänge in den Klassifikationsmerkmalen. Für Cluster k soll zum Beispiel gelten:
$x_{gj} \approx x_{g^*j}$,	$y_g = \beta_{0k} + \beta_{1k}x_{g1} + \beta_{2k}x_{g2} + \dots + \beta_{pk}x_{gp} + \varepsilon_g$,
für alle g und g^* aus k in allen Variablen j .	für alle g aus k . Charakteristisch für die Cluster sind in dem Beispiel die Regressionskoeffizienten und nicht die Werte der Objekte in den Variablen.
Die Objekte g und g^* unterschiedlicher Cluster k und k^* haben unterschiedliche Werte in den Klassifikationsmerkmalen. Es soll also gelten:	Für Objekte g aus Cluster k und Objekte g^* aus Cluster k^* Cluster soll gelten: $\beta_{kj} \neq \beta_{k^*j}$
$x_{gj} \neq x_{g^*j}$,	in zumindest einer Variablen.
für alle g aus k und g^* aus k^* in allen Variablen j .	<i>Beispiel:</i> Jugendliche werden aufgrund ihrer Freizeitaktivitäten zu Clustern zusammengefasst.
<i>Beispiel:</i> Jugendliche werden danach klassifiziert, wie sich bestimmte Stressoren auf den Alkoholkonsum auswirken. y ist als der Alkoholkonsum definiert, die x_j bilden die Stressoren (zum Beispiel Stress in der Schule, Stress in der Familie, Stress mit Freunden, ...)	

oder Aggregate aufgrund der Ähnlichkeit in bestimmten Variablen in homogene Gruppen zusammenzufassen. Bei einer *variablenorientierten Clusteranalyse* werden dagegen ähnliche Spalten zu Clustern zusammengefasst: Aus einer Menge von Variablen werden aufgrund der Ähnlichkeit in bestimmten Objekten (Zeilen) homogene Gruppen gebildet.

1.5 Konfirmatorische und explorative Clusteranalyse

Clusteranalyseverfahren werden vielfach als *explorative Verfahren* betrachtet. Dies ist zum Teil zutreffend. Viele Verfahren erfordern nur die Spezifikation von Klassifikationsmerkmalen und Objekten, die geclustert werden sollen. Die Zahl der Cluster kann ebenso offen gelassen werden wie die ein Cluster kennzeichnenden Merkmalsausprägungen (Clusterprofile). In diesem Sinn werden Clusteranalyseverfahren explorativ eingesetzt. Sie können aber auch *konfirmatorisch* verwendet werden. In diesem Fall werden die

Tab. 1.3: Explorative und konfirmatorische Clusteranalyse

explorative Clusteranalyse	konfirmatorische Clusteranalyse
<ul style="list-style-type: none"> Die Zahl der Cluster ist unbekannt. Sie muss bestimmt werden. Die Merkmale der Cluster (zum Beispiel Clustermittelwerte beim K-Means-Verfahren) sind unbekannt und müssen ermittelt werden. Die inhaltliche Interpretation der Cluster kann schwierig sein. Die Anpassung an die Daten wird maximiert. 	<ul style="list-style-type: none"> Die Zahl der Cluster ist bekannt und muss nicht bestimmt werden. Die Merkmale der Cluster (zum Beispiel Clustermittelwerte beim K-Means-Verfahren) sind zumindest teilweise bekannt und müssen nicht ermittelt werden. Die Cluster haben bereits vorab eine inhaltliche Bedeutung. Die Anpassung an die Daten kann schlecht sein.

Zahl der Cluster und die Charakteristiken der Cluster vorab definiert. Tabelle 1.3 veranschaulicht die Unterschiede zwischen explorativer und konfirmatorischer Analyse.

Konfirmatorische Verfahren werden in den Abschnitten »konfirmatorische K-Means-Verfahren«, »konfirmatorische Profilanalyse« usw. behandelt. Sie haben zwei zentrale Vorteile:

1. Sie vermeiden das Problem der Bestimmung der Clusterzahl. Dies ist nach wie vor ein kritischer Punkt explorativer Verfahren, der trotz zahlreicher Anläufe noch nicht befriedigend gelöst ist.
2. Die Cluster haben eine substantielle Interpretation. Bei explorativen Verfahren ist es mitunter schwierig, für alle Cluster inhaltlich sinnvoll Namen zu finden. Mitunter verbleibt ein inhaltlich schwer zu interpretierendes »Restcluster«.

Diesen Vorteilen stehen folgende Nachteile gegenüber.

1. Die Modellanpassung an die empirischen Daten ist im Regelfall geringer.
2. Standard-Statistikprogramme, wie IBM-SPSS, enthalten keine konfirmatorischen Clusteranalyseverfahren. Aus diesem Grund werden sie trotz der genannten Vorteile auch selten eingesetzt.

1.6 Anwendungsbeispiele

Nachfolgend sollen einige Anwendungsbeispiele aus dem Bereich der Sozialwissenschaften gegeben werden. Einen allgemeinen Überblick über die Anwendung in allen Wissenschaftsgebieten gibt Kettenring (2006).

Beispiel 1: Clusteranalyse von Variablen (Neumann, Frindte u. a. 1999). 2 500 Jugendliche im Alter von 15 bis 19 Jahren werden in Deutschland nach der Sympathie zu vorgegebenen Jugendgruppen befragt. Auf Basis der Urteile der Jugendlichen wird die Ähnlichkeit der Jugendgruppen ermittelt. Die Ähnlichkeiten werden anschließend mittels einer hierarchischen Clusteranalyse untersucht. In dem Beispiel wird die Ähnlichkeit von Objekten indirekt über die Urteile der Personen erfasst. Als ähnlich werden Jugendgruppen bezeichnet, die ähnliche Sympathieurteile erhalten. Diese Methoden kann auch für andere Aufgabenstellungen verwendet werden, so zum Beispiel kann nach demselben Vorgehen die Ähnlichkeit von politischen Parteien, von Freizeitaktivitäten oder von Produkten bestimmt werden. Ähnlichkeiten können auch direkt erhoben werden. Diesbezügliche Methoden werden in Abschnitt 4.1 dargestellt.

Beispiel 2: Objektorientierte Clusteranalyse einer Fragebatterie (Neumann, Frindte u. a. 1999). Für eine Fragebatterie zur Messung von Fremdenfeindlichkeit wird in der bereits im ersten Beispiel genannten Studie ein K-Means-Verfahren gerechnet, um unterschiedliche Einstellungstypen zu ermitteln. Es werden vier Cluster gefunden. Dieses Vorgehen eignet sich dann, wenn ein heterogenes Antwortverhalten angenommen wird. In diesem Fall ist die üblicherweise zum Einsatz kommende Faktorenanalyse nicht zweckdienlich. Das Vorgehen kann auch zum Auffinden von Response-Sets eingesetzt werden, zum Beispiel zur Bestimmung einer Gruppe von Personen mit Ja-Sage-Tendenz. Anstelle eines K-Means-Verfahrens ist der Einsatz eines probabilistischen Verfahrens zu überlegen, da vermutlich Überlappungen vorliegen.

Beispiel 3: Typen von Jugendlichen (Münchmeier 1997). Auf der Basis der 12. Shell Jugendstudie entwickelt Münchmeier fünf Typen von Jugendlichen: »Kids«, »Gesellschaftskritisch Loyale«, »Traditionelle«, »Konventionelle« und »(Noch-)Nicht-Integrierte«. Interessant ist an dem Beispiel, dass neben Einstellungs- und Interessensvariablen mit dem Geschlecht und dem Alter auch sozialstrukturelle Variablen als aktive Klassifikationsmerkmale eingesetzt werden. Ob dies sinnvoll ist, wird in Abschnitt 18.2 diskutiert.

Beispiel 4: Hierarchische Clusteranalyse von Ländern (Saint-Arnaud und Bernard 2003). Ausgehend von der bekannten Typologie von Esping-Andersen (1990) wird untersucht, ob sich diese auch angesichts der wirtschaftlichen Entwicklung der letzten Jahrzehnte noch nachweisen lassen oder ob eine Konvergenz der Typen beobachtbar ist. Dazu werden geeignete Indikatoren gebildet, die die wohlfahrtsstaatlichen Leistungen erfassen. Daran anschließend werden die Länder mittels hierarchisch-agglomerativer Verfahren geclustert. Die Ergebnisse zeigen, dass sich die unterschiedlichen Wohlfahrtsstaatsregime nach wie vor feststellen lassen. In Teil II werden wir ebenfalls Länder clustern.

Beispiel 5: Bestimmung sozialer Milieus mittels probabilistischer Clusteranalyseverfahren (Pöge 2007). In dieser Arbeit werden Jugendliche aus Münster und Duisburg anhand ihrer Werthaltungen und musikalischen Präferenzen klassifiziert. Die so gefundenen Typologien werden im Hinblick auf deviantes und delinquentes Verhalten untersucht und mit Hilfe weiterer sozialstruktureller Merkmale und Lebensstildimensionen näher beschrieben. Auf der Basis dreier Datensätze der Jahre 2003 und 2005 wird ein stabiler Kern an Musik- und Wertemilieutypen beschrieben. Die Befunde zeigen im Ergebnis, dass Erhebungsort und -zeitpunkt übergreifend, bestimmte Jugendmilieutypen generell stärker und schwächer kriminalitätsbelastet sind als andere, und dass spezielle Formen von abweichendem Verhalten für einige Jugendmilieus typisch sind. In den Kapiteln 12, 14 und 16 werden wir zwar nicht Milieus, aber Wertetypen bestimmen. Um die Berechnungen nachvollziehen zu können, werden nur zwei Merkmale verwendet. Zur Bestimmung einer guten Typologie sind mehr Variablen empfehlenswert.

Beispiel 6: Bestimmung von latenten Konflikten (Bacher 1995b). Mittels Faktorenanalysen werden zunächst für den *Sozialen Survey Österreich* (ssö) – wie in Beispiel 5 – zentrale Wertedimensionen entwickelt. Der ssö ist eine wissenschaftliche Omnibusumfrage, vergleichbar dem ALLBUS, findet aber in wesentlich größeren Abständen statt. Die zentralen Wertedimensionen decken die Einstellungen zu wichtigen gesellschaftspolitischen Themen ab. Als Verfahren wird die latente Profilanalyse eingesetzt. Latente Konflikte sind dadurch erkennbar, dass mehr als eine Klasse vorliegt; dies ist der Fall. Eine zentrale Konfliktlinie bildet die Einstellung zur Rolle der Frau.

Beispiel 7: Bestimmung von statistischen Zwillingen (Bacher 2002). Es wird untersucht, wie sich ein bestimmtes Ereignis oder ein bestimmtes Programm bzw. eine bestimmte Intervention auf definierte Zielvariablen auswirkt. Dazu stehen Daten für die Untersuchungsgruppe und einer Vergleichsgruppe zur Verfügung. Für die Personen der Untersuchungsgruppe sollen in einer Vergleichsgruppe statistische Zwillinge gesucht werden, die sich nur durch die Teilnahme an der Maßnahme oder das Eintreten eines Ereignisses unterscheiden, während sie sich in allen anderen Einflussfaktoren nicht unterscheiden. Dazu können die in Kapitel 8 behandelten Distanzmaße eingesetzt werden. Die Techniken können auch zur Imputation fehlender Werte oder zur Datenfusion eingesetzt werden.

Beispiel 8: Bestimmung von homogenen Ausgangsgruppen zur differentiellen Wirkungsmessung (Peck 2005). Die Teilnehmer und Nicht-Teilnehmer einer Maßnahme werden hinsichtlich der Ausgangsbedingungen, zum Beispiel hinsichtlich Alter, Geschlecht, Berufstätigkeit, Einkommen, zu homogenen Gruppen zusammengefasst. Für jede Gruppe wird anschließend getrennt analysiert, wie sich die Maßnahme auswirkt. Mit dieser Methode können differentielle Wirkungen erfasst werden,

so zum Beispiel kann die Situation eintreten, dass die untersuchte Maßnahme in einem Cluster wirkt, während sie in einem anderen Cluster keine Wirkung entfaltet. Analog zur Clusterung auf der Basis der Ausgangsbedingungen, können auch Cluster von homogenen Programmimplementierungen gebildet werden. Auch die hier verwendeten Techniken können zur Imputation fehlender Werte oder zur Datenfusion eingesetzt werden.

Kritisch anzumerken ist, dass in manchen Publikationen das eingesetzte Verfahren nicht genau spezifiziert wird, sondern nur sehr allgemein darauf verwiesen wird: »Es wird eine Clusteranalyse gerechnet«. Dies ist unbefriedigend und ermöglicht oft keine Einschätzung und Beurteilung des realisierten methodischen Vorgehens. Wünschenswert sind genauere Angaben zum gewählten Verfahren. Aus der methodischen Darstellung sollte ersichtlich sein:

- die ausgewählten Variablen
- die ausgewählten Objekte
- gegebenenfalls die eingesetzten Datentransformationen
- das gewählte (Un-)Ähnlichkeitsmaß
- das ausgewählte Verfahren (genaue Bezeichnung)
- das eingesetzte Computerprogramm (inklusive Versionsnummer)
- die technischen Voreinstellungen
- die verwendeten Kriterien zur Bestimmung der Clusterzahl
- die durchgeführte Stabilitätsprüfung
- die durchgeführte Validitätsprüfung

Eine methodische Beschreibung sollte – stichpunktartig skizziert – wie folgt aussehen: »Für die Analyse wurden die Variablen X_1 , X_2 , usw. ausgewählt. In die Analyse wurden alle Befragten einbezogen. Die Variablen wurden z-transformiert, um Vergleichbarkeit zu erreichen. Die Clusteranalyse wurde mit IBM-SPSS 18 gerechnet. Eingesetzt wurde das Ward-Verfahren mit quadrierten euklidischen Distanzen und den technischen Standardeinstellungen, da das Verschmelzungsschema eine klare Interpretation besitzt. Die Bestimmung der Clusterzahl erfolgte mittels des inversen Scree-Tests. Zur Stabilitätsprüfung wurden für die ausgewählte Clusterzahl noch drei weitere Analysen mit dem Complete-, Single- und Weighted-Average-Linkage gerechnet. Die Übereinstimmung wurde mittels des adjustierten Randindex (Hubert und Arabie 1985) beurteilt. Zur Validitätsprüfung wurde auf die Variablen Z_1 , Z_2 usw. zurückgegriffen.«

1.7 Modellprüfung und Validierung

Inwiefern die Vorstellungen, die mit der Wahl eines bestimmten Clusteranalyseverfahrens verbunden sind, zutreffen, ist in jedem konkreten Anwendungsfall zu prüfen. Die Prüfung umfasst folgende Schritte:

1. *Prüfung der Modellanpassung:* Bei der Anwendung eines Clusteranalyseverfahrens werden Maßzahlen berechnet, die entweder angeben, wie gut die (zentralen) Vorstellungen an eine gute Klassifikation erfüllt sind, oder die prüfen, wie gut die berechneten Cluster mit den Daten übereinstimmen. Der Fokus liegt in der Regel auf den Kriterien 1 bis 3 einer guten Klassifikation aus Abschnitt 1.2. Eine gute Modellanpassung liegt dann vor, wenn die Kriterien 1 bis 3 (und eventuell 7 und 8) erfüllt sind. Eine gute Modellanpassung ist eine notwendige Voraussetzung für eine inhaltliche Interpretation. Liegt sie nicht vor, ist eine Fehleranalyse für die schlechte Modellanpassung durchzuführen.
2. *Prüfung der inhaltlichen Interpretierbarkeit:* Liegt eine gute Modellanpassung vor, ist eine inhaltliche Interpretation durchzuführen. Ist eine inhaltliche Interpretation nicht möglich, ist ebenfalls eine Fehleranalyse erforderlich.
3. *Stabilitätsprüfung:* Es wird untersucht, wie stark sich die Ergebnisse ändern, wenn geringfügige Modifikationen in den Daten und/oder in den getroffenen Spezifikationen des gewählten Verfahrens vorgenommen werden. Führen geringfügige Änderungen zu stark abweichenden Ergebnissen, liegt Instabilität vor und eine Fehleranalyse ist durchzuführen.
4. *Inhaltliche Validitätsprüfung:* Die inhaltliche oder kriterienbezogene Validitätsprüfung besteht darin, dass Hypothesen über den Zusammenhang der ermittelten Cluster mit nicht in die Clusterbildung einbezogenen Merkmalen formuliert werden. Beispielsweise: »Cluster C_1 hat in der KriterienvARIABLE Z_1 einen höheren Wert als Cluster C_2 « oder »das Auftreten von C_1 hängt von Z_1, Z_2 usw. ab« oder » Z_3 tritt häufiger in C_3 auf«. Können die Hypothesen nicht bestätigt werden, ist wiederum eine Fehleranalyse erforderlich.

In der neueren clusteranalytischen Literatur (Everitt, Landau u. a. 2001; Gordon 1999) wird von einer *formalen Gültigkeits-* bzw. *formalen Validitätsprüfung* gesprochen. Die Verwendung dieses Begriffes ist jedoch nicht einheitlich. Wir wollen den Begriff als umfassendes Konzept verwenden und von einer formal gültigen Clusteranalyse dann sprechen, wenn alle formalen Kriterien aus Abschnitt 1.2 erfüllt sind, wenn also die gefundene Klassifikation folgende formale Kriterien erfüllt:

- Die Cluster sind in sich homogen.
- Die Cluster sind voneinander verschieden.
- Die Clusterstruktur erklärt die Daten.

- Die Cluster und die Clusterstruktur sind stabil.
- Die Cluster haben eine bestimmte Mindestgröße und ihre Anzahl ist überschaubar (optional).

Formale Gültigkeit und Modellanpassung unterscheiden sich darin, dass bei der formalen Gültigkeitsprüfung die Stabilität, die Clustergröße und die Clusterzahl als weitere Kriterien miteinfließen, während sich die Modellanpassung auf die Homogenität in den Clustern, die Heterogenität zwischen den Clustern und die Erklärungskraft der Clusterstruktur konzentriert.

Bei der formalen Gültigkeitsprüfung (und der Modellanpassung) ist wichtig, dass alle relevanten Kriterien für eine gute Clusterlösung betrachtet werden, unabhängig davon, ob sie in das konkrete Verfahren explizit einfließen oder nicht (siehe Abbildung 1.3). In dem fiktiven Beispiel der Abbildung werden an die gesuchte Klassifikation die Kriterien $K_1, K_2, K_3, \dots, K_q$ gestellt. In das Clusterverfahren fließen die Kriterien K_1, K_2 und K_3 ein, nicht berücksichtigt bei der Berechnung der Cluster werden die Kriterien K_4, K_5, \dots, K_q . Aufgabe der formalen Gültigkeitsprüfung ist die Ermittlung von Maßzahlen $M(K_i)$, die angeben, wie gut alle Kriterien K_i , also K_1, K_2, \dots, K_q , erfüllt sind. Absolute Schwellenwerte für diese Maßzahlen fehlen häufig, daher ist ein relativer Einsatz sinnvoll, bei dem mehrere Lösungen miteinander verglichen werden und schließlich jene Lösung ausgewählt wird, die insgesamt am besten abschneidet. In Kapitel 20 wird ein Vorgehen dargestellt, bei dem beliebig viele und unterschiedlich skalierte Maßzahlen in die Beurteilung der formalen Gültigkeit einbezogen werden.

1.8 Fehleranalyse

Eine *Fehleranalyse* ist immer dann erforderlich, wenn die mit der Datenanalyse verbundenen Zielsetzungen nicht erreicht werden können. So zum Beispiel kann das Ergebnis der Anwendung eines Clusteranalyseverfahrens darin bestehen, dass eine schlechte Modellanpassung vorliegt, dass eine inhaltliche Interpretation der Cluster nicht möglich ist, dass die Ergebnisse instabil oder dass die Ergebnisse inhaltlich nicht valide sind. Die Datenanalyse ist in diesen Fällen als gescheitert zu betrachten. Eine Diagnose möglicher Fehlerursachen ist erforderlich, um nicht auf einem unbefriedigenden Erkenntnisstand zu verweilen. Fehlerursachen können allgemein sein:

1. Die theoretischen Annahmen, die der untersuchten Fragestellung zugrunde liegen, sind »falsch«.

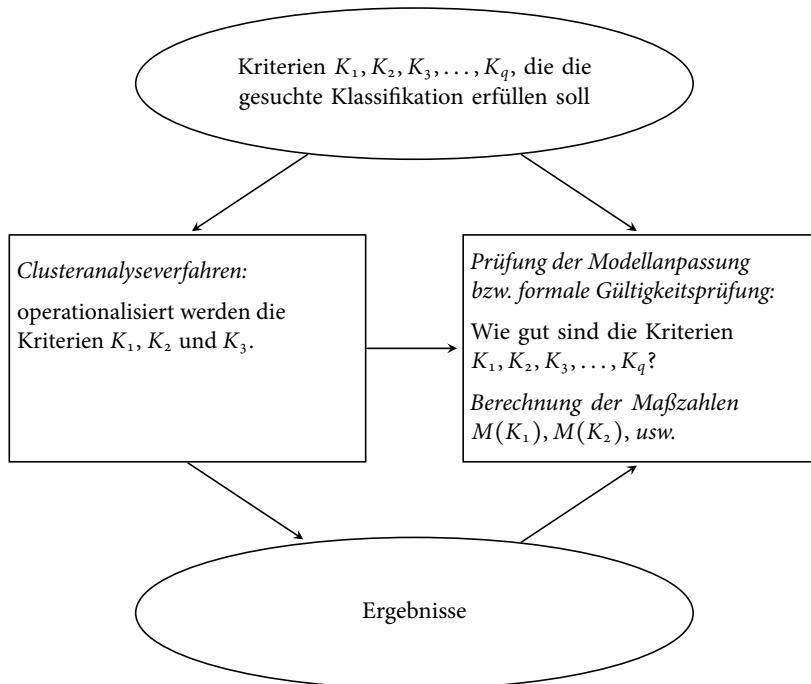


Abb. 1.3: Grundidee der Prüfung der Modellanpassung bzw. der formalen Gültigkeitsprüfung

2. Die Daten sind zur Beantwortung der untersuchten Fragestellung nicht geeignet. Ursachen hierfür können beispielsweise zu stark fehlerbehaftete Variablen oder Objekte sein. Aber auch eine falsche Aggregierung kann eine mögliche Ursache sein.
3. Das verwendete Verfahren ist zur Beantwortung der untersuchten Fragestellung ungeeignet. Das Verfahren trifft Annahmen, die mit den theoretischen Vorstellungen nicht übereinstimmen, oder es ist für die vorliegenden Daten ungeeignet.

Das Vorgehen einer Fehlerdiagnose besteht aus folgenden Schritten:

Schritt 1: Spezifiziere alle möglichen Fehlerursachen.

Schritt 2: Bestimme bei jeder Fehlerursache, ob sie sich anhand der vorliegenden Daten empirisch überprüfen lässt.

Abhängig von den Ergebnissen des Schritts 2 wird für jede Fehlerquelle entweder Schritt 3 oder Schritt 4 ausgeführt:

Schritt 3: Überprüfe die Fehlerquelle mit den vorliegenden Daten.

Schritt 4: Erzeuge und/oder erhebe weitere Daten für eine empirische Überprüfung der Fehlerquelle. Mit »erzeugen« ist das Durchführen von Simulationsrechnungen

mit fiktiven Modelldaten gemeint, etwa um die Eigenschaften eines Verfahrens zu ermitteln. Mit »erheben« ist die Erhebung zusätzlicher Daten gemeint.

Da mehrere Fehlerquellen untersucht werden, werden bei einer Fehleranalyse im Regelfall sowohl Schritt 3 als auch Schritt 4 ausgeführt. Es folgen:

- Schritt 5:* Beurteile aufgrund der vorausgehenden Ergebnisse sowie der verfügbaren Literatur die Wahrscheinlichkeit, mit der die einzelnen Fehlerquellen auftreten. Das Ergebnis kann darin bestehen, dass nur eine Fehlerquelle einen hohen Wahrscheinlichkeitswert besitzt oder mehrere Fehlerquellen gleich wahrscheinlich sind.
- Schritt 6:* Abhängig von dem wahrscheinlichen Fehler kann folgende Entscheidung getroffen werden:

1. Die Datenanalyse wird mit einer geänderten Theorie fortgesetzt. Die Ausgangstheorie ist falsifiziert.
2. Die Datenanalyse wird subgruppenspezifisch (Datenfehler: Personen) oder mit anderen Variablen (Datenfehler: Variable) fortgesetzt oder abgebrochen, da die Daten zu »schlecht« sind.
3. Die Datenanalyse wird mit einem anderen Verfahren fortgesetzt.

Das hier dargestellte Vorgehen entspricht der pragmatischen Modellkonzeption von Kreutz und Bacher (1991), bei der alle möglichen alternativen Erklärungen (in diesem Fall Fehlerquellen) in Erwägung gezogen und überprüft werden.

1.9 Datenanalyse als iterativer Prozess

Zwischen der Fehlerdiagnose bzw. Fehleranalyse und der inhaltlichen (kriterienbezogenen) Validitätsprüfung besteht kein formaler Unterschied bei der Vorgehensweise. In beiden Situationen werden Hypothesen formuliert, die empirisch geprüft werden. Bei der Validitätsprüfung sind es Hypothesen der folgenden Art: »Cluster C unterscheidet sich in der Variablen Z (Kriterienvariablen) von den anderen Clustern C*« oder »in Cluster C hat die Variable Z₁ einen höheren Wert als die Variable Z₂« usw. Bei der Fehleranalyse werden Hypothesen der Art: »Fehlerursache U erklärt die schlechte Modellanpassung« usw. spezifiziert. Wie bei der Fehleranalyse ist auch im Rahmen einer inhaltlichen Validitätsprüfung die Spezifikation aller möglichen Erklärungen sinnvoll.

Sowohl die Fehler- als auch die Validitätsprüfung stellen Schritte eines Datenanalyseprozesses dar, der in der Regel mehrfach durchlaufen wird, bis eine befriedigende Erklärung der Ergebnisse gefunden ist. Abbildung 1.4 verdeutlicht diesen Prozess. An das Datenmaterial wird eine bestimmte Fragestellung F₁ herangetragen, die mit einer geeigneten

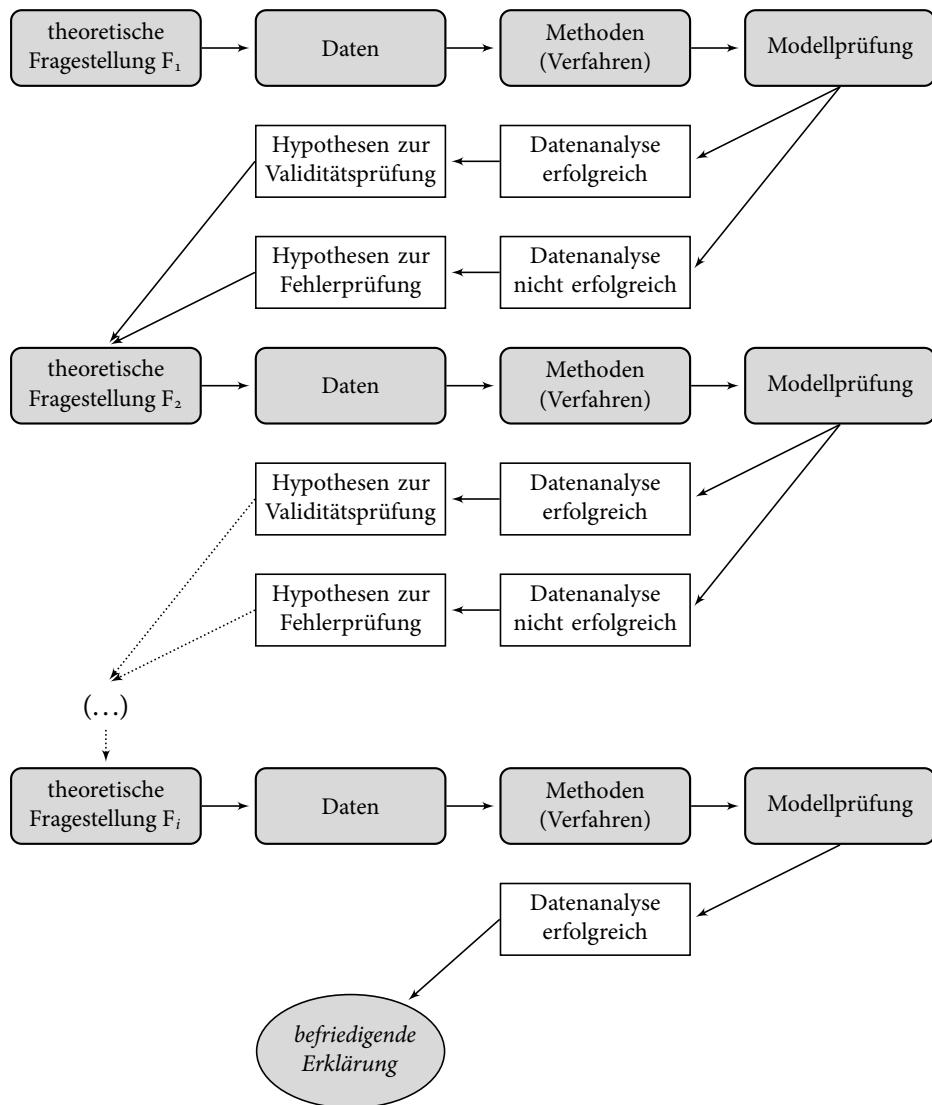


Abb. 1.4: Die Datenanalyse als iterativer Prozess

Methode untersucht wird. Für die Ergebnisse wird eine Modellprüfung durchgeführt, die dazu führen kann, dass die Datenanalyse als erfolgreich oder gescheitert zu betrachten ist. Abhängig von diesem Ergebnis werden Hypothesen für eine Validitätsprüfung oder Fehlerprüfung formuliert. Diese stellen dann die neue Fragestellung F_2 dar, die mit den Daten untersucht wird. Es wird ein geeignetes Verfahren ausgewählt. Dieser Prozess wird solange wiederholt, bis eine befriedigende Erklärung vorliegt. Damit ist eine Erklärung gemeint, die Bunge (1967, S. 25–44) mit »interpretativer« Erklärung umschreibt. Kennzeichen einer derartigen Erklärung ist, dass sie den »Modus operandi« angibt, dass sie also spezifiziert, wie das Ergebnis (das zu erklärende Phänomen) zustande gekommen sein kann. Dies kann nach Bunge nur durch einen bestimmten Grad an Tiefe erreicht werden, wenn also die Erklärung auf zugrunde liegende latente Dimensionen, Strukturen, Prozesse und Interaktionen zurückgreift.

Zum Auffinden einer befriedigenden Erklärung kann der ergänzende Einsatz von qualitativen Erhebungsmethoden (Leitfadeninterview, fokussiertes Interview, Gruppendiskussion) zweckmäßig sein. Das Vorgehen könnte wie folgt aussehen: Mittels einer Clusteranalyse werden Typen gebildet, die dann »qualitativ« befragt werden, indem zum Beispiel ein typischer Fall für jedes Cluster ausgewählt wird.

1.10 Computerprogramme

Die meisten Standardprogramme, wie IBM-SPSS, SAS, STATA und SYSTAT enthalten Programme für die hier behandelten unvollständigen und deterministischen Clusteranalyseverfahren. Allerdings werden häufig keine oder nur wenige Maßzahlen zur Modellbeurteilung und zur Validierung zur Verfügung gestellt. Probabilistische Verfahren sind teilweise implementiert. IBM-SPSS bietet mit TwoStep-Cluster ein hierarchisches modellbasiertes Verfahren an, das aber im Vergleich zu anderen modellbasierten Verfahren eine schlechtere Performanz zeigt (siehe Abschnitt 17.3). Für STATA gibt es mit GLLAMM² ein leistungsfähiges Zusatzmodul zur probabilistischen Clusteranalyse.

Neben den Standardprogrammen steht Einzelsoftware zur Verfügung – einen Softwareüberblick geben Everitt, Landau u. a. (2001, S. 197–208) –, die zum Teil entgeltlich verfügbar ist, zum Teil aber auch unentgeltlich aus dem Internet heruntergeladen werden kann. Entgeltliche Produkte sind beispielsweise:

- Clustan³: Dieser »Klassiker« hat seine Stärken in der graphischen Datenausgabe, in der Verfügbarkeit von Teststatistiken und in einem elaborierten K-Means-Verfahren,

² <http://www.gllamm.org/>

³ <http://www.clustan.com/>

das mit mehreren zufälligen Startwerten rechnet und die Definition von Ausreißern zulässt.

- Latent GOLD⁴: Die Stärken dieses Programmes liegen in der probabilistischen Modellierung. Es erlaubt die Analyse unterschiedlicher Skalentypen, enthält neben der Clusteranalyse auch ein latentes Regressionsmodell und eine ordinale Faktorenanalyse, vermeidet degenerierte Lösungen und vieles mehr. Es wird ausführlich in Kapitel 16 beschrieben.

Beispiele für unentgeltliche Produkte sind:

- Das freie Softwarepaket R⁵. Neben dem Paket »cluster« von Maechler, Rousseeuw u. a. (2005), das aktualisierte und erweiterte Algorithmen und Begleitmaterial zu Kaufman und Rousseeuw (1990) bereitstellt, gibt es für R eine Vielzahl von – zum Teil hochspezialisierten – Erweiterungen im Bereich Clusteranalyse⁶ und maschinellem Lernen. Mit einem dieser Erweiterungspakete, MCLUST, haben wir in Abschnitt 17.3 Simulationsrechnungen durchgeführt.
- WEKA⁷ ist eine in Java programmierte Data-Mining-Software, in die unter anderem auch Algorithmen aus dem Bereich des »unsupervised learning« integriert sind. Auch für sie liegen in Abschnitt 17.3 Simulationsergebnisse vor.
- Ein noch breiter aufgestelltes Projekt ist KNIME⁸, der »Konstanz Information Miner«, das zusätzliche Algorithmen integriert hat.
- Auch mit AutoClass⁹, einer Entwicklung der NASA für bayesianische Clusteranalyse haben wir im Rahmen von Simulationsrechnungen Erfahrungen sammeln können.
- CLUSTER 3.0¹⁰ bietet eine Software zum K-Median-Verfahren an (siehe Abschnitt 12.14).

Die meisten Beispiele im vorliegenden Buch wurden mit dem Programm Paket ALMO¹¹ gerechnet. ALMO enthält unvollständige, deterministische und probabilistische Verfahren und stellt die im Buch behandelten Maßzahlen zur Beurteilung der Modellanpassung zur Verfügung. ALMO kann unentgeltlich beim Erstautor angefordert werden. Für die probabilistische Clusteranalyse empfehlen wir Latent GOLD, ALMO bietet hier ein gutes Einstiegsmodell an.

4 http://www.statisticalinnovations.com/products/latentgold_v4.html

5 <http://www.r-project.org>

6 <http://cran.r-project.org/web/views/Cluster.html>

7 <http://www.cs.waikato.ac.nz/ml/weka/>

8 <http://knime.org>

9 <http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/>

10 <http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm>

11 <http://www.almo-statistik.de/>

Teil I

Unvollständige Clusteranalyseverfahren

2 Einleitende Übersicht

Unvollständige Clusteranalyseverfahren sind dadurch gekennzeichnet, dass sie keine Zuordnung der Klassifikationsobjekte zu Clustern vornehmen, sondern nur eine räumliche Darstellung der Klassifikationsobjekte in einem niedrigdimensionalen – idealer Weise in einem zweidimensionalen – Raum berechnen.¹ In diesem Kapitel werden folgende unvollständige Clusteranalyseverfahren behandelt:

1. *Multiple Korrespondenzanalyse* (Kapitel 3): Die multiple Korrespondenzanalyse ist eine Art *Faktorenanalyse* – genauer: eine Art *Hauptkomponentenanalyse* (siehe dazu später) – für nominalskalierte Variablen. Es wird eine dimensionale Darstellung der Ausprägungen der untersuchten nominalen Variablen berechnet. Die multiple Korrespondenzanalyse stellte in den 1980er und 1990er Jahren ein relativ »neues« Verfahren der Datenanalyse² dar, deren Anwendung und Interpretation oft mit Problemen und Fehlern verbunden war. In der Zwischenzeit sind einführende Bücher (Blasius 2001) verfügbar. Auch Weiterentwicklungen für ordinale Variablen existieren (SPSS Inc. 2008).
2. *Nichtmetrische mehrdimensionale Skalierung*³ (Kapitel 4): Bei der nichtmetrischen mehrdimensionalen Skalierung können Variablen beliebigen Messniveaus unter

¹ In der clusteranalytischen Literatur werden diese Verfahren daher auch als geometrische Methoden bezeichnet (Gordon 1981, S. 80–120).

² Das Adjektiv »neu« wurde in Anführungszeichen gesetzt, da die multiple Korrespondenzanalyse keinesfalls ein erst in jüngster Zeit entwickeltes Verfahren ist. Das Verfahren wurde unter unterschiedlichen Bezeichnungen bereits in den 1930er und 1940er Jahren von verschiedenen Autoren – unter anderem von Horst (1935) und Guttman (1941) – zur Analyse nominalskalierter Variablen entwickelt. Zur Geschichte der multiplen Korrespondenzanalyse siehe die Arbeit von Tenenhaus und Young (1985). Auch Greenacre (1984, S. 7–11) gibt einen kurzen historischen Überblick. Unter anderem werden folgende synonyme Bezeichnungen für die multiple Korrespondenzanalyse verwendet: »Appropriate Scaling Method«, »Dual Scaling«, »Homogeneity Analysis«, »Optimal Scaling«, »Quantification Method« und »Scalogram Analysis« (Tenenhaus und Young 1985).

³ Die nichtmetrische mehrdimensionale Skalierung, ordinale mehrdimensionale Skalierung oder nichtmetrische multidimensionale Skalierung wird auch als nichtlineares Projektionsverfahren (Jain und Dubes 1988), als Gradientenverfahren zur Entdeckung von Punktekonfigurationen in Minkowskiraumen (Sixtl 1982, S. 341–352), als Verfahren nach Kruskal (Hartung und Elpelt 1984, S. 405–420), Kruskalverfahren, Ähnlichkeitsstrukturanalyse oder englisch »nonmetrical multidimensional analysis« (MDA) bezeichnet.

Ausschöpfung des jeweiligen Informationsgehalts analysiert werden. In die nichtmetrische mehrdimensionale Skalierung geht nur die ordinale Information der Unähnlichkeiten oder Ähnlichkeiten zwischen den Objekten ein, also zum Beispiel – wenn Ähnlichkeiten mittels Korrelationen erfasst werden – die Information »Variable A und B korrelieren stärker als die Variablen A und C usw.« oder »Person A und B sind ähnlicher als Person D und F usw.«. Neben der nichtmetrischen mehrdimensionalen Skalierung gibt es auch metrische Ansätze, welche die Information »Variable A und B korrelieren zweimal so stark wie Variablen A und C« (proportionale mehrdimensionale Skalierung) oder die Information »Zwischen der Korrelation der Variablen A und B und den Korrelationen der Variablen A und C besteht ein linearer Zusammenhang« (lineare mehrdimensionale Skalierung) nutzen.

3. *Bivariate Korrespondenzanalyse* (Abschnitt 5.1): Sie ist eine Art kanonische Korrelationsanalyse für nominalskalierte Variablen. Im Unterschied zur multiplen Korrespondenzanalyse werden zwei Variablengruppen gebildet und eine dimensionale Darstellung der beiden Variablengruppen wird gesucht. Trotz dieser Differenz können die Ergebnisse beider Verfahren ineinander überführt werden. Ob man sich bei einer Analyse für die multiple Korrespondenzanalyse entscheidet oder für die bivariate Korrespondenzanalyse, hängt von der inhaltlichen Fragestellung ab.
4. *Nomiale Faktorenanalyse nach McDonald* (Abschnitt 5.2): Die nominale Faktorenanalyse kann man sich als Alternative zur multiplen Korrespondenzanalyse vorstellen. Ihr liegt wie der Faktorenanalyse ein klares faktorenanalytisches Modell zugrunde, was bei der multiplen und bivariaten Korrespondenzanalyse nicht der Fall ist.
5. *Hauptkomponentenanalyse und Faktorenanalyse für quantitative, ordinale und dichotome Variablen* (Abschnitt 5.3): Die Hauptkomponentenmethode stellt das Pendant zur multiplen Korrespondenzanalyse für quantitative und ordinale Variablen dar. Ihr Ziel ist das Auffinden einer graphischen Darstellung von Objekten. Die Faktorenanalyse geht darüber hinaus. Ihr Ziel ist das Auffinden von gemeinsamen, inhaltlich sinnvoll zu interpretierenden Dimensionen (Faktoren). In dem Modell, auf dem die Faktorenanalyse basiert, wird die Annahme getroffen, dass den untersuchten Variablen q gemeinsame Faktoren zugrunde liegen und dass zusätzlich jede Variable Messfehler und spezifische Eigenschaften aufweist. Bei der Hauptkomponentenmethode wird dagegen angenommen, dass nur gemeinsame Faktoren vorliegen, die nicht notwendigerweise inhaltlich interpretierbar sein müssen.

Forschungspraktisch können die Verfahren unter anderem für folgende Aufgabenstellungen eingesetzt werden:

1. *als unvollständige Clusteranalyseverfahren*. Bei der Interpretation werden Cluster von Objekten gebildet. Voraussetzung dafür ist, dass eine graphische Darstellung möglich ist, dass also maximal drei Dimensionen vorliegen und die Zahl der untersuchten

Objekte nicht zu groß ist. Das Ziel der Analyse besteht hier darin, graphisch eine Clusterbildung vorzunehmen.

2. als *Hilfsverfahren zur Berechnung einer neuen Datenmatrix* für ein deterministisches Clusteranalyseverfahren. Liegen mehr als drei Dimensionen oder eine große Objektzahl vor, ist eine graphische Clusterbildung nicht mehr möglich. In diesem Fall kann folgendes Verfahren gewählt werden:
 - Für die untersuchten Variablen oder Objekte wird ein unvollständiges Clusteranalyseverfahren angewendet.
 - Dabei zeigt sich zum Beispiel, dass vier Dimensionen für eine gute Modellanpassung erforderlich sind.
 - Die berechnete Koordinatenmatrix in den vier Dimensionen wird als neue Datenmatrix gespeichert.
 - Für diese neue Distanzmatrix wird eine deterministische Clusteranalyse durchgeführt.
3. als *Verfahren der Faktorenanalyse*. Im Unterschied zum Einsatz als unvollständiges Clusteranalyseverfahren wird keine Bildung von Clustern angestrebt, sondern eine inhaltliche Interpretation der berechneten Dimensionen. Bei der Analyse wird von der Annahme ausgegangen, dass gemeinsame, inhaltlich interpretierbare Dimensionen (Faktoren) vorliegen.

Die unter 1 und 2 dargestellte Vorgehensweise geht von der Annahme aus, dass den untersuchten Daten ein clusteranalytisches Modell mit inhaltlich interpretierbaren Clustern zugrunde liegt, während bei der Anwendung der faktorenanalytischen Interpretation von einem dimensionalen Modell mit inhaltlich interpretierbaren Dimensionen ausgegangen wird. Für eine faktorenanalytische Interpretation eignet sich insbesondere die Faktorenanalyse (die »gewöhnliche« Faktorenanalyse für quantitative, ordinale und dichotome Faktoren und die Faktorenanalyse nach McDonald für nominale Variablen). Aber auch bei den anderen Verfahren wird oft versucht, die ermittelten Dimensionen inhaltlich zu benennen. Ob in einem konkreten Anwendungsfall eine clusteranalytische und/oder faktorenanalytische Interpretation möglich ist, kann erst aufgrund der Ergebnisse entschieden werden. Formal sind vier Ergebnisse denkbar: Beide Interpretationen sind den Daten angemessen, eine faktorenanalytische Interpretation ist den Daten angemessener, eine clusteranalytische Interpretation ist den Daten angemessener und beide Interpretationen sind den Daten nicht angemessen. Die Abbildung 2.1 auf der nächsten Seite gibt ein Beispiel für zwei mögliche Ergebnisse. In der Teilabbildung 2.1a ist eine faktorenanalytische Interpretation möglich. Die beiden untersuchten Variablen (Schulbildung und Einkommen) laden fast ausschließlich auf einer Dimension, die sich als soziale Schichtungsdimension interpretieren lässt. Eine clusteranalytische Interpretation ist hier weniger angemessen. In der Teilabbildung 2.1b ist dagegen eine clusteranalyti-

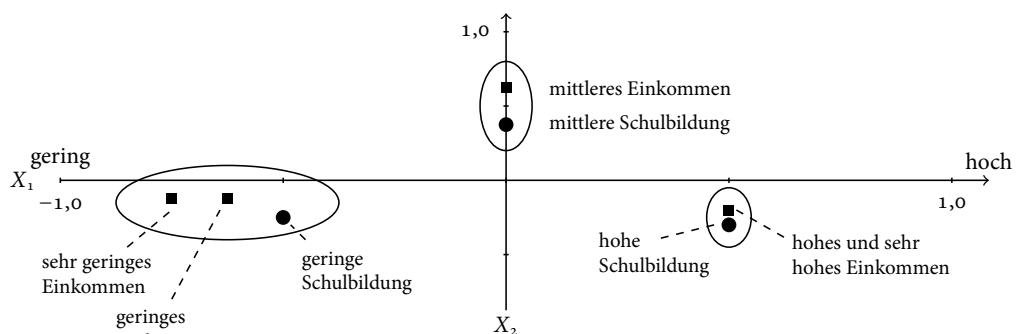
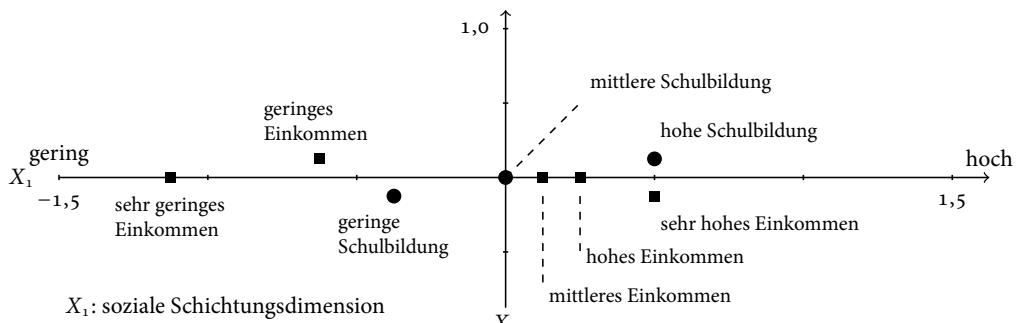


Abb. 2.1: Fiktive Ergebnisse einer multiplen Korrespondenzanalyse

sche Interpretation angemessen. In der zweidimensionalen Lösung sind drei Cluster gut erkennbar. Auch eine dimensionale Interpretation ist möglich. Die erste Dimension lässt sich als soziale Schichtungsdimension interpretieren, die zweite als Mittelschicht versus Ober- und Unterschicht.

Das Vorgehen der Datenanalyse ist bei allen Verfahren und Aufgabenstellungen dasselbe und besteht aus folgenden Schritten:

- 1 Auswahl der zu untersuchenden Variablen und Objekte (Zeilen der Datenmatrix).
- 2 Durchführen von Datentransformationen, sofern diese erforderlich sind.

Diese beiden Schritte müssen bei jeder Datenanalyse aufgrund inhaltlicher und empirischer Kriterien durchgeführt werden und sollen hier nicht näher behandelt werden. Die nächsten Schritte sind:

- 3 *Auswahl eines Unähnlichkeits- oder Ähnlichkeitsmaßes* bei der nichtmetrischen mehrdimensionalen Skalierung, sofern nicht eine direkt erhobene (Un-)Ähnlichkeitsmatrix untersucht wird. (Un-)Ähnlichkeitsmaße werden ausführlich in Kapitel 8 behandelt.
- 4 *Berechnung einer Zusammenhangs-, Ähnlichkeits- oder Unähnlichkeitsmatrix*. Dies wird im Regelfall von der verwendeten Software vorgenommen.
- 5 *Bestimmung der Zahl maximal möglicher Dimensionen* (Ausnahme: nichtmetrische mehrdimensionale Skalierung). Ist die maximale Dimensionszahl zum Beispiel gleich 10, sind zehn Lösungen, nämlich die eindimensionale, die zweidimensionale Lösung usw., möglich.
- 6 *Auswahl der bedeutsamen Dimensionen*. Hier wird die Zahl der möglichen Lösungen weiter eingegrenzt.

Die nächsten Schritte hängen davon ab, ob eine faktorenanalytische und/oder clusteranalytische Interpretation gesucht wird. Für eine *faktorenanalytische Interpretation* sind die weiteren Schritte:

- 7a *Auswahl erforderlicher Dimensionen für eine faktorenanalytische Interpretation*. Die Zahl der Lösungen wird weiter eingegrenzt. Es werden nur jene Lösungen mit einer guten Modellanpassung ausgewählt.
- 8a *Inhaltliche Interpretation der Dimensionen*: Eine gute Modellanpassung ist wertlos, wenn die entsprechende Lösung nicht inhaltlich interpretierbar ist. Liegen mehrere, inhaltlich interpretierbare Lösungen mit guter Modellanpassung vor, muss der Anwender oder die Anwenderin eine Entscheidung für eine bestimmte Lösung treffen – sinnvoll ist dabei eine Zuhilfenahme von Validierungsergebnissen. Es wird dann die valide Lösung ausgewählt.
- 9a *Validitätsprüfung der inhaltlichen Interpretation*. Erst bei einer erfolgreichen Validitätsprüfung ist die faktorenanalytische Interpretation abgeschlossen.

Bei einer *clusteranalytischen Interpretation* ist das Vorgehen ähnlich jenem der faktorenanalytischen Interpretation. Die Schritte sind:

- 7b *Auswahl der erforderlichen Dimensionszahl* für die clusteranalytische Interpretation (analog zu Schritt 7a).
- 8b *Graphische Darstellung der Ergebnisse* mit anschließender graphischer Clusterbildung, sofern dies möglich ist.
- 9b *Inhaltliche Interpretation der Cluster* (analog zu Schritt 8a).
- 10b *Validitätsprüfung der Cluster* (analog zu Schritt 9a).

- 11b Berechnung einer Distanzmatrix, wenn keine graphische Darstellung möglich ist.
12b Durchführen einer hierarchischen Clusteranalyse mit den dort angegebenen Schritten
(siehe Kapitel 6 als einleitende Übersicht).

Werden keine validen Ergebnisse gefunden, ist eine Fehleranalyse nach der in Kapitel 1 angeführten Logik durchzuführen.

3 Multiple Korrespondenzanalyse

3.1 Ein Anwendungsbeispiel

Bei der Auswertung von sozialwissenschaftlichen Umfragen tritt häufig das Problem auf, aus Variablen, die den sozialen Status einer Person charakterisieren, wie zum Beispiel ausgeübte berufliche Tätigkeit, abgeschlossene Schulbildung und Einkommen, einen sozialen Schichtindex zu bilden oder eine Einteilung (Klassifikation) in soziale Schichten vorzunehmen oder die soziale Schichtungsstruktur dimensional darzustellen. Sind die Variablen quantitativ, kann zur Bestimmung der Dimensionalität eine Faktorenanalyse gerechnet werden. Bei nominalskalierten Variablen, wie zum Beispiel Berufsbezeichnungen, ist die Anwendung der Faktorenanalyse nicht mehr möglich. In dieser Situation kann die *multiple Korrespondenzanalyse* (»multiple correspondence analysis«, MCA) eingesetzt werden. Ihre primäre Aufgabe besteht darin, für die Zusammenhangsstruktur zwischen zwei oder mehreren nominalen Variablen eine räumliche Darstellung in einem quantitativen Merkmalsraum zu finden. Sie eignet sich aber auch zum Auffinden von abgeleiteten Skalenwerten (faktorenanalytische Interpretation).

Wir wollen dazu als Beispiel die Daten einer Untersuchung über die Lebensbedingungen von Kindern in Österreich verwenden (Wilk und Bacher 1993). In der Untersuchung wurden Kinder im Alter von ungefähr zehn Jahren ($n = 2745$) sowie deren Eltern und Klassenlehrer befragt. Die Angaben zur sozialen Schichtzugehörigkeit der Kinder wurden über die Eltern erfragt, da sich in Pretests zeigte, dass die Kinder nur sehr unvollständige Angaben über soziale Schichtungsmerkmale machen konnten. Bei den Eltern wurden die in der Abbildung 3.1 auf Seite 45 dargestellten Informationen zur sozialen Schichtzugehörigkeit erhoben.

In die Analyse werden nur jene Kinder einbezogen, für die Angaben über beide Elternteile vorliegen ($n = 1823$). Die Einkommensangaben müssen für die multiple Korrespondenzanalyse zusammengefasst werden, da bei den Müttern die oberen Einkommenskategorien nur schwach besetzt sind und bei ihren Partnern die unteren. Ferner werden Antwortverweigerungen (ungültige Angaben) als selbständige Ausprägung (k. A.) aufgenommen, um Skalenwerte für ungültige Angaben zu erhalten. Durch diese Festlegung müssen alle

Variablen als nominalskaliert betrachtet werden. Es werden nun folgende Entscheidungen getroffen (Details werden später behandelt):

Auswahl der Variablen und Objekte: Es sollen sechs soziale Schichtungsmerkmale untersucht werden. In die Analyse sollen nur Kinder einbezogen werden, für die soziale Schichtungsmerkmale beider Elternteile vorliegen.

Datentransformationen: Die Einkommensvariablen sollen zusammengefasst werden. Bei jeder Variablen sollen ungültige Angaben als selbständige Ausprägung in die Analyse aufgenommen werden.

Zentrale Fragestellungen der Analyse sind:

- *Faktorenanalytische Fragestellung:* Reicht eine Dimension, die sich als soziale Schichtungsdimension bezeichnen lässt, zur Darstellung der Zusammenhangsstruktur der sechs sozialen Schichtungsvariablen aus?
- *Clusteranalytische Fragestellung:* Lassen sich abgegrenzte soziale Schichten (Cluster) erkennen?

Auswahl eines Ähnlichkeits- oder Unähnlichkeitsmaßes: Dieser Schritt ist nur bei der nichtmetrischen mehrdimensionalen Skalierung erforderlich. Bei der multiplen Korrespondenzanalyse muss ein bestimmtes Zusammenhangsmaß, nämlich die so genannten standardisierten Residuen, verwendet werden (siehe Abschnitt 3.2.1). Diese messen das überzufällig gemeinsame Auftreten bzw. Nichtauftreten von zwei Ausprägungen. Werden die standardisierten Residuen quadriert und mit der Fallzahl multipliziert, ergibt sich der χ^2 -Beitrag einer Zelle (siehe Abschnitt 3.3.1). Aufsummieren der Beiträge führt zum χ^2 -Wert zur Prüfung der statistischen Unabhängigkeit. In unserem Beispiel ergibt sich ein χ^2 -Wert von 8 483,90 mit 915 Freiheitsgraden. Er weicht signifikant ($p < 0,001$) von 0 ab. Die Zusammenhänge zwischen den Variablen sind »überzufällig«. Eine multiple Korrespondenzanalyse ist somit sinnvoll. Die Gefahr, dass nur »weißes Rauschen« untersucht wird, ist nicht gegeben.

Berechnen einer Zusammenhangsmatrix (siehe Abschnitt 3.2.1): Für alle Ausprägungskombinationen wird eine Zusammenhangsmatrix berechnet. In den Zellen der Zusammenhangsmatrix stehen die standardisierten Residuen.

Zahl maximal möglicher Dimensionen (siehe Abschnitt 3.2.3): Die Zahl der maximal möglichen Dimensionen bei der multiplen Korrespondenzanalyse ist gleich der Zahl der Ausprägungen aller nominalen Variablen minus der Zahl der nominalen Variablen. In unserem Beispiel sind maximal 42 Dimensionen möglich (siehe Tabelle 3.1 auf Seite 46).

Bestimmung der bedeutsamen Dimensionen (siehe Abschnitt 3.3): Die Zahl der »bedeutsamen« Dimensionen bei der multiplen Korrespondenzanalyse ist gleich der Zahl der

Variable »Einkommen der Mutter« (EinkM) und »Einkommen des Partners« (EinkP) mit den Ausprägungen:

- hat kein Einkommen (k. Eink.)
- bezieht Arbeitslosenunterstützung, Karenz usw. und zwar ... öS (k. Eink.)
- bis 4 000 öS (-4 000)
- 4 001 bis 5 500 öS (-5 500)
- 5 501 bis 7 500 öS (-7 500)
- 7 501 bis 10 000 öS (-10 000)
- 10 001 bis 15 000 öS (-15 000)
- 15 001 bis 20 000 öS (-20 000)
- 20 001 bis 25 000 öS (20 001+)
- 25 001 bis 30 000 öS (20 001+)
- 30 001 bis 35 001 (20 001+)
- weiß ich nicht (k. A.)

(Da nur ein Elternteil (Mutter oder ihr Partner) den Fragebogen ausfüllte, wurde diese Antwortkategorie für den Fall, dass das Einkommen des Partners nicht bekannt ist, aufgenommen)

Variable »abgeschlossene Schulbildung der Mutter« (SchulbM) und »abgeschlossene Schulbildung des Partners« (SchulbP) mit den Ausprägungen:

- Pflichtschule ohne Lehre (Pfl. o. L.)
- Pflichtschule mit Lehre (Pfl. m. L.)
- Berufsbildende mittlere Schule (BMS)
- Berufsbildende höhere Schule (BHS)
- Allgemeinbildende höhere Schule / Gymnasium (AHS)
- Hochschule bzw. Akademie (Uni)

Variable »derzeitige oder letzte berufliche Stellung der Mutter« (BerufM) und »derzeitige oder letzte berufliche Stellung des Partners« (BerufP) mit den Ausprägungen:

- leitender Angestellter oder höherer Beamter (leitAngBea)
- Facharbeiter, Vorarbeiter (Facharb)
- freiberufliche Tätigkeit (Freiber)
- angelernter Arbeiter, Hilfsarbeiter (angelArb)
- mittlerer Angestellter oder Beamter (mittAngBea)
- selbständiger Landwirt (selbLandw)
- einfacher Angestellter oder Beamter (einfAngBea)
- Selbständiger im Handel oder Gewerbe (Selbst)

Abb. 3.1: Erfasste Informationen zur sozialen Schicht der Eltern in einem Forschungsprojekt über die Lebensbedingungen von Kindern in Österreich (In Klammern stehen die verwendeten Abkürzungen für die Variablen und die (zusammengefassten) Ausprägungen)

Tab. 3.1: Variablenausprägungen und Dimensionszahl

nominale Variable	Ausprägungen (einschließlich k. A.)
EinkM	8
EinkP	8
SchulbM	7
SchulbP	7
BerufM	9
BerufP	9
Summe	48
Zahl der nominalen Variablen	-6
maximal mögliche Dimensionen	42

Eigenwerte, die einen Wert größer 1,0 haben. Zur Bestimmung der Zahl bedeutsamer Dimensionen müsste theoretisch eine Analyse mit einer maximalen Dimensionszahl von 42 gerechnet werden. Dies würde einen enormen Rechenaufwand bedeuten. Für das Beispiel wurde daher mit einer Startzahl von 17 Dimensionen für die Zahl bedeutsamer Dimensionen gerechnet. Für die 17. Dimension ergab sich ein Eigenwert von 1,0307. Da dieser nahe bei dem Schwellwert von 1,0 liegt, wurde auf eine Erhöhung der Dimensionszahl und auf ein erneutes Durchrechnen verzichtet.¹

3.1.1 Faktorenanalytische Interpretation

In Bezug auf die *Auswahl der erforderlichen Dimensionen* für die faktorenanalytische Interpretation kann der *Anpassungsindex für die mittleren Residuenabweichungen* (»Goodness-of-Fit Index for Residuals«, GFIR*-Index, siehe Abschnitt 3.3.2) berechnet werden. Der GFIR*-Index für eine bestimmte Dimensionszahl gibt an, wie viele Prozentpunkte des Gesamt- χ^2 -Wertes durch die entsprechende Dimensionszahl erklärt werden. Ein Wert nahe 1 bedeutet, dass der Gesamt- χ^2 -Wert durch die entsprechende Dimensionszahl erklärt wird. Für das Beispiel ergeben sich die in Tabelle 3.2 dargestellten Werte.

Für den GFIR*-Index ergibt sich für die erste Dimension ein Wert von 0,437 bzw. 43,7 Prozent. Die eindimensionale Lösung erklärt somit 43,7 Prozent des Gesamt- χ^2 -Wertes. Bei zwei Dimensionen erhöht sich der GFIR*-Index von 0,437 bzw. 43,7 Prozent auf 0,546 bzw. 54,6 Prozent. Dies bedeutet eine Verbesserung um über 10 Prozentpunkte. Der größte Anpassungsindex ergibt sich bei fünf Dimensionen. Daran anschließend nimmt

¹ Wäre er deutlich größer 1,0, dann müsste erneut – mit einer größeren Dimensionszahl – gerechnet werden.

Tab. 3.2: Modellprüfgrößen für die faktorenanalytische Interpretation

Zahl der Dimensionen	GFIR*	Zahl der Dimensionen	GFIR*	Zahl der Dimensionen	GFIR*
0	0,000	6	0,586	12	0,443
1	0,437	7	0,574	13	0,429
2	0,546	8	0,535	14	0,406
3	0,604	9	0,489	15	0,393
4	0,613	10	0,479	16	0,389
5	0,614	11	0,472	17	0,400

der Anpassungsindex ab. Die Zahl der zu interpretierenden Dimensionen kann somit aufgrund des GFIR*-Index auf ein bis fünf Dimensionen eingeschränkt werden. Es kommen somit fünf Lösungen in Betracht, die eindimensionale Lösung, die zweidimensionale Lösung usw. Die Koordinatenwerte der ersten fünf Dimensionen sind in Tabelle 3.3 auf der nächsten Seite wiedergegeben.

Inhaltliche Interpretation: Für welche Dimensionszahl man sich letztendlich entscheidet, wird von der inhaltlichen Interpretierbarkeit der Dimensionen abhängen. Betrachten wir dazu die berechneten Koordinatenwerte für die ersten fünf Dimensionen (siehe Tabelle 3.3 auf der nächsten Seite). Zur Interpretation der Dimensionen ist es zweckmäßig, hohe und niedrige Koordinatenwerte zu markieren und nur diese zur Interpretation zu verwenden. In dem Beispiel wurden Koordinatenwerte größer 1 bzw. kleiner -1 ausgewählt. Welche Koordinatenwerte man als »hoch« bezeichnet, hängt von den Ergebnissen ab. Treten Koordinatenwerte größer 2 auf, wird man Koordinatenwerte von 1 nicht mehr als »hoch« bezeichnen. Umgekehrt wird man unter Umständen bereits Koordinatenwerte größer 0,6 bzw. kleiner -0,6 als »hoch« bezeichnen, wenn die meisten Koordinatenwerte zwischen 0,5 und -0,5 liegen.

Für die erste Dimension ergibt sich folgendes Bild (siehe Tabelle 3.4 auf Seite 49): Der negative Pol ist durch Ausprägungen gekennzeichnet, die mit einer hohen sozialen Schichtzugehörigkeit assoziiert werden können. Der Gegenpol wird durch den Beruf des selbständigen Landwirtes und durch eine geringe Schulbildung des Partners gebildet. Insgesamt kann die *erste Dimension als allgemeine soziale Schichtungsdimension* interpretiert werden. Für die einzelnen Ausprägungen ergibt sich hinsichtlich der sozialen Schichtzugehörigkeit die in Tabelle 3.5 auf Seite 49 dargestellte Rangordnung.

Man sieht, dass die in der Variablen »abgeschlossene Schulbildung« enthaltene ordinale Information (Pfl. o. L. < Pfl. m. L. < BMS < BHS, AHS < Uni) auf der sozialen Schichtungsdimension abgebildet wird. Auch hinsichtlich des Einkommens (-5 500 < -7 500 < -10 000 < -15 000 < -20 000 < 20 001+) ist dies der Fall. Die Tatsache, dass die Ausprägungen »k. Eink.« und »k. A.« eine höhere soziale Position als Einkommen bis zu

Tab. 3.3: Koordinatenwerte für die ersten fünf Dimensionen

Variable	Ausprägung	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5
EinkM	k. A.	0,4987	-1,3574	-1,0971	-0,3847	-0,0766
EinkM	k. Eink.	0,2450	0,1002	0,0821	-0,0846	0,6359
EinkM	-5 500	0,2023	0,4960	-0,0433	-0,0254	-0,5563
EinkM	-7 500	-0,0182	0,4653	0,2617	-0,1458	-0,9791
EinkM	-10 000	-0,2904	0,2241	0,3351	0,5743	-0,7504
EinkM	-15 000	-0,5908	0,1401	-0,0017	-0,0534	-0,8950
EinkM	-20 000	-1,0978	-0,4743	0,1927	0,4636	0,1257
EinkM	20 001+	-1,4729	-0,8940	0,2321	0,7171	-0,2739
EinkP	k. A.	0,3663	-0,9783	-0,4977	-0,3891	0,0496
EinkP	k. Eink.	0,0808	-0,7148	0,3461	-0,3419	-0,4543
EinkP	-5 500	0,9993	-1,6349	1,5791	-0,3358	-0,2685
EinkP	-7 500	0,8458	-0,8744	2,2922	-0,8644	-0,9782
EinkP	-10 000	0,7712	0,1648	0,1028	0,9273	-0,9569
EinkP	-15 000	0,4456	0,6354	-0,0483	0,1889	-0,0556
EinkP	-20 000	-0,2840	0,5194	-0,1092	-0,3174	0,2375
EinkP	20 001+	-1,1910	-0,2266	0,1995	0,2780	0,4113
SchulbM	k. A.	0,9490	-2,5503	-2,4961	0,0165	-0,6048
SchulbM	Pfl. o. L.	0,8817	0,0164	0,3028	0,6115	0,1658
SchulbM	Pfl. m. L.	0,2617	0,3383	-0,1994	-0,2844	0,1802
SchulbM	BMS	-0,3591	0,1523	0,1112	-0,7722	-0,3232
SchulbM	BHS	-0,7391	0,1043	0,0588	-0,1022	-0,1643
SchulbM	AHS	-1,0452	-0,3282	0,0923	-0,2029	-0,3701
SchulbM	Uni	-1,5185	-0,5654	0,3490	1,1075	0,1558
SchulbP	k. A.	0,8360	-1,9683	-2,1626	0,1338	-0,6942
SchulbP	Pfl. o. L.	1,1352	-0,5913	1,0263	0,6755	-0,0799
SchulbP	Pfl. m. L.	0,3133	0,4880	-0,1787	-0,0620	0,0839
SchulbP	BMS	-0,2524	0,1057	0,3392	-0,8952	-0,5839
SchulbP	BHS	-0,8507	-0,2283	0,0981	-0,3447	0,1753
SchulbP	AHS	-0,8411	-0,0739	-0,0795	-0,8039	-0,2911
SchulbP	Uni	-1,4448	-0,5215	0,1867	0,8583	0,4891
BerufM	k. A.	0,3818	-0,3270	-0,6376	-0,1011	0,6878
BerufM	leitAngBea	-1,4458	-0,6740	0,2212	1,1057	-0,0165
BerufM	Facharb	0,2029	0,6539	-0,1144	-0,5397	0,6936
BerufM	Freiber	-0,5932	-0,3982	-0,2434	0,8514	0,4588
BerufM	angelArb	0,8466	0,4793	0,0284	1,1954	-0,4058
BerufM	mitAngBea	-0,6821	0,1989	0,0809	-0,3015	-0,2806
BerufM	selbLandw	1,3909	-1,5193	2,2254	-0,8313	0,3207
BerufM	einfAngBea	-0,0565	0,5359	-0,1195	-0,4076	-0,1482
BerufM	Selbst	-0,3643	-0,3619	-0,1362	-0,5651	-2,0879
BerufP	k. A.	0,5516	-1,4345	-1,3831	0,2312	-0,1032
BerufP	leitAngBea	-1,1587	-0,3236	0,1696	0,2185	0,6303
BerufP	Facharb	0,4283	0,7417	-0,3217	-0,0661	0,3048
BerufP	Freiber	-0,8236	-0,4559	0,1837	1,2097	-0,2257
BerufP	angelArb	0,9933	0,1226	0,3342	1,2827	-0,6801
BerufP	mitAngBea	-0,2634	0,4196	-0,0901	-0,4276	0,0932
BerufP	selbLandw	1,2893	-1,7163	2,4124	-1,2809	0,4767
BerufP	einfAngBea	0,5774	0,3637	-0,3677	0,2037	0,0659
BerufP	Selbst	-0,5700	-0,1911	0,0769	-0,5485	-1,8195

grau hinterlegt: Koordinatenwerte mit einem Absolutbetrag größer 1,0

Tab. 3.4: Negativer und positiver Pol der ersten Dimension

negativer Pol der 1. Dimension (hohe soziale Schicht)	positiver Pol der 1. Dimension (geringe soziale Schicht)
EinkM 15 001 bis 20 000 öS	
EinkM über 20 001 öS	
EinkP über 20 001 öS	
SchulbM AHS	
SchulbM Universität (Hochschule/Akademie)	SchulbP Pflichtschule ohne Lehre
SchulbP Universität (Hochschule/Akademie)	BerufM selbständige Landwirtin
BerufM leitende Angestellte/Beamtin	BerufP selbständiger Landwirt
BerufP leitender Angestellter/Beamter	

15 000 öS bei den Partnern abbilden, kann dadurch erklärt werden, dass in der Ausprägung »k. Eink.« auch Arbeitslosenunterstützungen enthalten sind, die möglicherweise über 15 000 öS liegen. Problematisch an dieser Interpretation erscheint nur die geringe soziale Stellung des selbständigen Landwirtes. Dies lässt sich aber leicht erklären: Die selbständigen Landwirte gab im Durchschnitt ein geringes Einkommen an und weisen oft nur einen geringen Schulabschluss auf (Pflichtschule ohne Lehre). Dies erklärt den geringen sozialen Status dieser Berufsgruppe. Allerdings lassen sich selbständige Landwirte und Antwortverweigerungen nur bedingt auf dieser ersten Dimension skalieren, da sie eigenständige Dimensionen bilden (siehe dazu nachfolgende Ausführungen).

Die zweite Dimension kann als Antwortverweigerungstendenz interpretiert werden. Hohe Antwortverweigerungstendenzen ergeben sich demnach für die in Tabelle 3.6 auf der nächsten Seite aufgelisteten Ausprägungen. Auf dem positiven Pol sind keine hohen Werte auffindbar. Eine besonders hohe Antwortbereitschaft liegt somit für keine der anderen Ausprägungen vor.

Tab. 3.5: Rangordnung hinsichtlich der sozialen Schichtzugehörigkeit für einzelne Ausprägungen

soziale Schicht	EinkM	EinkP	SchulbM	SchulbP	BerufM	BerufP
hoch	20 001+	20 001+	Uni	Uni	leitAngBea	leitAngBea
	-20 000	-20 000	AHS	AHS	mittAngBea	Freiber
	-15 000	k. Eink. ^{a)}	BHS	BHS	Freiber	Selbst
	-10 000	k. A.	BMS	BMS	Selbst	mittAngBea
	-7 500	-15 000	Pfl. m. L.	Pfl. m. L.	einfAngBea	Facharb
	-5 500	-10 000	Pfl. o. L.	k. A.	Facharb	k. A.
	k. Eink. ^{a)}	-7 500	k. A.	Pfl. o. L.	k. A.	einfAngBea
	k. A.	-5 500			angelArb	angelArb
niedrig					selbLandw	selbLandw

a) derzeit kein Einkommen oder Arbeitslosenunterstützung oder Karentgeldzahlung.

Tab. 3.6: Negativer Pol der zweiten Dimension

negativer Pol der 2. Dimension (hohe Antwortverweigerungstendenz)
EinkM keine Angabe
EinkP bis 5 500 öS
SchulbM keine Angabe
SchulbP keine Angabe
BerufM selbständige Landwirtin
BerufP keine Angabe
BerufP selbständiger Landwirt

Wir haben bisher die Interpretation der ein- und zweidimensionalen Lösung behandelt. Hinsichtlich der verbleibenden Dimensionen ist analog vorzugehen. Die dritte Dimension besteht aus zwei Polen (siehe Tabelle 3.7): Hier werden die selbständigen Landwirte von den Antwortverweigerungen getrennt. Der negative Pol ist durch Antwortverweigerungen in den Variablen »mütterliches Einkommen« (EinkM), »Schulbildung der Mutter« (SchulbM) und »Schulbildung des Partners« (SchulbP) gekennzeichnet. Ihm steht eine Erwerbstätigkeit der Mutter und ihres Partners als »selbständiger Landwirt« gegenüber. Die Partner der Mütter haben eine Pflichtschule ohne Lehre abgeschlossen und ein Einkommen bis zu 7 500 öS. Allerdings ist es schwer, die dritte Dimension inhaltlich als kontinuierliche Dimension mit den Polen »selbständige Landwirte« und »Antwortverweigerungen« zu interpretieren. Dies würde bedeuten, dass selbständige Landwirte besonders selten die Antwort verweigert hätten.

Die vierte und fünfte Dimension sind nur schwer inhaltlich interpretierbar. Bei einer faktorenanalytischen Interpretation wird man sich daher für die ein- oder zweidimensionale Lösung entscheiden, da hier eine Interpretation der Dimensionen als Kontinuum möglich ist. Die bisherigen Ergebnisse lassen sich wie folgt zusammenfassen: Es liegt eine allgemeine soziale Schichtungsdimension vor. Antwortverweigerungen und selbständige Landwirte lassen sich in dieses eindimensionale Schichtungskonzept nur bedingt einordnen.

Tab. 3.7: Negativer und positiver Pol der dritten Dimension

negativer Pol der 3. Dimension	positiver Pol der 3. Dimension
EinkM keine Angabe	EinkP bis 5 500 öS
SchulbM keine Angabe	EinkP 5 501 bis 7 500 öS
SchulbP keine Angabe	SchulbP Pflichtschule ohne Lehre
	BerufM selbständige Landwirtin
	BerufP selbständiger Landwirt

Tab. 3.8: Hypothesen und erwartete Korrelationsmuster

	Hypothese	erwartetes Korrelationsmuster
H ₁	Höhere soziale Schichten haben weniger finanzielle Sorgen.	Negative Korrelation zwischen der 1. Dimension (negativer Pol: hohe soziale Schicht; positiver Pol: niedrige soziale Schicht) und der Variablen »finanzielle Sorgen« mit den Ausprägungen »sehr oft« (1) bis »nie« (4).
H ₂	Höhere soziale Schichten verfügen über mehr Wohnraum.	Negative Korrelation zwischen der 1. Dimension und der Variablen »Wohnungsgröße« (Wohnungsgröße in Quadratmetern).
H ₃	Höhere soziale Schichten leben in größeren Gemeinden.	Negative Korrelation zwischen der 1. Dimension (negativer Pol: hohe soziale Schicht) mit der Gemeindegröße (1 = Land, 11 = Wien).
H ₄	Höhere Antwortverweigerungstendenz in den sozialen Schichtungsmerkmalen ist mit höherer Antwortverweigerung bei anderen Fragen verbunden.	Negative Korrelation zwischen der 2. Dimension (negativer Pol: hohe Antwortverweigerungstendenz in den sozialen Schichtungsmerkmalen) mit der Zahl der Antwortverweigerungen bei anderen Fragen.

Validitätstests für die faktorenanalytische Interpretation: Die inhaltliche Interpretierbarkeit der Dimensionen sowie die internen Modellprüfgrößen stellen ein wichtiges Entscheidungsmittel für die Auswahl der Zahl der Dimensionen dar. Für weitere Analysen ist es aber unerlässlich, inhaltliche Validitätstests zur Absicherung der inhaltlichen Interpretation durchzuführen. Entsprechend dem in Kapitel 1 dargestellten Vorgehen müssen zunächst Hypothesen formuliert werden, wie die berechneten Dimensionen mit anderen Variablen zusammenhängen. In unserem Beispiel können die in Tabelle 3.8 dargestellten Hypothesen verwendet werden. Die erwarteten Korrelationsmuster lassen sich graphisch, wie in Abbildung 3.2 auf der nächsten Seite dargestellt, verdeutlichen.

Die empirische Prüfung dieser Hypothesen besteht aus zwei Schritten:

1. Für die Personen werden die Skalenwerte der Personen in den latenten Dimensionen berechnet. Es werden also die neuen Variablen »soziale Schichtzugehörigkeit« und »Antwortverweigerungstendenz« gebildet.
2. Die neuen Variablen werden mit den Validitätskriterien korreliert.

Für unser Beispiel ergeben sich die in Tabelle 3.9 auf Seite 53 ausgewiesenen Korrelationen. Für die erste Dimension treten die erwarteten Korrelationsmuster auf. Es ergeben sich relativ starke Korrelationen mit den Variablen »Gemeindegröße« und »finanzielle Sorgen«. Mit der Wohnungsgröße ist die Korrelation dagegen etwas schwächer ausgeprägt, das Vorzeichen entspricht aber den Erwartungen. Die schwache Korrelation lässt sich dadurch erklären, dass die selbständigen Landwirte eine geringe soziale Position aufweisen, aber in großen »Wohnungen« (ihren Bauernhöfen) leben. Die positive Korrelation zwischen

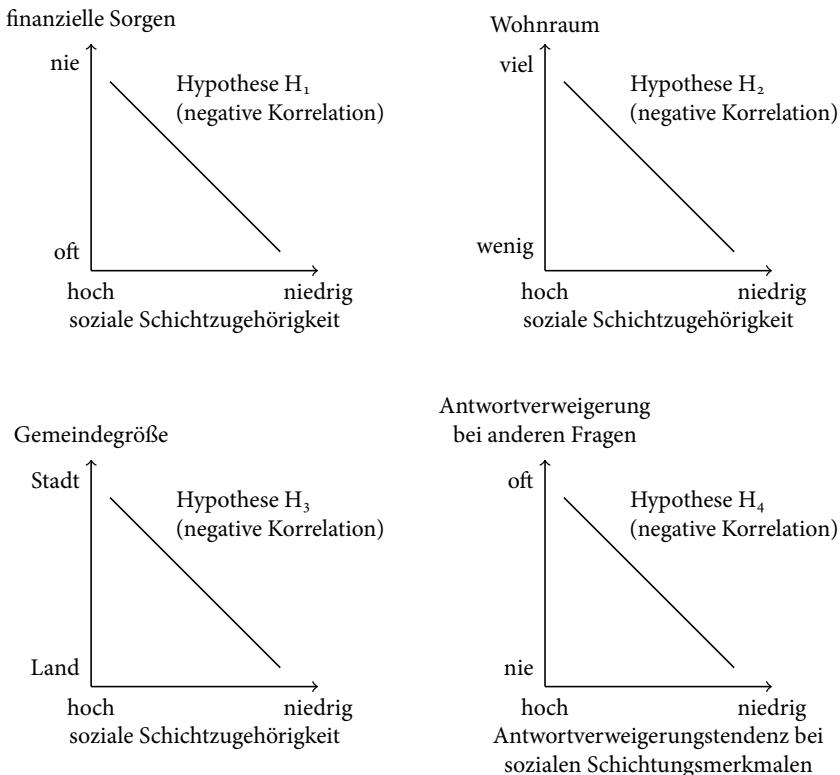


Abb. 3.2: Erwartete Korrelationsmuster

der sozialen Schichtzugehörigkeit und der Zahl der Antwortverweigerungen bedeutet, dass bei höheren Schichten in der Tendenz weniger Antwortverweigerungen auftreten. Antwortverweigerungen sind in dieser Untersuchung also ein Unterschichtphänomen. Auch hinsichtlich der zweiten Dimension stimmt das Vorzeichen der berechneten Korrelation mit dem erwarteten Vorzeichen überein. Die Korrelation ist jedoch – im Vergleich zur Gemeindegröße und den finanziellen Sorgen – nicht besonders hoch.

Aufgrund dieses ersten Validitätstests scheint eine Verwendung der beiden ersten Dimensionen für weitere Analysen gerechtfertigt. Für diese wird man die Skalenwerte der Personen in den beiden Dimensionen berechnen (siehe Abschnitt 3.2.3).

3.1.2 Clusteranalytische Interpretation

Auswahl der für eine clusteranalytische Interpretation erforderlichen Dimensionen (zu Einzelheiten siehe unten): Allgemein ist bei der multiplen Korrespondenzanalyse zu

Tab. 3.9: Koordinatenwerte für die ersten zwei Dimensionen

Kriteriumsvariable	empirische Korrelationen		erwartete Korrelationen	
	Dim. 1	Dim. 2	Dim. 1	Dim. 2
Wohnungsgröße	-0,1228	-0,1597	-(H ₂)	?
Gemeindegröße	-0,3086	0,0309	-(H ₃)	?
finanzielle Sorgen	-0,2670	-0,0388	-(H ₁)	?
Anzahl der Antwortverweigerungen	0,1108	-0,1190	?	-(H ₄)

»?«: nicht spezifiziert

beachten, dass die Dimensionszahl für eine clusteranalytische Interpretation nicht gleich jener einer faktorenanalytischen Interpretation ist. Es würde somit ein schwerwiegender Interpretationsfehler vorliegen, wenn wir aus einer guten Modellanpassung für die faktorenanalytische Interpretation auf eine gute Modellanpassung für eine clusteranalytische Interpretation schließen. Voraussetzung für eine clusteranalytische Interpretation ist eine gute Modellanpassung zwischen den empirischen und den bei einer bestimmten Dimensionszahl berechneten Distanzen. Zur Beurteilung der Modellanpassung kann der *Anpassungsindex für die Distanzabweichungen* (»Goodness-of-Fit Index for Distance«, GFID*-Index) verwendet werden (hierzu siehe ausführlicher Abschnitt 3.3). Der GFID*-Index ist analog dem GFIR*-Index zu interpretieren. Ein Wert von 1 bzw. 100 Prozent bedeutet eine perfekte Modellanpassung. Für das Beispiel ergeben sich die in Tabelle 3.10 dargestellten Größen. Es zeigt sich, dass für eine Interpretation der Distanzen eine wesentlich höhere Dimensionszahl erforderlich ist als für eine faktorenanalytische Interpretation. Erst bei zehn Dimensionen überschreitet der GFID*-Index einen Wert von 0,50 bzw. 50 Prozent.

Graphische Darstellung der Ergebnisse mit anschließender graphischer Clusterbildung, sofern dies möglich ist: Eine graphische Bestimmung von Clustern ist aufgrund der hohen

Tab. 3.10: Modellprüfgrößen für die clusteranalytische Interpretation

Zahl der Dimensionen	GFID*	Zahl der Dimensionen	GFID*	Zahl der Dimensionen	GFID*
0	0,000	6	0,335	12	0,587
1	0,067	7	0,361	13	0,614
2	0,127	8	0,403	14	0,671
3	0,216	9	0,439	15	0,696
4	0,249	10	0,516	16	0,713
5	0,277	11	0,561	17	0,732

Dimensionszahl (10) und der Zahl der Ausprägungen (48) nicht mehr möglich. Somit entfallen die *inhaltliche Interpretation* und die *Validitätsprüfung der Cluster*.

Berechnung einer neuen Datenmatrix: Da eine graphische Clusterbildung nicht möglich ist, wird die Matrix der Koordinatenwerte der ersten zehn Dimensionen als neue Datenmatrix abgespeichert.

Durchführen einer hierarchischen Clusteranalyse (siehe Abschnitt 6.5): Für die neue Datenmatrix wird eine hierarchische Clusteranalyse mit dem Weighted-Average-Linkage-Verfahren (siehe Abschnitt 9.5) und der quadrierten euklidischen Distanz durchgeführt. Die Verwendung der quadrierten euklidischen Distanz ist erforderlich, da diese als Distanzmaß der multiplen Korrespondenzanalyse definiert ist (siehe Abschnitt 3.3.4). Ohne hier auf das Detail der Analyse einzugehen, soll die dabei ermittelte hierarchische Ähnlichkeitsstruktur in Form eines sogenannten *Dendrogramms* wiedergegeben werden. In dem Dendrogramm (Abbildung 3.3 auf Seite 56) lassen sich folgende sechs Cluster erkennen:

Cluster 1: Untere Einkommensschichten. Das Einkommen des Partners liegt zwischen 5 500 und 7 500 öS.

Cluster 2: Selbständige Landwirte. Die Landwirte geben ebenfalls ein sehr geringes Einkommen an.

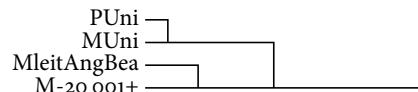
Cluster 3: Cluster der Antwortverweigerungen.

Cluster 4: Selbständige Erwerbstätige. Beide Elternteile sind selbstständig erwerbstätig.

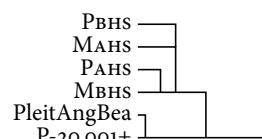
Cluster 5: Mittelschicht.

Cluster 6: Freiberuflich Erwerbstätige.

Die Grobclusterstruktur wird im Wesentlichen durch den derzeit oder zuletzt ausgeübten Beruf bestimmt. Auch innerhalb des relativ großen Clusters der Mittelschicht lassen sich spezifische, inhaltlich gut interpretierbare Ähnlichkeitsmuster feststellen. So zum Beispiel dieses:



Die Partner der Mütter und die Mütter besitzen Universitätsabschlüsse. Die Mütter sind als leitende Angestellte mit einem Einkommen von über 20 001 öS tätig. Bei diesem Muster:



haben beide Elternteile entweder eine BHS oder AHS abgeschlossen, die Partner der Mütter sind als leitende Angestellte mit einem Einkommen von über 20 001 öS tätig. Neben der 6-Clusterlösung führt auch noch eine 14-Clusterlösung, bei der die Mittelschicht ausdifferenziert wird, zu einer guten Modellanpassung (siehe Abschnitt 9.5).

Zusammenfassend können wir festhalten: Die hierarchische Clusteranalyse erbringt eine inhaltlich interpretierbare 6- und 14-Clusterlösung, die vorwiegend durch die berufliche Tätigkeit charakterisiert sind, wobei in der 6-Clusterlösung eine breite Mittelschicht vorliegt.

Soll mit einer Clusterlösung weiter gerechnet werden, ist eine Zuordnung der Kinder zu den sozialen Schichten erforderlich. Dabei könnten beispielsweise für die 6-Clusterlösung folgende hierarchische Regeln angewendet werden:

1. Übt die Mutter oder ihr Partner eine freiberuflche Tätigkeit aus, wird das Kind der sozialen Schicht der freiberuflch Erwerbstätigen zugeordnet.
2. Ist die Mutter oder ihr Partner als selbständiger Landwirt tätig, wird das Kind der sozialen Schicht der selbständigen Landwirte zugeordnet.
3. Ist die Mutter oder ihr Partner selbständig erwerbstätig, wird das Kind der sozialen Schicht der Selbständigen zugeordnet.
4. Liegt das Einkommen des Partners zwischen 5 500 und 7 500 öS, wird das Kind der sozialen Schicht der unteren Einkommensgruppe zugeordnet.
5. Liegt für die Mutter keine Angabe bezüglich der abgeschlossenen Schulbildung oder für den Partner keine Angabe bezüglich der abgeschlossenen Schulbildung oder des Berufes vor, wird die soziale Schichtzugehörigkeit als fehlender Wert (keine Angabe) behandelt.
6. Wurden die Kinder bisher noch nicht zugeordnet, findet eine Zuordnung zur Mittelschicht statt.

»Hierarchisch« bedeutet dabei, dass die Regeln von oben nach unten abgearbeitet werden. Eine einmal vorgenommene Zuordnung zu einer Schicht wird nicht mehr aufgelöst. Ist beispielsweise die Mutter als selbständige Landwirtin tätig (Regel 2) und übt ihr Partner nebenberuflch den Beruf eines angelernten Arbeiters mit einem Einkommen zwischen 5 500 und 7 500 öS aus (Regel 4), findet eine Zuordnung zur sozialen Schicht der Landwirte statt, da diese Zuordnungsmöglichkeit zuerst geprüft wird.

Insgesamt ist in dem vorgestellten Beispiel eine faktoren- und eine clusteranalytische Interpretation möglich. Die clusteranalytische Interpretation hat in dem Beispiel den Nachteil, dass eine sehr breite Mittelschicht entsteht, während die anderen sozialen Schichten nur schwach besetzt sind. Man wird sich daher bei weiterführenden Analysen für die Verwendung der beiden bei der faktorenanalytischen Interpretation gefundenen Dimensionen entscheiden, da dies ein differenzierteres Bild eröffnet, und zusätzlich für

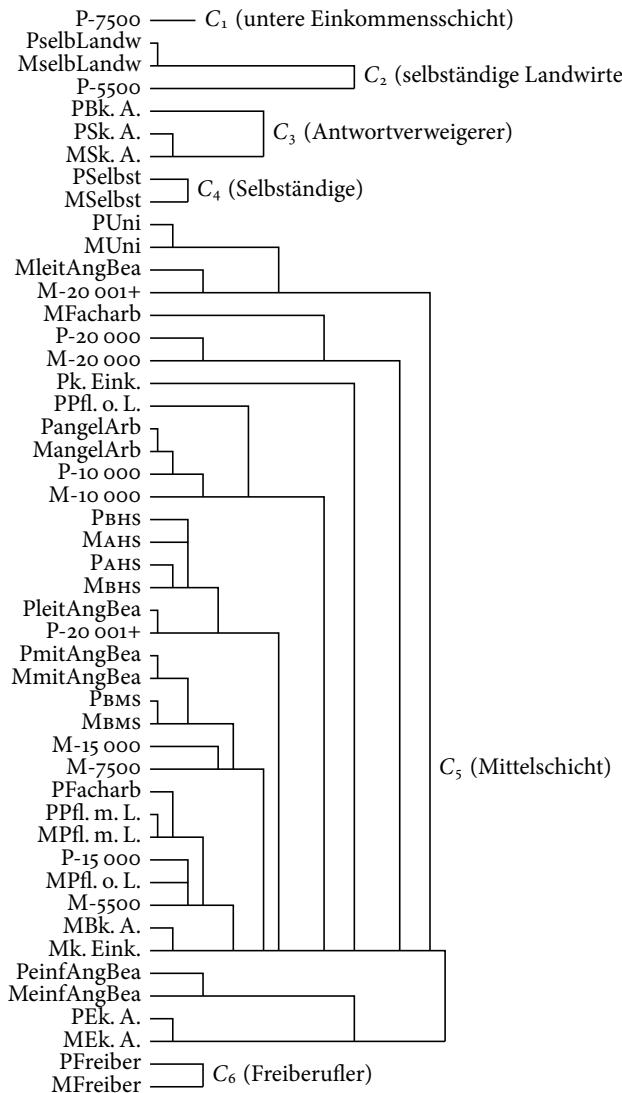


Abb. 3.3: Dendrogramm einer Clusteranalyse für die Distanzen

die Landwirte eine getrennte Analyse durchführen, da diese auch noch auf einer weiteren, dritten Dimension laden.

Tab. 3.11: Zusammenhang zwischen der abgeschlossenen Schulbildung der Mutter und des Vaters (Gesamtanteils-werte, Prozentbasis = 1 823 Fälle; $\chi^2 = 1\,055,471$, $p < 0,001$, Kontingenzkoeffizient $C_{\text{korr}} = 0,6541$)

SchulbP	SchulbM							gesamt
	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	
B ₁	0,0148	0,0126	0,0109	0,0043	0,0000	0,0011	0,0005	0,0444
B ₂	0,0038	0,0647	0,0301	0,0115	0,0027	0,0021	0,0016	0,1168
B ₃	0,0060	0,1371	0,2122	0,0658	0,0235	0,0159	0,0170	0,4777
B ₄	0,0011	0,0159	0,0235	0,0367	0,0087	0,0093	0,0065	0,1020
B ₅	0,0005	0,0076	0,0181	0,0192	0,0137	0,0142	0,0076	0,0811
B ₆	0,0005	0,0021	0,0120	0,0126	0,0065	0,0142	0,0049	0,0532
B ₇	0,0011	0,0038	0,0164	0,0186	0,0137	0,0224	0,0482	0,1245
gesamt	0,0279	0,2441	0,3236	0,1689	0,0691	0,0795	0,0866	1,0000

Abkürzungen: A = abgeschl. Schulbildung der Mutter, B = abgeschl. Schulbildung des Partners

1 = keine Angabe 5 = berufsbildende höhere Schule

2 = Pflichtschule ohne Lehre 6 = allgemeinbildende höhere Schule

3 = Pflichtschule mit Lehre 7 = Hochschule/Akademie

4 = berufsbildende mittlere Schule

Lesehilfe: Bei 1,48 Prozent (= $100 \cdot 0,0148$) der untersuchten Datensätze (100 Prozent = 1823) liegt in beiden Variablen keine Angabe (A₁ und B₁) vor. Insgesamt traten bei der Variablen SchulbM 2,79 Prozent (= $100 \cdot 0,0279$) fehlende Werte (Spaltensumme in A₁) auf. In der Variablen SchulbP waren dies 4,44 Prozent (Zeilensumme in B₁).

3.2 Das Modell der multiplen Korrespondenzanalyse

Primär besteht die Aufgabe der multiplen Korrespondenzanalyse darin, die Zusammenhangsstruktur zwischen zwei oder mehreren nominalskalierten Variablen durch eine räumliche Darstellung der Ausprägungen in einem q -dimensionalen Raum, der von quantitativen Dimensionen aufgespannt wird, darzustellen. Das *Kalkül zur Berechnung der gesuchten Koordinatenwerte $x_{i(k)h}$* besteht aus drei Schritten:

1. Berechnung einer empirischen Zusammenhangsmatrix G für die nominalen Variablen.
2. Berechnung der Eigenwerte und -vektoren der empirischen Zusammenhangsmatrix.
3. Reskalierung der Eigenvektoren; so ergeben sich die gesuchten Koordinatenwerte.

In dem in der Tabelle 3.11 wiedergegebenen Illustrationsbeispiel ist unter Verwendung der Terminologie der Tabelle 3.12 auf der nächsten Seite $n = 1\,823$ und die Zahl der Variablen $m = 2$. Wenn wir mit $i = 1$ die abgeschlossene Schulbildung des Partners und mit $j = 2$ die abgeschlossene Schulbildung der Mutter bezeichnen, ist $p_{1(k=1)} = 0,0444$,

Tab. 3.12: Notation für die multiple Korrespondenzanalyse

Symbol	Beschreibung
$i(k)$	Ausprägung k der nominalen Variablen i
$j(k^*)$	Ausprägung k^* der nominalen Variablen j
h	Index für die Dimension h
m	Zahl der in die Analyse einbezogenen nominalen Variablen
m_i	Zahl der Ausprägungen der nominalen Variablen i
M	Zahl der Ausprägungen aller nominalen Variablen
q	Zahl der Dimensionen
n	Zahl der in die Analyse einbezogenen Fälle
$x_{i(k)h}$	gesuchter Koordinatenwert der Ausprägung k der nominalen Variablen i auf der Dim. h
$x_{j(k^*)h}$	gesuchter Koordinatenwert der Ausprägung k^* der nominalen Variablen j auf der Dim. h
$p_{i(k)}$	Anteilswert (relative Auftrittshäufigkeit) der Ausprägung k der nominalen Variablen i
$p_{j(k^*)}$	Anteilswert (relative Auftrittshäufigkeit) der Ausprägung k^* der nominalen Variablen j
$p_{i(k)j(k^*)}$	Anteilswert (relative Auftrittshäufigkeit) der Ausprägung k der nominalen Variablen i und der Ausprägung k^* der nominalen Variablen j

$p_{1(k=2)} = 0,1168$ usw., $m_1 = 7$ und $m_2 = 7$ und die Zahl der Ausprägungen $M = 14$. Für $p_{1(1)2(1)}$ ergibt sich ein Wert von $0,0148$ usw.

3.2.1 Berechnung der empirischen Zusammenhangsmatrix \mathbf{G}

Die in der multiplen Korrespondenzanalyse untersuchte empirische Zusammenhangsmatrix \mathbf{G} der nominalen Variablen besitzt allgemein folgenden Aufbau:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \dots & \mathbf{G}_{1j} & \dots & \mathbf{G}_{1m} \\ \vdots & & \vdots & & \vdots \\ \mathbf{G}_{i1} & \dots & \mathbf{G}_{ij} & \dots & \mathbf{G}_{im} \\ \vdots & & \vdots & & \vdots \\ \mathbf{G}_{m1} & \dots & \mathbf{G}_{mj} & \dots & \mathbf{G}_{mm} \end{bmatrix}.$$

Die empirische Zusammenhangsmatrix ist symmetrisch und besteht aus $(m \times m)$ -Submatrizen. In der Submatrix \mathbf{G}_{ij} stehen die Zusammenhänge zwischen den Ausprägungen der nominalen Variablen i und j . Als Zusammenhangsmaß werden die standardisierten Residuen verwendet. Diese sind definiert als:²

$$g_{i(k)j(k^*)} = \frac{p_{i(k)j(k^*)} - p_{i(k)} \cdot p_{j(k^*)}}{\sqrt{p_{i(k)} \cdot p_{j(k^*)}}}. \quad (3.1)$$

² Siehe dazu die Ableitung der bivariaten Korrespondenzanalyse bei Dunn und Everitt (1982).

Für Ausprägungen von unterschiedlichen nominalen Variablen enthält die Matrix \mathbf{G} das überzufällige gemeinsame Auftreten bzw. Nichtauftreten der untersuchten Ausprägungen. Ein Wert größer 0 bedeutet, dass die beiden Ausprägungen überzufällig gemeinsam auftreten, ein Wert kleiner 0, dass die beiden Ausprägungen überzufällig nicht gemeinsam auftreten. »Überzufällig« bedeutet dabei »von der bei statistischer Unabhängigkeit erwarteten relativen Auftrittshäufigkeit abweichend«. Umso größer der Absolutwert der Größe $g_{i(k)j(k^*)}$, desto stärker ist die Abweichung von der statistischen Unabhängigkeit.

Für dieselben Ausprägungen in einer nominalen Variablen ($i = j$ und $k = k^*$) ergeben sich folgende Werte:

$$g_{i(k)i(k)} = \frac{p_{i(k)i(k)} - p_{i(k)} \cdot p_{i(k)}}{\sqrt{p_{i(k)} \cdot p_{i(k)}}} = \frac{p_{i(k)} - p_{i(k)} \cdot p_{i(k)}}{p_{i(k)}} \quad (3.2)$$

$$= 1 - p_{i(k)}, \quad (3.3)$$

da $p_{i(k)i(k)} = p_{i(k)}$.

Für unterschiedliche Ausprägungen in einer nominalen Variablen ($i = j$ und $k \neq k^*$) vereinfacht sich Gleichung 3.1 zu:

$$g_{i(k)i(k^*)} = \frac{p_{i(k)i(k^*)} - p_{i(k)} \cdot p_{i(k^*)}}{\sqrt{p_{i(k)} \cdot p_{i(k^*)}}} = \frac{0 - p_{i(k)} \cdot p_{i(k^*)}}{\sqrt{p_{i(k)} \cdot p_{i(k^*)}}} \\ = -\sqrt{p_{i(k)} \cdot p_{i(k^*)}},$$

da $p_{i(k)i(k^*)} = 0$. Da in einer nominalen Variablen nur eine Ausprägung auftreten kann, ist die gemeinsame Auftrittshäufigkeit gleich 0.

Die *Zusammenhänge innerhalb einer nominalen Variablen* ($i = j$) stellen rein *rechnerische Größen* dar, die nicht empirisch erfasst wurden. Diese Tatsache ist bei den Modelltests zu berücksichtigen (siehe dazu Abschnitt 3.3).

Für das Beispiel ergibt sich die in der Tabelle 3.13 auf der nächsten Seite dargestellte Zusammenhangsmatrix. Der Wert von 0,3848 für die Ausprägung B₁ ($k = 1$) in der Variablen SchulbP ($i = 1$) und A₁ ($k^* = 1$) in der Variablen SchulbM ($j = 2$) bedeutet, dass die fehlenden Werte überzufällig gemeinsam auftreten. Er berechnet sich entsprechend Gleichung 3.1 wie folgt:

$$g_{1(1)2(1)} = \frac{0,0148 - 0,0279 \cdot 0,0444}{\sqrt{0,0279 \cdot 0,0444}} = \frac{0,0136}{0,0352} = 0,3863,$$

da 0,0148 die relative gemeinsame Auftrittshäufigkeit der Ausprägungen B₁ und A₁ ist und die Ausprägungen A₁ bzw. B₁ Anteilsraten von 0,0279 bzw. 0,0444 besitzen. Diese Werte können der Tabelle 3.13 auf der nächsten Seite entnommen werden.³

3.2.2 Berechnung der Eigenwerte, Faktorladungen und Koordinatenwerte

Für die Matrix \mathbf{G} wird eine *Eigenwertzerlegung* (»single value decomposition«) durchgeführt:

$$\mathbf{G} = \mathbf{V} \cdot \mathbf{D} \cdot \mathbf{V}^T$$

wobei \mathbf{V} die Matrix der Eigenvektoren und \mathbf{D} die Diagonalmatrix der Eigenwerte ist. \mathbf{V}^T ist die transponierte Eigenvektormatrix. In der ersten Spalte von \mathbf{V} steht der erste Eigenvektor, in der zweiten Spalte der zweite Eigenvektor usw. Die Eigenvektoren sind auf die Länge 1 normiert. Technisch wird in diesem Schritt nichts anderes als eine *Hauptkomponentenanalyse*, also eine Art Faktorenanalyse ohne Rotation und Kommunalitäts schätzung, für die Zusammenhangsmatrix \mathbf{G} durchgeführt.

Anstelle der Eigenvektoren können auch Faktorladungen betrachtet werden. Diese erhält man dadurch, dass die Elemente der Eigenvektoren mit der Wurzel des dazugehörigen Eigenwertes multipliziert werden:

$$f_{i(k)h} = v_{i(k)h} \cdot \sqrt{d_h},$$

wobei $f_{i(k)h}$ die Faktorladung der Ausprägung k der nominalen Variablen i auf der Dimension h ist, $v_{i(k)h}$ ist der Eigenvektorwert der Ausprägung k der nominalen Variablen i auf der Dimension h und d_h der zu der Dimension h korrespondierende Eigenwert. In Matrixschreibweise lässt sich die Berechnung der Faktorladungen darstellen als:

$$\mathbf{F} = \mathbf{V} \cdot \mathbf{D}^{1/2},$$

wobei \mathbf{F} die Matrix der Faktorladungen ist.

Um die gesuchten Koordinatenwerte zu erhalten, müssen die Eigenvektoren bzw. die Faktorladungen *reskaliert* werden. Dazu werden die Faktorladungen mit dem Kehrwert der Wurzel aus den Anteilsraten der einzelnen Ausprägungen multipliziert:

$$x_{i(k)h} = f_{i(k)h} \cdot \left(\frac{1}{\sqrt{p_{i(k)}}} \right) = \frac{f_{i(k)h}}{\sqrt{p_{i(k)}}} = p_{i(k)}^{-1/2} \cdot v_{i(k)h} \cdot d^{1/2}. \quad (3.4)$$

³ Die Abweichungen für die Zelle (B₁,A₁) an der dritten und vierten Kommastelle entstehen durch Rundung.

Tab. 3.13: Zusammenhangsmatrix \mathbf{G} ; abgeschlossene Schulbildung der Mutter (SchulbM: A₁ bis A₇) und abgeschlossene Schulbildung des Partners (SchulbP: B₁ bis B₇), zu den Abkürzungen der Ausprägungen siehe Tabelle 3.11 auf Seite 57

	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
A ₁	0,9720	-0,0826	-0,0951	-0,0687	-0,0439	-0,0471	-0,0492
A ₂	-0,0826	0,7559	-0,2810	-0,2030	-0,1298	-0,1393	-0,1454
A ₃	-0,0951	-0,2810	0,6763	-0,2338	-0,1495	-0,1604	-0,1674
A ₄	-0,0687	-0,2030	-0,2338	0,8310	-0,1080	-0,1159	-0,1210
A ₅	-0,0439	-0,1298	-0,1495	-0,1080	0,9308	-0,0741	-0,0774
A ₆	-0,0471	-0,1393	-0,1604	-0,1159	-0,0741	0,9204	-0,0830
A ₇	-0,0492	-0,1454	-0,1674	-0,1210	-0,0774	-0,0830	0,9133
B ₁	0,3848	0,0170	-0,0284	-0,0359	-0,0554	-0,0409	-0,0532
B ₂	0,0099	0,2144	-0,0393	-0,0585	-0,0593	-0,0736	-0,0842
B ₃	-0,0634	0,0600	0,1466	-0,0524	-0,0519	-0,1133	-0,1199
B ₄	-0,0328	-0,0570	-0,0519	0,1486	0,0205	0,0134	-0,0240
B ₅	-0,0361	-0,0862	-0,0504	0,0468	0,1081	0,0971	0,0076
B ₆	-0,0243	-0,0947	-0,0392	0,0382	0,0479	0,1541	0,0047
B ₇	-0,0404	-0,1523	-0,1187	-0,0164	0,0550	0,1264	0,3607

(a) Teil 1, Submatrix \mathbf{G}_{11} und \mathbf{G}_{21}

	B ₁	B ₂	B ₃	B ₄	B ₅	B ₆	B ₇
B ₁	0,3848	0,0099	-0,0634	-0,0328	-0,0361	-0,0243	-0,0404
B ₂	0,0170	0,2144	0,0600	-0,0570	-0,0862	-0,0947	-0,1523
B ₃	-0,0284	-0,0393	0,1466	-0,0519	-0,0504	-0,0392	-0,1187
B ₄	-0,0359	-0,0585	-0,0524	0,1486	0,0468	0,0382	-0,0164
B ₅	-0,0554	-0,0593	-0,0519	0,0205	0,1081	0,0479	0,0550
B ₆	-0,0409	-0,0736	-0,1133	0,0134	0,0971	0,1541	0,1264
B ₇	-0,0532	-0,0842	-0,1199	-0,0240	0,0076	0,0047	0,3607
B ₁	0,9555	-0,0720	-0,1457	-0,0673	-0,0600	-0,0486	-0,0743
B ₂	-0,0720	0,8831	-0,2362	-0,1091	-0,0973	-0,0788	-0,1206
B ₃	-0,1457	-0,2362	0,5222	-0,2207	-0,1969	-0,1594	-0,2439
B ₄	-0,0673	-0,1091	-0,2207	0,8979	-0,0910	-0,0736	-0,1127
B ₅	-0,0600	-0,0973	-0,1969	-0,0910	0,9188	-0,0657	-0,1005
B ₆	-0,0486	-0,0788	-0,1594	-0,0736	-0,0657	0,9467	-0,0814
B ₇	-0,0743	-0,1206	-0,2439	-0,1127	-0,1005	-0,0814	0,8754

(b) Teil 2, Submatrix \mathbf{G}_{12} und \mathbf{G}_{22}

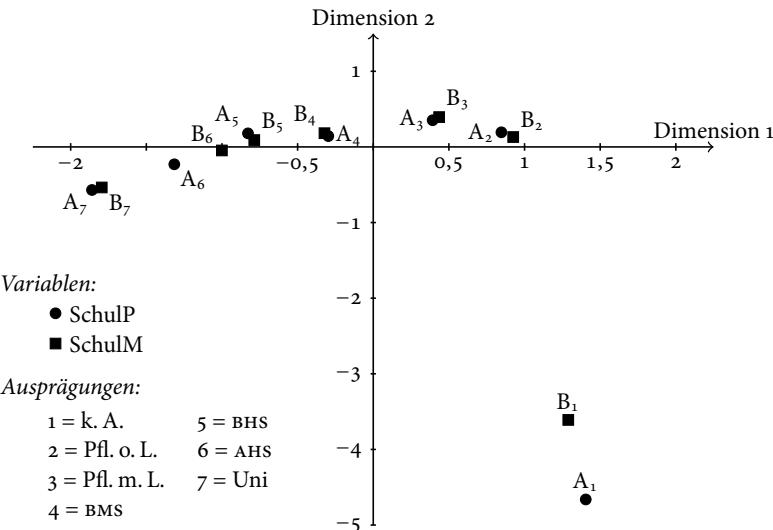


Abb. 3.4: Zweidimensionales Ergebnis der Korrespondenzanalyse für das Illustrationsbeispiel

Dabei ist $x_{i(k)h}$ der gesuchte Koordinatenwert der Ausprägung k in der nominalen Variablen i auf der Dimension h . In Matrixschreibweise lässt sich die Reskalierung darstellen als:

$$\mathbf{X} = \mathbf{H}^{-1/2} \cdot \mathbf{F} = \mathbf{H}^{-1/2} \cdot \mathbf{V} \cdot \mathbf{D}^{1/2},$$

wobei \mathbf{H} eine $(M \times M)$ -Diagonalmatrix ist. In der Diagonalen stehen die Anteilswerte der einzelnen Ausprägungen ($M = \text{Zahl der Ausprägungen insgesamt}$). In $h_{1(1)1(1)}$ steht also der Anteilswert der Ausprägung 1 in der nominalen Variablen 1 ($h_{1(1)1(1)} = p_{1(1)}$), in $h_{1(2)1(2)}$ der Anteilswert der Ausprägung 2 in der nominalen Variablen 1 ($h_{1(2)1(2)} = p_{1(2)}$) usw.

Die sich bei einer Eigenwertzerlegung der Zusammenhangsmatrix \mathbf{G} ergebenden Eigenwerte und Eigenvektoren enthält die Tabelle 3.14. Durch Multiplikation der Spalten mit der Wurzel des entsprechenden Eigenwertes ergeben sich die Faktorladungen. Für die Ausprägung A₁ ($i = 1; k = 1$) ergibt sich eine Faktorladung von $0,1881 \cdot \sqrt{1,5992} = 0,2349$ auf der ersten Dimension. Durch Multiplikation mit dem Kehrwert der Wurzel aus dem Anteilswert der Ausprägung A₁ ($p_{1(1)} = 0,0279$) erhält man den gesuchten Koordinatenwert für die Ausprägung A₁ auf der ersten Dimension: $x_{1(1)1} = 0,2349 / \sqrt{0,0279} = 1,4063$. Die Abweichungen an der dritten und vierten Kommastelle von dem in der Tabelle 3.14 angegebenen Wert entstehen durch Rundung.

Die erste Dimension kann wie in dem Anwendungsbeispiel des Abschnitts 3.1 als soziale Schichtungsdimension interpretiert werden. Ein negativer Wert bedeutet eine hohe

Tab. 3.14: Eigenvektoren, Faktorladungen und Koordinatenwerte für das Illustrationsbeispiel

Anteils-werte	Eigenvektoren		Faktorladungen		Koordinatenwerte	
	Dim. 1	Dim. 2	Dim. 1	Dim. 2	Dim. 1	Dim. 2
<i>Schulbildung Mutter (SchulbM)</i>						
A ₁	0,0279	0,1881	-0,6611	0,2349	-0,7796	1,4045
A ₂	0,2441	0,3351	0,0810	0,4184	0,0955	0,8470
A ₃	0,3236	0,1789	0,1702	0,2234	0,2007	0,3927
A ₄	0,1689	-0,0977	0,0499	-0,1221	0,0588	-0,2970
A ₅	0,0691	-0,1743	0,0399	-0,2176	0,0471	-0,8280
A ₆	0,0795	-0,2969	-0,0549	-0,3708	-0,0647	-1,3149
A ₇	0,0866	-0,4383	-0,1421	-0,5473	-0,1676	-1,8591
<i>Schulbildung Partner (SchulbP)</i>						
B ₁	0,0444	0,2176	-0,6452	0,2717	-0,7609	1,2892
B ₂	0,1168	0,2531	0,0379	0,3161	0,0446	0,9249
B ₃	0,4777	0,2411	0,2312	0,3011	0,2727	0,4356
B ₄	0,1020	-0,0825	0,0491	-0,1030	0,0579	-0,3226
B ₅	0,0811	-0,1794	0,0219	-0,2241	0,0258	-0,7865
B ₆	0,0532	-0,1847	-0,0091	-0,2307	-0,0108	-1,0003
B ₇	0,1245	-0,4383	-0,1421	-0,6332	-0,1892	-1,7946
Eigenwerte	1,5592	1,39087				

Abkürzungen: siehe Tabelle 3.11 auf Seite 57

Schichtzugehörigkeit, ein positiver Wert eine geringe Schichtzugehörigkeit (die Polung der berechneten Dimensionen ist willkürlich und kann umgedreht werden). Die zweite Dimension wird ausschließlich durch die Antwortverweigerungen gebildet. Die Antwortverweigerungen lassen sich nicht auf der sozialen Schichtungsdimension einordnen. Auch in dieser Hinsicht stimmt das Ergebnis mit dem Anwendungsbeispiel des Abschnitts 3.1 überein. Da die Objektmenge nicht groß ist und eine zweidimensionale Lösung berechnet wurde, ist eine graphische Darstellung der Koordinaten möglich (siehe Abbildung 3.4). Gut erkennbar ist, dass die Antwortverweigerungen (A₁, B₁) die zweite Dimension bilden.

3.2.3 Berechnung der Skalenwerte und Interpretation der Koordinaten

Die Reskalierung führt dazu, dass sich die *Koordinatenwerte als Mittelwerte der untersuchten Ausprägungen auf den Dimensionen* interpretieren lassen. Der Koordinatenwert $x_{i(k)h}$ der Ausprägung k der nominalen Variablen i auf der Dimension h ist gleich dem

Tab. 3.15: Dummy-Auflösung der Schulbildungsvariablen

Schulbildung der Mutter							
Ausprägung	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
Dummies	$a_{1(1)}$	$a_{1(2)}$	$a_{1(3)}$	$a_{1(4)}$	$a_{1(5)}$	$a_{1(6)}$	$a_{1(7)}$
Wert	0	0	0	0	0	0	1
Schulbildung des Partners							
Ausprägung	B ₁	B ₂	B ₃	B ₄	B ₅	B ₆	B ₇
Dummies	$a_{2(1)}$	$a_{2(2)}$	$a_{2(3)}$	$a_{2(4)}$	$a_{2(5)}$	$a_{2(6)}$	$a_{2(7)}$
Wert	0	0	0	0	0	0	1

Mittelwert $\bar{y}_{i(k)h}$ der Personen bzw. allgemein der Objekte mit der Ausprägung k in der nominalen Variablen i auf der Dimension h . Es gilt also:

$$\bar{y}_{i(k)h} = x_{i(k)h} .$$

Wir können uns also die berechneten Dimensionen als latente (unbekannte) quantitative Variablen h (mit Mittelwerten von 0 und Standardabweichungen von 1) vorstellen, auf denen jede Person g einen bestimmten Skalenwert y_{gh} besitzt. Wird nun für die Personen g mit der Ausprägung k in der nominalen Variablen i der Mittelwert

$$\bar{y}_{i(k)h} = \frac{1}{n_{i(k)}} \cdot \sum_{g \in i(k)} y_{gh}$$

in der latenten Variablen h berechnet, so ist dieser gleich dem Koordinatenwert $x_{i(k)h}$ der Ausprägung k der nominalen Variablen i auf der Dimension h .

Der Skalenwert einer Person g auf der Dimension h wird wie folgt berechnet:

$$y_{gh} = \frac{\sum_i \sum_k a_{gi(k)} \cdot x_{i(k)h}}{d_h} ,$$

wobei $a_{gi(k)}$ der Wert der Person g in der zu der Ausprägung k in der Variablen i gehörenden Dummy-Variablen $a_{i(k)}$ ist. Den Wert 1 nimmt a also an, wenn die Person g in der Variablen i die Ausprägung k besitzt, ansonsten den Wert 0. Haben beide Elternteile einen Universitätsabschluss, ergibt sich die in Tabelle 3.15 dargestellte Dummy-Auflösung. Der Skalenwert auf der sozialen Schichtungsdimension ist daher gleich:

$$y = (0 \cdot 1,4045 + \dots + 1 \cdot (-1,8591) + 0 \cdot 1,2892 + \dots + 1 \cdot (-0,5361)) / 1,5592 = 1,1976.$$

In Matrixschreibweise lässt sich die Berechnung der Skalenwerte wie folgt darstellen:

$$\mathbf{Y} = \mathbf{A} \cdot \underbrace{\mathbf{X}}_{=\mathbf{H}^{-1/2} \cdot \mathbf{F}} \cdot \mathbf{D}^{-1} = \mathbf{A} \cdot \mathbf{H}^{1/2} \cdot \underbrace{\mathbf{F}}_{=\mathbf{V} \cdot \mathbf{D}^{1/2}} \cdot \mathbf{D}^{-1} = \mathbf{A} \cdot \mathbf{H}^{-1/2} \cdot \mathbf{V} \cdot \mathbf{D}^{-1/2}$$

wobei \mathbf{Y} die Matrix der Skalenwerte ist. Sie ist von der Ordnung $n \times q$ ($n =$ Zahl der Personen, $q =$ Zahl der Dimensionen). In der ersten Spalte stehen die Skalenwerte der Personen in der ersten Dimension, in der zweiten Spalte die Skalenwerte der Personen in der zweiten Dimension. Die Matrix \mathbf{A} ist die $(n \times M)$ -Datenmatrix der Dummies. In der ersten Zeile stehen die Ausprägungen der Person 1 in den Dummy-Variablen, in der zweiten Zeile stehen die Ausprägungen der Person 2 in den Dummy-Variablen usw. Ein Beweis findet sich bei Bacher (1995a).

3.2.4 Unerwünschter Effekt der Reskalierung der Faktorladungen und Rotation der Faktoren

Die Reskalierung der Faktorladungen führt zu einer eleganten und einfachen Interpretation der Ergebnisse der multiplen Korrespondenzanalyse. Sie hat aber den Nachteil, dass in die Berechnung der Koordinatenwerte die Anteilswerte der Ausprägungen eingehen. Kleinere Anteilswerte, also schwach besetzte Ausprägungen, führen bei gleichen Faktorladungen zu höheren Koordinatenwerten. Insbesondere sehr kleine Anteilswerte (kleiner 5 Prozent) haben einen sehr starken Einfluss auf die Ergebnisse. Bei der Namensgebung der Dimensionen ist daher darauf zu achten, dass ein sehr hoher oder ein sehr niedriger Koordinatenwert nicht ausschließlich durch die Reskalierung entsteht. Ob dies der Fall ist, kann durch eine Inspektion der Varianzbeiträge der einzelnen Ausprägungen beantwortet werden. Hat eine Ausprägung einen hohen Koordinatenwert (zum Beispiel 4,5), aber nur einen geringen Varianzbeitrag (zum Beispiel 0,04), sollte sie nicht zur Namensgebung verwendet werden. Der Varianzbeitrag $v_{i(k)h}^2$ einer Ausprägung k der Variablen i auf der Dimension h ist das Quadrat der entsprechenden Faktorladung $f_{i(k)h}$:

$$v_{i(k)h}^2 = f_{i(k)h}^2 \cdot$$

Schwellenwerte für den Varianzbeitrag existieren nicht. Als Mindestgröße kann aber ein Wert von 0,09 (9 Prozent) angesehen werden. Dies entspricht einer Faktorladung von 0,30, und eine geringere Faktorladung wird niemand mehr als brauchbar betrachten. Für das Illustrationsbeispiel der Tabelle 3.14 auf Seite 62 ergeben sich die in Tabelle 3.16 auf der nächsten Seite dargestellten Varianzbeiträge.

Zur Namensgebung auf der ersten Dimension sind somit die Koordinatenwerte der Ausprägungen A₁ und B₁ (keine Angaben) nur bedingt geeignet. Hierzu wird man die Ausprägungen A₆ (SchulbM = AHS), A₇ (SchulbM = Uni), B₂ (SchulbP = Pfl. o. L.), B₆ (SchulbP = AHS) und B₇ (SchulbP = Uni) verwenden. Da B₂ einen positiven Koordinatenwert besitzt und die anderen Ausprägungen einen negativen Wert, wird man die erste Dimension als soziale Schichtungsdimension interpretieren. Betrachtet man alle

Tab. 3.16: Varianzbeiträge der Ausprägungen zu den Dimensionen

	Koordinatenwerte		Varianzbeiträge	
	Dim. 1	Dim. 2	Dim. 1	Dim. 2
<i>Schulbildung Mutter (SchulbM)</i>				
A ₁	1,4045	-4,6615	0,0551	0,6079
A ₂	0,8470	0,1934	0,1751	0,0091
A ₃	0,3927	0,3529	0,0499	0,0403
A ₄	-0,2970	0,1432	0,0149	0,0034
A ₅	-0,8280	0,1792	0,0473	0,0022
A ₆	-1,3149	-0,2297	0,1375	0,0042
A ₇	-1,8591	-0,5693	0,2995	0,0281
<i>Schulbildung Partner (SchulbP)</i>				
B ₁	1,2892	-3,6100	0,0738	0,5790
B ₂	0,9249	0,1307	0,0999	0,0020
B ₃	0,4356	0,3945	0,0906	0,0743
B ₄	-0,3226	0,1815	0,0106	0,0033
B ₅	-0,7865	0,0907	0,0502	0,0006
B ₆	-1,0003	-0,0468	0,0532	0,0001
B ₇	-1,7946	-0,5361	0,4010	0,0358
Summe			1,5592	1,3907

grau hinterlegt: Koordinatenwerte mit einem Absolutbetrag größer 1

Abkürzungen: siehe Tabelle 3.11 auf Seite 57

Koordinatenwerte der Ausprägungen A₂ bis A₇ und B₂ bis B₇, so wird – wie in dem Anwendungsbeispiel des Abschnitts 3.1 – die ordinale Information der Ausprägungen (A₂ < A₃ < A₄ < A₅, A₆ < A₇, B₂ < B₃ < B₄ < B₅, B₆ < B₇) abgebildet. Zur Namensgebung der zweiten Dimension eignen sich die Ausprägungen A₁ und B₁ (keine Angaben). Sie besitzen hohe Koordinatenwerte und ihre Varianzbeiträge sind größer 9 Prozent.

Im Unterschied zur Faktorenanalyse ist eine *Rotation der Faktoren* (Dimensionen) problematisch. Eine Rotation hat nämlich den Effekt, dass sich die rotierten Koordinatenwerte nicht mehr als Mittelwerte der Ausprägungen auf den (rotierten) latenten Dimensionen interpretieren lassen.⁴ Daher sollte bei einer faktorenanalytischen Interpretation auf eine Rotation verzichtet werden, da die Interpretation der rotierten Koordinatenwerte ungeklärt ist.

⁴ Eine – in der Forschungspraxis aber so gut wie nie auftretende – Ausnahme besteht allerdings, wenn alle Eigenvektoren gleich sind und orthogonal rotiert wird.

3.3 Modellprüfgrößen

In diesem Abschnitt werden die *Modellprüfgrößen für die multiple Korrespondenzanalyse* vorgestellt. Dabei wird zunächst auf die *Signifikanz der Zusammenhangsstruktur* und die *Dimensionszahl* eingegangen. Darauf folgen Ausführungen zur *Überprüfung einer faktoren- und clusteranalytischen Interpretation* sowie zu *Schwellenwerten für eine Interpretation*.

3.3.1 Signifikanz der Zusammenhangsstruktur

Bevor die multiple Korrespondenzanalyse angewendet wird, sollte untersucht werden, ob überhaupt eine *überzufällige Zusammenhangsstruktur* vorliegt und somit eine multiple Korrespondenzanalyse sinnvoll ist. Dazu kann die Beziehung der Elemente der empirischen Zusammenhangsmatrix \mathbf{G} zu dem aus der Tabellenanalyse bekannten χ^2 -Test verwendet werden. Werden die Elemente g der Zusammenhangsmatrix \mathbf{G} quadriert und mit der Fallzahl multipliziert, ergibt sich der χ^2 -Beitrag der Zelle $(i(k), j(k^*))$:

$$\begin{aligned}\chi_{i(k)j(k^*)}^2 &= n \cdot g_{i(k)j(k^*)}^2 = n \cdot \left(\frac{p_{i(k)j(k^*)} - p_{i(k)} \cdot p_{j(k^*)}}{\sqrt{p_{i(k)} \cdot p_{j(k^*)}}} \right)^2 \\ &= n \cdot \frac{(p_{i(k)j(k^*)} - p_{i(k)} \cdot p_{j(k^*)})^2}{p_{i(k)} \cdot p_{j(k^*)}}.\end{aligned}$$

Aus der empirischen Zusammenhangsmatrix \mathbf{G} kann daher eine χ^2 -Prüfgröße berechnet werden, mit

$$\chi^2 = \sum_i \sum_{j>i} \sum_{k,k^*} \chi_{i(k)j(k^*)}^2.$$

In die Berechnung werden nur die Ausprägungen unterschiedlicher nominaler Variablen einbezogen, da die berechneten Zusammenhänge zwischen den Ausprägungen innerhalb einer nominalen Variablen rein rechnerische Größen darstellen (siehe dazu Abschnitt 3.2).

Der χ^2 -Wert prüft die paarweise Unabhängigkeit der Variablen. Bei drei nominalen Variablen V_1 , V_2 und V_3 beispielsweise wird also geprüft, ob insgesamt die Zusammenhänge zwischen V_1 und V_2 , zwischen V_1 und V_3 sowie zwischen V_2 und V_3 unabhängig sind. Für den Fall, dass nur zwei nominale Variablen in die Analyse einbezogen werden, kann die Zahl der Freiheitsgrade df nach der aus der Tabellenanalyse bekannten Formel berechnet werden, mit

$$df = (\text{Zeilenzahl} - 1) \cdot (\text{Spaltenzahl} - 1) = (m_1 - 1) \cdot (m_2 - 1).$$

Bei mehr als zwei nominalen Variablen berechnet sich die Zahl der Freiheitsgrade wie folgt:

$$df = \left(\sum_i \sum_{j>i} m_i \cdot m_j - \sum_i (m_i - 1) \right) - 1.$$

Für unser Beispiel ergibt sich ein χ^2 -Wert von 1 055,471 mit 36 Freiheitsgraden. Das Fehlerniveau p ist kleiner 0,001. Es liegen somit »überzufällige« Zusammenhänge vor und der Einsatz der multiplen Korrespondenzanalyse ist sinnvoll. Wie beim χ^2 -Test in der Tabellenanalyse ist auch in der multiplen Korrespondenzanalyse bei der Interpretation Vorsicht angebracht: Bei großen Stichproben – wie in unserem Fall – liefert der χ^2 -Test beinahe immer signifikante Ergebnisse, bei kleinen Stichproben dagegen werden kaum signifikante Ergebnisse ausgewiesen.

Im Prinzip lässt sich der χ^2 -Test zur Prüfung der Signifikanz einer bestimmten Dimensionszahl verallgemeinern (siehe dazu für die bivariate Korrespondenzanalyse Andersen 1991, S. 374–376). Bei der multiplen Korrespondenzanalyse tritt allerdings häufig das Problem auf, dass erwartete Häufigkeiten kleiner 0 auftreten. Diese müssen dann aus der Berechnung eliminiert werden, so dass nicht alle Zellen in die Berechnung eingehen. Hinzu kommt, dass auch für diesen χ^2 -Test die Abhängigkeit von der Stichprobengröße gilt. Aus diesen Gründen soll auf eine Darstellung des χ^2 -Tests für eine bestimmte Dimensionszahl verzichtet werden; er ist allerdings in dem Programm Paket ALMO enthalten (Holm 1993).

3.3.2 Die Zahl maximal möglicher und bedeutsamer Dimensionen

Aus der Faktorenanalyse wissen wir, dass die maximale Dimensionszahl bei der Verwendung der Hauptkomponentenanalyse als Schätzverfahren für die Anzahl der Faktoren gleich dem Rang der untersuchten Matrix ist (Holm 1976, S. 70–71). Der Rang der empirischen Zusammenhangsmatrix G ist gleich »Zahl der Ausprägungen minus Zahl der nominalen Variablen« (Greenacre 1989, S. 139). Wie bei der Hauptkomponenten- oder der Faktorenanalyse ist es nicht sinnvoll, alle möglichen Faktoren zu betrachten. Nach Greenacre (1989, S. 145) sind nur jene Faktoren (Dimensionen) »bedeutsam«, deren mittlerer Eigenwert größer $1/m$ ist. Der mittlere Eigenwert ist der berechnete Eigenwert dividiert durch die Zahl der nominalen Variablen (m). Dieses Abbruchskriterium ist gleich dem *Kaiserkriterium*, das bei der Hauptkomponentenanalyse verwendet wird (Arninger 1979, S. 37–38; Holm 1976, S. 70–71). Entsprechend dem Kaiserkriterium werden jene Faktoren ausgewählt, deren Eigenwerte größer 1 sind. Es gilt somit: Die Zahl der bedeutsamen Dimensionen (Faktoren) lässt sich mit dem Kaiserkriterium bestimmen.

Das Kaiserkriterium $d_h > 1$ ist gleich dem von Greenacre (1989, S. 145) angeführten Kriterium: $d_h/m > 1/m$. Für das Beispiel der Tabelle 3.11 auf Seite 57 ergeben sich folgende Werte:

maximal mögliche Eigenwerte:	$7 + 7 - 2 = 12$
Größe der Eigenwerte:	1,5592, 1,3907, 1,2567, 1,1713, 1,1236, 1,0561, 0,9439, 0,8764, 0,8287, 0,7433, 0,6093, 0,4408
Eigenwerte größer 1:	6

3.3.3 Überprüfung der faktorenanalytischen Interpretation

Im Unterschied zur Faktorenanalyse gibt die durch die Dimension h erklärte Varianz

$$\nu_h^2 = \frac{d_h}{M - m}$$

ein zu »pessimistisches« Bild über die Modellanpassung (Andersen 1991, S. 390; Greenacre 1989, S. 145; Lebart, Morineau u. a. 1984, S. 175–176). Ein Grund ist, dass im Unterschied zur bivariaten Korrespondenzanalyse keine eindeutige Beziehung zur so genannten *Trägheit* (»inertia«) besteht.⁵ Hinzu kommt, dass in der Diagonalen – im Unterschied zur Faktorenanalyse – nicht die Varianzen der untersuchten Variablen stehen, sondern rein rechnerische Größen. Zur Beurteilung der Brauchbarkeit der faktorenanalytischen Interpretation empfehlen wir daher die Verwendung des *Anpassungsindex für die mittleren Residuenabweichungen* (»Goodness-of-Fit Index for Residuals«, GFIR^{*}). Dieser ist definiert als:

$$\text{GFIR}^*(q) = 1 - \frac{(\text{RMR}^*(q))^2}{(\text{RMR}^*(0))^2}.$$

wobei $\text{RMR}^*(q)$ die *Wurzel aus der mittleren Residuenquadratsumme* (»Root-Mean-Square-Residuals«) bei q Dimensionen ist (Bollen 1989, S. 257; Jöreskog und Sörbom 1984, I.41). Sie ist wie folgt definiert:

$$\text{RMR}^*(q) = \left[\frac{1}{p} \sum_i \sum_{j>i} \sum_{k,k^*} r(q)_{i(k)j(k^*)}^2 \right]^{1/2}, \quad (3.5)$$

⁵ Die Trägheit ist in der bivariaten Korrespondenzanalyse definiert als »Gesamt- χ^2 -Wert dividiert durch die Fallzahl« (Greenacre 1989, S. 86–87). Es lässt sich zeigen, dass bei der bivariaten Korrespondenzanalyse die Summe der Eigenwerte gleich der Trägheit ist (Greenacre 1989, S. 91). Die erklärte Varianz eines Faktors lässt sich somit als χ^2 -Beitrag des Faktors zum Gesamt- χ^2 -Wert interpretieren. Bei der multiplen Korrespondenzanalyse ist dies nicht der Fall. In diesem Abschnitt wird ein Anpassungsindex GFIR^{*} für die mittlere Residuenabweichung definiert, der eine analoge Interpretation für die multiple Korrespondenzanalyse erlaubt.

wobei $r(q)_{i(k)j(k^*)} = (g_{i(k)j(k^*)} - \sum_{h=0}^q f_{i(k)h} \cdot f_{j(k)h})$ die Differenz zwischen dem Element $g_{i(k)j(k^*)}$ der empirischen Zusammhangsmatrix \mathbf{G} und dem bei q erwarteten Zusammenhang $\sum_{h=0}^q f_{i(k)h} \cdot f_{j(k)h}$ ist. Die Zahl der in die Summenberechnung einbezogenen Elemente ist p . (Die Faktorladungen $f_{i(k)0}$ und $f_{j(k^*)0}$ für die Dimension 0 sind gleich 0). Summiert wird über alle Ausprägungen zwischen den Variablen. Die sich rein aufgrund des Kalküls ergebenden Zusammenhänge innerhalb der nominalen Variablen werden nicht berücksichtigt. Dies soll der hochgestellte Stern »*« zum Ausdruck bringen.

Die Wurzel der mittleren Residuenquadratabweichung wird in der (konfirmatorischen) Faktorenanalyse als Modellanpassungsgröße verwendet (Bollen 1989, S. 257; Jöreskog und Sörbom 1984, I.41). Sie lässt sich bei der multiplen Korrespondenzanalyse als erklärter χ^2 -Beitrag interpretieren, da die Größe $\text{RMR}^*(o)$ im Nenner gleich dem mittleren χ^2 -Wert ist:

$$\text{RMR}^*(o)^2 = \frac{1}{p} \cdot \underbrace{\sum_i \sum_{j>i} \sum_{k,l} g_{i(l)j(k)}^2}_{\chi^2/n} = \frac{\chi^2}{p \cdot n} .$$

Im Nenner steht also der durchschnittliche χ^2 -Beitrag für eine Zelle bezogen auf einen Fall. Diese Größe ist – abgesehen von der Skalierungskonstante p^* – identisch mit dem zentralen Konzept der Trägheit in der bivariaten Korrespondenzanalyse.

In unserem Beispiel der Tabelle 3.11 auf Seite 57 ergibt sich bei o Dimensionen eine mittlere Residuenabweichung von

$$\text{RMR}^*(o) = \frac{\chi^2}{p \cdot n} = \frac{1055,471}{49 \cdot 1823} = 0,0118 .$$

Für die erste Dimension ergeben sich die in Tabelle 3.17 wiedergegebenen Residuen zwischen den Ausprägungen unterschiedlicher nominaler Variablen. Ein hoher positiver Wert bedeutet, dass der empirische Zusammenhangswert durch den erwarteten Zusammenhangswert unterschätzt wird. Ein hoher negativer Wert bedeutet dagegen eine Überschätzung des empirischen Zusammenhangs. Ein hoher positiver Wert liegt für die Ausprägungen A₁ und B₁ vor, also für die Fälle mit einer Antwortverweigerung. Dies ist leicht zu erklären, da beide Ausprägungen den zweiten Faktor bilden. Eine Überschätzung der empirischen Zusammenhänge (Residuen größer 0,1) liegt für folgende Ausprägungspaare vor: (A₁,B₃), (A₃,B₂), (A₆,B₇), (A₇,B₅) und (A₇,B₆).

Tab. 3.17: Residuenmatrix für das Illustrationsbeispiel

SchulbP	SchulbM						
	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
B ₁	0,3209	-0,0967	-0,0891	-0,0028	0,0037	0,0597	0,0955
B ₂	-0,0642	0,0820	-0,1099	-0,0199	0,0094	0,0436	0,0887
B ₃	-0,1341	-0,0659	0,0793	-0,0156	0,0136	-0,0016	0,0448
B ₄	-0,0086	-0,0138	-0,0288	0,1360	-0,0019	-0,0247	-0,0804
B ₅	0,0165	0,0075	-0,0003	0,0194	0,0593	0,0140	-0,1149
B ₆	0,0298	0,0018	0,0122	0,0100	-0,0023	0,0686	-0,1215
B ₇	0,1083	0,1126	0,0227	-0,0937	-0,0828	-0,1083	0,0141

Abkürzungen: siehe Tabelle 3.11 auf Seite 57

Die Berechnung der Werte soll exemplarisch für die Zelle (A₁,B₁) dargestellt werden. Der bei einer Dimension erwartete Zusammenhang berechnet sich mit (siehe Ausführungen zur Formel 3.5 auf Seite 69):

$$\hat{g}(q=1)_{i(k)j(k^*)} = \sum_{h=0}^1 f_{i(k)h} \cdot f_{j(k)h} = f_{i(k)0} \cdot f_{j(k)0} + f_{i(k)1} \cdot f_{j(k)1} .$$

Da $i = 1$, $k = 1$, $j = 2$ und $k^* = 1$, ergibt sich ein Wert von

$$\hat{g}(q=1)_{1(1)2(1)} = 0 \cdot 0 + 0,2349 \cdot 0,2717 = 0,0638$$

für den erwarteten Zusammenhang, wenn die Faktorladungen der Tabelle 3.14 auf Seite 62 für A₁ und B₁ in der ersten Dimension eingesetzt werden. Der empirische Zusammenhang $g_{1(1)2(1)}$ ist entsprechend Tabelle 3.13 auf Seite 60 gleich 0,3848. Die Abweichung zwischen empirischem und bei einer Dimension erwartetem Zusammenhang ist somit gleich $0,3848 - 0,0638 = 0,3210$.

Werden alle Elemente der Residuenmatrix der Tabelle 3.17 quadriert und addiert, ergibt sich ein Wert von 0,3136. Insgesamt gehen $p = 49$ Zellen in die Berechnung ein. Die mitt-

Tab. 3.18: Anpassungsindizes für das Illustrationsbeispiel

	Zahl der Dimensionen						
	0	1	2	3	4	5	6
erklärte Varianz	0,000	0,130	0,246	0,351	0,449	0,543	0,631
GFIR*	0,000	0,456	0,559	0,439	0,189	-0,116	-0,496

lere quadrierte Residuenabweichung ist daher gleich 0,0064. Für den Anpassungsindex $\text{GFIR}^*(1)$ für die erste Dimension ergibt sich somit ein Wert von

$$\text{GFIR}^*(1) = 1 - \frac{0,0064^{1/2}}{0,0118} = 0,4576.$$

Die erste Dimension erklärt also ungefähr 46 Prozent des Gesamt- χ^2 -Wertes. Für alle sechs bedeutsamen Dimensionen ergeben sich die in Tabelle 3.18 auf der vorherigen Seite dargestellten Werte. Es ist ersichtlich, dass die beste Modellanpassung bei zwei Dimensionen erreicht wird. Die zweidimensionale Lösung erklärt 55,9 Prozent des χ^2 -Wertes (Tabellenwert multipliziert mit 100). Die Modellanpassung würden wir als befriedigend bezeichnen (siehe Tabelle 3.20 auf Seite 74). Bei fünf und sechs Dimensionen ergibt sich sogar ein negativer Anpassungsindex. Eine faktorenanalytische Interpretation von fünf oder sechs Dimensionen ist somit auf keinen Fall zulässig. Für die weitere Analyse wird man sich aufgrund der Anpassungsindizes für die ein-, zwei- oder dreidimensionale Lösung entscheiden. Aus formalen Gesichtspunkten (höchster GFIR^* -Index) ist die zweidimensionale Lösung zu bevorzugen. Aus Vergleichsgründen wurde auch die durch q Dimensionen erklärte Varianz, wie sie bei der Faktorenanalyse berechnet wird, in die Tabelle aufgenommen. Für die ersten drei Dimensionen ist der bereits erwähnte Effekt der »Unterschätzung« der Modellanpassung ersichtlich. Während die zweidimensionale Lösung 55,9 Prozent des χ^2 -Wertes bzw. der »inertia« erklärt, beträgt der durch zwei Dimensionen erklärte Varianzanteil nur 24,6 Prozent (Tabellenwert multipliziert mit 100).

3.3.4 Modellprüfgrößen für die clusteranalytische Interpretation

Grundlage der clusteranalytischen Interpretation sind die quadrierten euklidischen Distanzen zwischen den Ausprägungen. Für die clusteranalytische Interpretation ist es daher naheliegend, zunächst zu prüfen, wie gut die für eine bestimmte Dimensionszahl berechneten Distanzen mit den empirischen Distanzen oder Ähnlichkeiten übereinstimmen.

Zur Beantwortung der Fragestellung, wie gut die für q Dimensionen berechneten Distanzen mit den empirischen Distanzen übereinstimmen, kann nach der Logik von GFIR^* ein *Anpassungsindex für die Distanzen* konstruiert werden (»Goodness-of-Fit Index for Distances«, GFID^*). Der Stern »*« bedeutet, dass in die Berechnung nur die Ausprä-

gungen unterschiedlicher nominaler Variablen eingehen. Der GFID^{*}-Index ist wie folgt definiert:

$$\begin{aligned}\text{GFID}^*(q) &= 1 - \frac{\text{RMD}^*(q)}{\text{RMD}^*(o)} \text{ mit} \\ \text{RMD}^*(q) &= \sum_i \sum_{j>i} \sum_{k,k^*} \left(d_{i(k)j(k^*)}^{\text{emp}} - d^2(q)_{i(k)j(k^*)} \right)^2 \text{ und} \\ \text{RMD}^*(o) &= \sum_i \sum_{j>i} \sum_{k,k^*} \left(d_{i(k)j(k^*)}^{\text{emp}} \right)^2,\end{aligned}$$

wobei $d_{i(k)j(k^*)}^{\text{emp}}$ die empirischen Distanzen und $d^2(q)_{i(k)j(k^*)}$ die berechneten quadrierten euklidischen Distanzen für q Dimensionen sind. Diese sind wie folgt definiert:

$$\begin{aligned}d_{i(k)j(k^*)}^{\text{emp}} &= \frac{1}{p_{i(k)}} (1 - p_{i(k)}) + \frac{1}{p_{j(k^*)}} \cdot (1 - p_{j(k^*)}) - 2 \cdot \left(\frac{p_{i(k)j(k^*)}}{p_{i(k)} \cdot p_{j(k^*)}} - 1 \right) \\ &= \frac{p_{j(k^*)} \cdot (1 - p_{i(k)}) + p_{i(k)} \cdot (1 - p_{j(k^*)})}{p_{i(k)} \cdot p_{j(k^*)}} - \frac{2 \cdot p_{i(k)j(k^*)} + 2 \cdot p_{i(k)} \cdot p_{j(k^*)}}{p_{i(k)} \cdot p_{j(k^*)}} \\ &= \frac{p_{j(k^*)} - p_{i(k)j(k^*)}}{p_{i(k)} \cdot p_{j(k^*)}} + \frac{p_{i(k)} - p_{i(k)j(k^*)}}{p_{i(k)} \cdot p_{j(k^*)}},\end{aligned}$$

sowie

$$d^2(q)_{i(k)j(k^*)} = \sum_{h=1}^q (x_{i(k)h} - x_{j(k^*)h})^2.$$

Die empirischen Distanzen lassen sich als gewichtete City-Block-Metrik interpretieren (Bacher 1996, S. 70–71). Entgegen den Ausführungen von Greenacre (1989) sind somit sowohl die empirischen als auch die im Merkmalsraum berechneten (theoretischen) Distanzen inhaltlich interpretierbar (Carroll, Green u. a. 1989). Für das Beispiel der Tabelle 3.11 auf Seite 57 ergeben sich die in Tabelle 3.19 auf der nächsten Seite dargestellten Werte. Bei zwei Dimensionen wird eine Anpassung von 55,4 Prozent erzielt. Diese ist wie folgt zu interpretieren: Bezogen auf die Quadratsumme der berechneten Distanzen ist die Quadratsumme der Differenzen zwischen berechneten und empirischen Distanzen um 55,4 Prozent kleiner. Wir können auch sagen: Die zweidimensionale Lösung bildet die empirischen Distanzen zu 55,4 Prozent ab. Bei drei Dimensionen ergibt sich eine Anpassung von 63,6 Prozent, bei vier von 66,6 Prozent usw. Im Unterschied zur faktorenanalytischen Interpretation ist eine eindimensionale Lösung für die Interpretation der Distanzen nicht geeignet ($\text{GFIR}^*(1) = 0,456$ für die faktorenanalytische Interpretation; $\text{GFID}^*(1) = 0,181$ für die clusteranalytische Interpretation).

Schwellenwerte, ab denen ein Anpassungsindex als gut zu beurteilen ist, fehlen. In Anlehnung an die faktorenanalytische Praxis, bei der oft Faktorladungen mit einem Abso-

Tab. 3.19: $GFID^*$ -Koeffizienten für das Illustrationsbeispiel

Zahl der Dimensionen	$GFID^*$	Zahl der Dimensionen	$GFID^*$
0	0,000	7	0,885
1	0,181	8	0,913
2	0,554	9	0,921
3	0,636	10	0,936
4	0,666	11	0,998
5	0,739	12	1,000
6	0,828		

lubetrag größer 0,5 als Mindestgrößen⁶ betrachtet werden, lassen sich die in Tabelle 3.20 dargestellten Richtwerte angeben, wenn das Intervall zwischen 0,5 und 1,0 in fünf gleich große Intervalle unterteilt wird. Da die Anpassungsindizes $GFIR^*$ und $GFID^*$ quadrierte Größen sind (erklärte Varianz), werden die Schwellenwerte quadriert. Von einer sehr guten Modellanpassung beispielsweise kann dann gesprochen werden, wenn der Wert des Index zwischen 0,81 und 1,00 liegt.

Tab. 3.20: Schwellenwerte zur Interpretation der Anpassungsindizes

Schwellenwerte für den Absolutbetrag der Faktorladung	Schwellenwerte $GFIR^*$ und $GFID^*$	Interpretation
0,9 bis 1,0	0,81 bis 1,00	sehr gute Modellanpassung
0,8 bis 0,9	0,64 bis 0,81	gute Modellanpassung
0,7 bis 0,8	0,49 bis 0,64	befriedigende Modellanpassung
0,6 bis 0,7	0,36 bis 0,49	ausreichende Modellanpassung
0,5 bis 0,6	0,25 bis 0,36	noch ausreichende Modellanpassung
kleiner 0,5	kleiner 0,25	Modellanpassung ist nicht ausreichend

3.4 Anwendungsempfehlungen

1. Es sollte zunächst mittels des χ^2 -Unabhängigkeitstests geprüft werden, ob zwischen den untersuchten Variablen ein signifikanter Zusammenhang vorliegt und eine multiple Korrespondenzanalyse überhaupt sinnvoll ist (siehe Abschnitt 3.3.1).
2. Zur Bestimmung der Dimensionalität bei der faktorenanalytischen Interpretation sollte der $GFIR^*$ -Index eingesetzt werden (siehe Abschnitt 3.3.3). Dabei kann die Dimension

⁶ Mitunter werden strengere Schwellenwerte von 0,6 gefordert, in der Forschungspraxis werden aber auch oft noch Faktorladungen größer 0,4 als brauchbar erachtet.

mit dem größten GFIR^{*}-Wert ausgewählt werden; zusätzlich können noch Lösungen mit einem befriedigenden Wert betrachtet werden (siehe Tabelle 3.20).

3. Zur inhaltlichen Benennung der Dimensionen sollten Ausprägungen mit hohen positiven oder negativen Skalenwerten verwendet werden, sofern ihre Varianzbeiträge ausreichend groß sind (Richtwert: 0,09, siehe Abschnitt 3.2.4).
4. Für die clusteranalytische Interpretation sollte der GFID^{*}-Index eingesetzt werden.
5. Sind die Anpassungsindizes GFID^{*} und GFIR^{*} nicht verfügbar, können die Eigenwerte bzw. die erklärten Varianzen verwendet werden. Der Anwender sollte sich aber der Problematik dieser Kenngrößen bewusst sein.
6. Sind die Ausprägungen ordinal geordnet, sollte geprüft werden, ob diese Ordnung in den Skalenwerten abgebildet wird (siehe Abschnitt 3.2.4 und Tabelle 3.5 auf Seite 49).
7. Zur inhaltlichen Validitätsprüfung sollten Hypothesen formuliert und empirisch überprüft werden (siehe Abschnitt 3.1.1).

4 Nichtmetrische mehrdimensionale Skalierung

4.1 Aufgabenstellung und Ähnlichkeitsmessung

Die *nichtmetrische mehrdimensionale Skalierung* (MDS) verfolgt im Prinzip dieselbe Aufgabenstellung wie die multiple Korrespondenzanalyse: Für eine Ähnlichkeits- oder Unähnlichkeitsmatrix zwischen Objekten soll eine räumliche Darstellung gefunden werden. Die zentralen Unterschiede zur multiplen Korrespondenzanalyse sind:

1. Es können (Un-)Ähnlichkeitsmaße verwendet werden, die die Information des jeweiligen Messniveaus ausschöpfen. Ferner können direkt erhobene (Un-)Ähnlichkeitsmatrizen verwendet werden.
2. Die Koordinatenwerte werden iterativ berechnet. Daher besteht die Gefahr von lokalen Minima und von instabilen Lösungen.
3. In die Berechnung geht nur die ordinale (nichtmetrische) Information der (Un-)Ähnlichkeiten ein.
4. Die Dimensionen werden nach keinem formalen Kriterium angeordnet. Die erste Dimension muss also nicht die größte Streuung besitzen.
5. Die Koordinatenwerte der Objekte auf einer Dimension ändern sich, wenn eine weitere Dimension hinzugenommen wird.
6. Eine formale Rechtfertigung der Interpretation der Dimensionen ist nicht möglich, da die Dimensionen nach keinem formalen Kriterium berechnet werden. Eine faktorenanalytische Interpretation ist möglich, sie kann aber durch die Eingabe von inhaltlich begründeten Startwerten für die Dimensionen oder durch die Projektion von externen Variablen in die berechnete Punktekonfiguration erleichtert werden.

Die nichtmetrische mehrdimensionale Skalierung besitzt gegenüber Anwendungssituationen eine größere Offenheit als die multiple Korrespondenzanalyse. Sie setzt nur voraus, dass eine (Un-)Ähnlichkeitsmatrix mit ordinaler Information vorliegt, dass also zwischen den Ähnlichkeiten der untersuchten Klassifikationsobjekte eine Ordnungsrelation

der Art $\ddot{a}_{12} \geq \ddot{a}_{34} \geq \ddot{a}_{23}$ usw. besteht, wobei \ddot{a}_{ij} die Ähnlichkeit zwischen den Klassifikationsobjekten i und j ist. Grundsätzlich bestehen zwei Möglichkeiten, eine (Un-)Ähnlichkeitsmatrix zu gewinnen:

1. *Die (Un-)Ähnlichkeitsmatrix wird direkt empirisch erhoben*, zum Beispiel durch die Methode des Paarvergleichs, des Triaden- oder Tetradenvergleichs (Sixtl 1982, S. 294–311). Für die Methode des Paarvergleichs könnte das Befragungsdesign wie folgt aussehen: Den Personen werden alle Paare von Objekten (zum Beispiel politische Parteien) oder eine Auswahl der Objektpaare vorgelegt. Für jedes Paar sollen die Befragten auf einer neunstufigen Skala (1 »sehr ähnlich« bis 9 »überhaupt nicht ähnlich«) ein Ähnlichkeitsurteil abgeben.
2. Die (Un-)Ähnlichkeitsmatrix wird aus einer Datenmatrix berechnet. Hier sind wiederum zwei Zugänge möglich:
 - a) *Indirekte Methode durch Stimulusskalierung*: Die Befragten (Experten und Expertinnen) werden aufgefordert, jedes Objekt (zum Beispiel politische Parteien) hinsichtlich ihres Wertes auf den einzelnen Dimensionen (zum Beispiel Kompetenz in der Bildungspolitik, Kompetenz in der Arbeitsmarktpolitik usw.) einzustufen.
 - b) *Indirekte Methode durch Responseskalierung*: Es wird erhoben, bei welchen Personen die Objekte (zum Beispiel Sympathie mit bestimmten politischen Parteien) auftreten. Ein gemeinsames Auftreten oder Nichtauftreten wird als Ähnlichkeit interpretiert. Durch Aggregierung über alle Befragte oder eine Befragtengruppe werden Ähnlichkeitsmaße berechnet.

Die drei Zugänge werden in der Übersicht 4.1 am Beispiel der Analyse von Freizeitaktivitäten verdeutlicht. Ziel der Analyse ist, die dimensionale Struktur von Freizeitaktivitäten zu ermitteln. Aus Gründen der Übersichtlichkeit werden nur vier Freizeitaktivitäten dargestellt. Formal unterscheiden sich die drei dargestellten Methoden dadurch, dass bei dem ersten Erhebungsdesign direkt eine Ähnlichkeitsmatrix vorliegt, während sie bei den beiden zuletzt genannten berechnet werden muss. Dazu kann ein in Kapitel 8 behandeltes (Un-)Ähnlichkeitsmaß verwendet werden.¹

Die nichtmetrische mehrdimensionale Skalierung setzt voraus, dass eine direkt erhobene oder aus den Daten berechnete *Unähnlichkeitsmatrix* vorliegt. Wir wollen diese im

¹ Die drei Zugänge unterscheiden sich auch noch in folgender Hinsicht: Bei den Zugängen 1 und 2a liegt für jede Person eine Un- oder Ähnlichkeitsmatrix vor. Zur mehrdimensionalen Skalierung können dann spezielle Verfahren eingesetzt werden, die individuell gewichtete Merkmalsräume berechnen und mit INDSCAL (»individual scaling«; Carroll und Chang 1970) abgekürzt werden. Beim Zugang 2b liegen keine individuellen Un- oder Ähnlichkeitsmatrizen vor. Sie können nur für alle Befragten oder Befragtengruppen berechnet werden. INDSCAL-Techniken können dann für (Un-)Ähnlichkeitsmatrizen von Befragtengruppen durchgeführt werden. INDSCAL-Techniken werden hier nicht behandelt, sie sind zum Beispiel in IBM-SPSS verfügbar.

Auf der nachfolgenden Liste werden jeweils zwei Freizeitaktivitäten als Paar dargestellt. Geben Sie bitte für jedes Paar an, ob die beiden Freizeitaktivitäten sehr ähnlich (1) oder überhaupt nicht ähnlich (7) sind.

			sehr ähnlich		überhaupt nicht ähnlich	
faulenzen	—	ein Buch lesen	①	②	③	④
faulenzen	—	Sport betreiben	①	②	③	④
faulenzen	—	fernsehen	①	②	③	④
usw.			①	②	③	④
					⑤	⑥
					⑦	

(a) Direkte Methode durch Befragung von Ähnlichkeitsurteilen

Freizeitaktivitäten können von unterschiedlichen Perspektiven aus betrachtet werden. Sie können zum Beispiel im Freien oder in geschlossenen Räumen stattfinden, sie können gemeinsam mit der Familie unternommen werden, sie können unterschiedlich anstrengend sein usw. Wie würden Sie nachfolgende Freizeitaktivitäten hinsichtlich der nachfolgenden Merkmale beurteilen?

		sehr anstrengend		überhaupt nicht anstrengend	
faulenzen	①	②	③	④	⑤
spazierengehen	①	②	③	④	⑤
Sport betreiben	①	②	③	④	⑤
usw.	①	②	③	④	⑤
				⑥	⑦

		immer gemeins. mit Familie		nie gemeins. mit Familie	
faulenzen	①	②	③	④	⑤
spazierengehen	①	②	③	④	⑤
ein Buch lesen	①	②	③	④	⑤
usw.	①	②	③	④	⑤
				⑥	⑦

(b) Indirekte Methode durch Stimulusskalierung

Die nachfolgende Liste enthält eine Reihe von Aktivitäten, die man in der Freizeit machen kann. Welche von diesen Aktivitäten haben Sie in der letzten Woche ausgeübt?

	habe ich gemacht	habe ich nicht gemacht
gefaulenzt	①	②
spazierengangen	①	②
ein Buch gelesen	①	②
usw.	①	②

(c) Indirekte Methode durch Responseskalierung

Abb. 4.1: Unterschiedliche Erhebungsdesigns zur Erfassung der Ähnlichkeit von Freizeitaktivitäten

Tab. 4.1: Beispiel für eine unvollständige Unähnlichkeitsmatrix

	A	B	C	D	E	F
A	—	4	2	1	6	7
B	4	—	1	3	5	4
C	2	1	—	KW	6	6
D	1	3	KW	—	5	4
E	6	5	6	5	—	3
F	7	4	6	4	3	—

Abkürzungen: KW: fehlender Wert

Folgenden mit \mathbf{U} bezeichnen und anstelle der umständlichen, aber präzisen Bezeichnung »Klassifikationsobjekte« nur von Objekten sprechen. Damit können entweder Variablen (Spalten einer Datenmatrix) oder Objekte (Zeilen einer Datenmatrix) gemeint sein. Liegt eine Ähnlichkeitsmatrix $\tilde{\mathbf{A}}$ vor, kann diese in der nichtmetrischen mehrdimensionalen Skalierung durch Multiplikation mit -1 in eine Unähnlichkeitsmatrix transformiert werden ($\mathbf{U} = -1 \cdot \tilde{\mathbf{A}}$), da nur die ordinale Information der Unähnlichkeiten verwendet wird.

Die zu analysierende Unähnlichkeitsmatrix muss nicht vollständig sein. Einige Unähnlichkeiten können fehlende Werte aufweisen. In der fiktiven unvollständigen Unähnlichkeitsmatrix der Tabelle 4.1 ist die Unähnlichkeit zwischen den Objekten C und D nicht bekannt. Die Unähnlichkeitsmatrix ist wie folgt zu lesen (größere Werte drücken eine geringere Ähnlichkeit aus): Die Unähnlichkeit für die Objekte A und D und jene für die Objekte B und C sind gleich. Die Objektpaare (A,D) und (B,C) sind sich am ähnlichsten. Die Unähnlichkeit zwischen A und C ist größer als jene zwischen A und D bzw. zwischen B und C usw.

4.2 Schätzalgorithmus

Für die Darstellung des Kalküls der nichtmetrischen mehrdimensionalen Skalierung soll die in Tabelle 4.2 dargestellte Notation eingeführt werden. Die Koordinatenwerte x_{ih} der Objekte i ($i = 1, 2, \dots, m$) auf den Dimensionen h ($h = 1, 2, \dots, q$) werden bei der nichtmetrischen mehrdimensionalen Skalierung so bestimmt, dass die Größenanordnung der berechneten Distanzen $d(q,p)_{ij}$ zwischen den Objekten im gesuchten Merkmalsraum mit der Größenanordnung der empirischen Unähnlichkeiten u_{ij} bestmöglich übereinstimmt. Für den Fall einer perfekten Übereinstimmung soll gelten:

$$u_{ij} \leq u_{i^*j^*} \leq u_{i^{**}j^{**}} \leq \dots \Leftrightarrow d(q,p)_{ij} \leq d(q,p)_{i^*j^*} \leq d(q,p)_{i^{**}j^{**}} \leq \dots \quad (4.1)$$

Tab. 4.2: Notation für die mehrdimensionale Skalierung

Symbol	Beschreibung
U	zu untersuchende Unähnlichkeitsmatrix
i, j	Indizes für die Objekte der Matrix U
$u_{i,j}$	Unähnlichkeit zwischen Objekt i und j
q	Zahl der untersuchten Dimensionen
h	Index für eine Dimension
m	Zahl der Objekte
x_{ih}	Koordinatenwert des Objekts i auf der Dimension h
x_{jh}	Koordinatenwert des Objekts j auf der Dimension h
X	$(m \times q)$ -Matrix der Koordinatenwerte (auch als Punktekonfiguration bezeichnet)
p	Metrikparameter

Hinsichtlich der Behandlung von Bindungen $u_{ij} = u_{i^*j^*}$ gibt es zwei Ansätze (Coxon 1982, S. 52–53): (1) Es wird nur *schwache Monotonie* gefordert (schwache Monotoniebedingung):

$$u_{ij} = u_{i^*j^*} \Rightarrow d_{ij} \leq d_{i^*j^*} \quad \text{oder} \quad d_{ij} \geq d_{i^*j^*}. \quad (4.2)$$

Sind die Unähnlichkeiten zwischen zwei Objektpaaren (i, j) und (i^*, j^*) gleich, so müssen die berechneten Distanzen nicht gleich sein. (2) Bei der *starken Monotoniebedingung* wird dagegen gefordert, dass bei Gleichheit der empirischen Unähnlichkeiten auch die berechneten Distanzen gleich sein müssen (starke Monotoniebedingung):

$$u_{ij} = u_{i^*j^*} \Rightarrow d_{ij} = d_{i^*j^*}. \quad (4.3)$$

Als Maß für die Übereinstimmung der beiden Ordnungsrelationen wird in der nichtmetrischen mehrdimensionalen Skalierung der *sogenannte STRESS* verwendet. Er ist wie folgt definiert:

$$\text{STRESS}(q,p) = \sqrt{\frac{\sum_i \sum_{j>i} (d(q,p)_{ij} - \hat{d}(q,p)_{ij})^2}{\sum_i \sum_{j>i} d(q,p)_{ij}^2}}, \quad (4.4)$$

wobei $\hat{d}(q,p)_{ij}$ die erwarteten oder theoretischen Distanzen und $d(q,p)_{ij}$ die im q -dimensionalen Raum berechneten Distanzen sind. Die berechneten Distanzen $d(q,p)_{ij}$ gehören der Gruppe der so genannten Minkowski-Metrik an. Die Minkowski-Metrik (siehe Formel 8.12 auf Seite 219) setzt voraus, dass ein *Metrikparameter* spezifiziert ist. Als Symbol für diesen Metrikparameter wird der Buchstabe » p « verwendet. Das Symbol » q « steht für die Dimensionszahl. Allgemein berechnet sich die Distanz zwischen zwei Objekten i und j in der Minkowski-Metrik wie folgt:

$$d(q,p)_{ij} = \left[\sum_{h=1}^q |x_{ih} - x_{jh}|^p \right]^{1/p}. \quad (4.5)$$

Tab. 4.3: Berechnete und empirische Unähnlichkeiten

	berechnete Distanzen				emp. Unähnlichkeit			
	1	2	3	4	1	2	3	4
1	0,00	2,00	2,00	2,82	—	1	1	2
2	2,00	0,00	2,82	2,00	1	—	3	3
3	2,00	2,82	0,00	2,00	1	3	—	4
4	2,82	2,00	2,00	0,00	2	4	3	—

Die absoluten Abweichungen $|x_{ih} - x_{jh}|$ werden mit p potenziert und über die q Dimensionen hinweg summiert. Aus dem Summenwert wird die p -te Wurzel gezogen. Für $p = 1$ ergibt sich die sogenannte *City-Block-Metrik* und für $p = 2$ die *euklidische Distanz*. Die Bedeutung des Metrikparameters wird ausführlich in Abschnitt 8.5 behandelt. Wir wollen hier anhand eines fiktiven Beispiels die Berechnung der euklidischen Distanzen in einem zweidimensionalen Raum ($q = 2$) darstellen (siehe Abbildung 4.2). Das Objekt 1 besitzt in der ersten Dimension einen Koordinatenwert von -1 und in der zweiten Dimension von 1 . Somit ist x_{11} also gleich -1 und x_{12} gleich 1 . Das Objekt 2 besitzt in beiden Dimensionen einen Wert von 1 , x_{21} und x_{22} sind also gleich 1 . Die euklidische Distanz (Metrikparameter $p = 2$) zwischen dem Objekt 1 und 2 ist gleich:

$$\begin{aligned} d(q=2, p=2)_{12} &= \left[|x_{11} - x_{21}|^2 + |x_{12} - x_{22}|^2 \right]^{1/2} \\ &= \left[|-1 - 1|^2 + |1 - 1|^2 \right]^{1/2} \\ &= 2 . \end{aligned}$$

Zwischen den Objekten 1 und 4 ergibt sich eine euklidische Distanz von

$$\begin{aligned} d(q=2, p=2)_{14} &= \left[|x_{11} - x_{41}|^2 + |x_{12} - x_{42}|^2 \right]^{1/2} \\ &= \left[|-1 - 1|^2 + |1 - (-1)|^2 \right]^{1/2} \\ &= 8^{1/2} = 2,82 . \end{aligned}$$

Insgesamt ergeben sich die in Tabelle 4.3 berechneten Distanzen. Diesen werden die in der Tabelle ebenfalls wiedergegebenen empirischen Unähnlichkeiten gegenübergestellt.

Die erwarteten Distanzen $\hat{d}(q,p)_{ij}$ der Gleichung 4.4 auf der vorherigen Seite werden nun so berechnet, dass die Bedingung 4.1 erfüllt ist. Dazu ist es zweckmäßig, die empirischen Unähnlichkeiten der Größe nach anzugeordnen (siehe Tabelle 4.4 auf Seite 84). Wir prüfen für die aufeinander folgenden Objektpaare die Monotoniebedingung 4.1, wobei Bindungen die starke Monotoniebedingung (Gleichung 4.3 auf der vorherigen Seite)

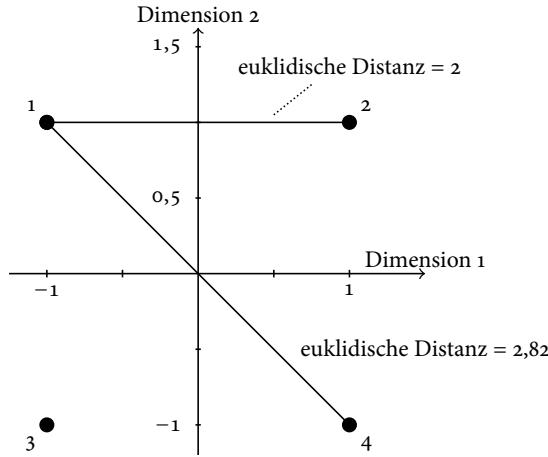


Abb. 4.2: Fiktive Punktekonfiguration in einem zweidimensionalen Raum

erfüllen sollen. Das dem ersten Objektpaar (1,2) nachfolgende Objektpaar (1,3) erfüllt die Monotoniebedingung, da $u_{12} = u_{13}$ und $d(q,p)_{12} = d(q,p)_{13}$. Das dem zweiten Objektpaar (1,3) nachfolgende Objektpaar (1,4) erfüllt ebenfalls die Monotoniebedingung, da $u_{13} < u_{14}$ und $d(q,p)_{13} \leq d(q,p)_{14}$. Auch für das dem Objektpaar (1,4) nachfolgende Objektpaar (2,3) ist die Monotoniebedingung erfüllt, da $u_{14} < u_{23}$ und $d_{14} \leq d_{23}$. Eine Verletzung der Monotoniebedingung tritt dagegen beim Objektpaar (3,4) auf. Die strenge Monotonieforderung: $u_{23} = u_{34} \Rightarrow d_{23} = d_{34}$ ist nicht erfüllt, da $d_{34} < d_{23}$. Für d_{23} und d_{34} würde sich ein Wert von 2,41 (= $(2,82 + 2,00)/2$) ergeben, wenn der Durchschnitt bzw. Mittelwert als Schätzwert eingesetzt werden soll. Dieser Wert ist aber zu klein und führt dazu, dass die Monotoniebedingung zwischen den Objektpaaren (1,4) und (2,3) bzw. (3,4) verletzt ist, da $d(q,p)_{14} > d(q,p)_{23} = d(q,p)_{34}$ ist, während $u_{14} < u_{23}$. Wir nehmen daher in die Berechnung der erwarteten Distanzen für die Objektpaare (2,3) und (3,4) auch das Objektpaar (1,4) auf und erhalten erwartete Distanzen von 2,547 (= $(2,82 + 2,82 + 2,00)/3$) für diese drei Objektpaare.

Wenn wir mit der Prüfung der Monotoniebedingung fortfahren, so ist diese für die Objektpaare (3,4) und (2,4) verletzt. In die Berechnung der erwarteten Distanz von (3,4) gingen die Objektpaare (1,4), (2,3) und (3,4) ein. Die erwartete Distanz für die Objektpaare (1,4) bis (3,4) wird daher als Mittelwert aus den Objektpaaren (1,4), (2,3), (3,4) und (2,4) berechnet. Es ergibt sich ein Wert von 2,410 (= $(2,82 + 2,82 + 2,00 + 2,00)/4$) für die erwartete Distanz der Objektpaare (1,4) bis (2,4). Damit sind alle aufeinander folgenden Objektpaare geprüft, und es ergeben sich die in der zweiten Spalte (Tabelle 4.4 auf der nächsten Seite) für die erwarteten Distanzen angeführten Werte. Diese erfüllen – wie man leicht nachprüfen kann – die Monotoniebedingung (Gleichung 4.1 auf Seite 80).

Tab. 4.4: Berechnung der erwarteten Distanzen

Objekt-paare <i>i,j</i>	emp. Unähnl. <i>u_{ij}</i>	berech. Distanzen $d(q,p)_{ij}$	erwart. Distanzen $\hat{d}(q,p)_{ij}$	Nenner STRESS $(d(q,p)_{ij})^2$	Zähler STRESS $(d(q,p)_{ij} - \hat{d}(q,p)_{ij})^2$
1,2	1	2,00	2,000	4,0	0,000
1,3	1	2,00	2,000	4,0	0,000
1,4	2	2,82	2,410	8,0	0,168
2,3	3	2,82	2,410	8,0	0,168
3,4	3	2,00	2,410	4,0	0,168
2,4	4	2,00	2,410	4,0	0,168
Summe				32,0	0,672

Anmerkung: Die Objektpaare sind entsprechend den empirischen Unähnlichkeiten geordnet.

Der Algorithmus zur Berechnung der erwarteten Distanzen \hat{d} ist allgemein:

Schritt 1: Ordne die Objektpaare entsprechend den empirischen Unähnlichkeiten.

Schritt 2: Setze die erwarteten Distanzen gleich den berechneten Distanzen.

Schritt 3: Wähle das erste Objektpaar (*i,j*) in der geordneten Liste aus.

Schritt 4: Prüfe für das nachfolgende Objektpaar (*i⁺,j⁺*) die Monotoniebedingung. Ist diese verletzt, fahre mit Schritt 5 fort, andernfalls gehe zu Schritt 8.

Schritt 5: Bestimme die Zahl der Objektpaare, die in die Berechnung der erwarteten Distanz $\hat{d}(q,p)_{ij}$ des Objektpaares (*i,j*) eingegangen sind. Diese Zahl soll mit *k* und das erste Objektpaar, das in die Berechnung einging, mit (*i⁻,j⁻*) bezeichnet werden.

Schritt 6: Bilde den Durchschnitt aus $(\sum_{i'=i^-, j'=j^-} d(q,p)_{i'j'} + d(q,p)_{i^+j^+}) / (k + 1)$ als neuen Schätzwert für die erwarteten Distanzen der Objektpaare (*i⁻,j⁻*) bis (*i⁺,j⁺*).

Schritt 7: Prüfe, ob dieser Durchschnittswert größer/gleich der vorausgehenden erwarteten Distanz $\hat{d}(q,p)_{i^--j^{--}}$ ist. Bei »ja« ist die Berechnung abgeschlossen. Bei »nein« setze $i^- = i^{--}$ und $j^- = j^{--}$ und gehe zu Schritt 6.

Schritt 8: Gehe zum nächsten Objektpaar über. Ist kein Objektpaar mehr vorhanden, beende die Berechnung, andernfalls gehe zu Schritt 3.

Sind die erwarteten und berechneten Distanzen bekannt, kann entsprechend Gleichung 4.4 auf Seite 81 der STRESS berechnet werden. Die entsprechende Information für das Beispiel enthält Tabelle 4.4. Die Summe im Nenner ist gleich der Summe der quadrierten berechneten Distanzen. Für unser Beispiel ergibt sich ein Wert von 32,0. Im

Tab. 4.5: Richtwerte zur Interpretation des STRESS-Koeffizienten

Wert des STRESS-Koeffizienten	Interpretation
STRESS = 0,000	perfekt (»perfect«)
0,000 < STRESS ≤ 0,025	ausgezeichnet (»excellent«)
0,025 < STRESS ≤ 0,050	gut (»good«)
0,050 < STRESS ≤ 0,100	befriedigend (»fair«)
0,100 < STRESS ≤ 0,200	ausreichend (»poor«)
STRESS > 0,200	nicht ausreichend

Zähler steht die Summe der quadrierten Abweichungen zwischen den berechneten und erwarteten Distanzen. Diese ist gleich 0,672. Es ergibt sich daher ein STRESS von

$$\text{STRESS} = \sqrt{\frac{0,672}{32}} = 0,145 .$$

Zur Interpretation des STRESS-Koeffizienten können die in der Tabelle 4.5 angegebenen Richtwerte von Kruskal (1964a, S. 3) verwendet werden. In unserem Beispiel würde also eine ausreichende Modellanpassung vorliegen, da der STRESS zwischen 0,1 und 0,2 liegt.

Unser bisheriges Vorgehen bestand darin, für eine gegebene Unähnlichkeitsmatrix eine Punktekonfiguration in einem zweidimensionalen Raum anzunehmen und zu prüfen, wie gut die berechneten Distanzen mit den empirischen Unähnlichkeiten übereinstimmen. Wir können uns nun fragen, ob durch eine Änderung der Koordinatenwerte ein geringerer STRESS-Wert erzielt wird und wie gegebenenfalls die Koordinatenwerte geändert werden sollen. Dabei ist es naheliegend, die Koordinatenwerte so zu ändern, dass der STRESS minimiert wird. Dies wird durch die so genannte *Gradientenmethode* erreicht. Die Koordinatenwerte werden dabei wie folgt neu berechnet:

$$x_{ih}^{\text{neu}} = x_{ih}^{\text{alt}} - \alpha \cdot \text{grad}(x_{ih}^{\text{alt}}) , \quad (4.6)$$

wobei x_{ih}^{alt} der alte Koordinatenwert des Objekts i in der Dimension h ist und x_{ih}^{neu} der neue. Die Schrittlänge wird mit α bezeichnet und $\text{grad}(x_{ih}^{\text{alt}})$ ist die erste partielle Ableitung des STRESS nach dem Koordinatenwert x_{ih} :

$$\text{grad}(x_{ih}^{\text{alt}}) = \frac{\partial \text{STRESS}(q, p)}{\partial x_{ih}^{\text{alt}}} . \quad (4.7)$$

Zur Berechnung der partiellen Ableitungen siehe Hartung und Elpelt (1984, S. 410), Kruskal (1964b) oder Sixtl (1982, S. 346–347). Die Schrittlänge α hat die Funktion, die Konvergenz des Verfahrens zu beschleunigen. Das eben dargestellte Vorgehen kann so

lange wiederholt werden, bis sich der STRESS-Wert nicht mehr ändert oder ein vorgegebener Schwellenwert unterschritten wird. Als Algorithmus dargestellt, sieht das Vorgehen bei einer vorgegebenen Dimensionszahl q und einem vorgegebenen Metrikparameter p folgendermaßen aus:

Schritt 1: Berechne zufällige oder verwende vom Benutzer eingegebene Startwerte.

Schritt 2: Normiere die Punktekonfiguration X mit: $\sum_i x_{ih} = 0$. Die Normierung hat keinen Einfluss auf den STRESS. Sie dient nur dazu, eine graphische Darstellung zu ermöglichen.

Schritt 3: Berechne die Distanzen zwischen den Objekten entsprechend Gleichung 4.5 auf Seite 81.

Schritt 4: Berechne die erwarteten Distanzen.

Schritt 5: Berechne den STRESS entsprechend Gleichung 4.4 auf Seite 81. Überprüfe, ob eine weitere Iteration erforderlich ist. Bei »nein« beende das Verfahren, andernfalls gehe zu Schritt 6.

Schritt 6: Berechne neue Koordinatenwerte mit dem Gradientenverfahren entsprechend Gleichung 4.6 auf der vorherigen Seite mit $x_{ih}^{\text{neu}} = x_{ih}^{\text{alt}} - \alpha \cdot \text{grad}(x_{ih}^{\text{alt}})$

Schritt 7: Gehe zu Schritt 2.

Abbruchskriterien im Schritt 5 können beispielsweise (Kruskal 1964b) sein:

1. Es wurde eine konvergente Lösung gefunden. Der berechnete STRESS liegt unter einem vorgegebenen Schwellenwert, zum Beispiel 0,05.
2. Es wurde eine nichtkonvergente Lösung gefunden. Beim Algorithmus kann der Fall eintreten, dass nach einer bestimmten Iterationszahl keine wesentliche Verbesserung gegenüber vorausgehenden Lösungen erzielt wird. In diesem Fall ist es »sinnvoll«, den Algorithmus abzubrechen, da nicht zu erwarten ist, dass bei einer Fortsetzung der Iteration entscheidende Verbesserungen erzielt werden. Nichtkonvergenz liegt auch vor, wenn die Zahl der Iterationen einen bestimmten Maximalwert überschreitet.

Bei einer nichtkonvergenten Lösung kann eine neue zufällige Ausgangskonfiguration erzeugt und der Algorithmus erneut durchlaufen werden. Bei eingegebenen Startwerten ist dies nicht möglich. Da in der Forschungspraxis die Zahl der Dimensionen q und in bestimmten Anwendungssituationen der Metrikparameter p nicht bekannt sind, wird die nichtmetrische mehrdimensionale Skalierung mit einer unterschiedlichen Dimensionszahl q und unterschiedlichen Metrikparametern p durchgerechnet. Für jede Parameterkonstellation (p,q) empfiehlt sich das Durchrechnen mehrerer Lösungen, um die Stabilität zu prüfen und lokale Minima zu vermeiden.

4.3 Maximale und angemessene Dimensionszahl

Bei der nichtmetrischen mehrdimensionalen Skalierung nimmt die Stabilität der Ergebnisse ab, wenn sich die Dimensionszahl der Zahl der Objekte annähert. Es ist daher sinnvoll, die maximale Dimensionszahl in einer Analyse zu begrenzen. Sixtl (1982, S. 349) gibt auf der Grundlage einer Sichtung von Simulationsstudien folgende Richtwerte für die *maximal zu untersuchende Dimensionszahl* an:

1. Bei der Analyse einer vollständigen Unähnlichkeitsmatrix sollte das Verhältnis der Objektzahl zur Dimensionszahl nicht kleiner $5 : 1$ sein. Werden also zehn Objekte untersucht, ist die maximale Dimensionszahl gleich zwei, da bei drei Dimensionen ein Verhältnis von $10 : 3 < 5 : 1$ vorliegen würde.
2. Bei der Analyse einer unvollständigen Unähnlichkeitsmatrix sollte das Verhältnis der Objektzahl zur Dimensionszahl nicht kleiner $7 : 1$ sein. Werden also zehn Objekte untersucht und liegt eine unvollständige Unähnlichkeitsmatrix vor, ist die maximale Dimensionszahl gleich eins, da bei zwei Dimensionen ein Verhältnis von $5 : 1$ vorliegen würde.

Die Richtwerte von Sixtl geben bei unvollständigen Unähnlichkeitsmatrizen nur eine grobe Orientierungshilfe, da die Anzahl fehlender Werte nicht berücksichtigt wird. Dies ist dagegen beim Verhältnis der Freiheitsgrade (»degree of freedom ratio«) der Fall (Spence und Domoney 1974; Young 1970). Das Verhältnis der Freiheitsgrade r_{df} ist wie folgt definiert:

$$r_{df} = \frac{\text{Zahl der Elemente mit validen Werten in der Unähnlichkeitsmatrix}}{\text{Zahl der zu schätzenden Parameter}} .$$

Die Zahl der zu schätzenden Parameter berechnet sich wie folgt:

$$\text{Zahl der zu schätzenden Parameter} = q \cdot (m - 1) - q \cdot (q - 1)/2 ,$$

wobei m die Zahl der Objekte und q die Zahl der Dimensionen ist. Da die Koordinatenwerte in jeder Dimension auf den Mittelwert 0 normiert sind, sind für jede Dimension $m - 1$ Koordinatenwerte zu schätzen. Wegen der Orthogonalität der Dimensionen sind weitere $q \cdot (q - 1)/2$ Koordinatenwerte zu fixieren. Damit beträgt die Zahl der zu schätzenden Parameter $q \cdot (m - 1) - q \cdot (q - 1)/2$. Die Normierung der Gesamtstreuung der Koordinatenwerte wird nicht berücksichtigt. Soll sie mitberücksichtigt werden, ist die Zahl 1 zu subtrahieren. Nach Spence und Domoney (1974) und Young (1970) sollte das Verhältnis der Freiheitsgrade abhängig vom Fehlerausmaß zwischen 2,5 (fehlerfreie Daten) und 3,5 (stark fehlerbehaftete Daten) liegen. Bei zehn Objekten und drei fehlenden Werten liegen für 42 Elemente ($= 10 \cdot (10 - 1)/2 - 3$) der Unähnlichkeitsmatrix Informationen vor. Bei einer Dimension sind 9 ($= 1 \cdot (10 - 1) - 1 \cdot (1 - 1)/2$) Parameter zu schätzen.

Tab. 4.6: Verhältnis der Freiheitsgrade in Abhängigkeit von der Dimensionszahl für die Zusammenhangsmatrix \mathbf{G} der Tabelle 3.13 auf Seite 60

Zahl der Dim.	Zahl der Objekte	Zahl der Elemente mit validen Werten in der Unähnlichkeitsmatrix	Zahl zu schätzender Parameter	Verhältnis der Freiheitsgrade
1	14	49	$1 \cdot (14 - 1) - 1 \cdot (1 - 1)/2 = 13$	$49/13 = 3,87$
2	14	49	$2 \cdot (14 - 1) - 2 \cdot (2 - 1)/2 = 25$	$49/25 = 1,96$
3	14	49	$3 \cdot (14 - 1) - 3 \cdot (3 - 1)/2 = 36$	$49/36 = 1,36$

Das Verhältnis der Freiheitsgrade ist $42/9 = 4,67$. Die Ermittlung einer Dimension ist somit zulässig. Für zwei Dimensionen nimmt das Verhältnis der Freiheitsgrade einen Wert von $42/17 = 2,47$ an, das knapp unter dem Schwellenwert von 2,5 bei fehlerfreien Daten liegt.

Die genannten Richtwerte sind als Orientierungshilfe bei der Auswahl der maximalen Dimensionszahl zu verstehen. In bestimmten Anwendungsfällen kann man aber aus inhaltlichen Gründen an einer größeren Dimensionszahl interessiert sein. In diesem Fall sollte dann die Stabilität der Ergebnisse empirisch überprüft werden, indem mit unterschiedlichen Startkonfigurationen gerechnet wird. Betrachten wir dazu ein Beispiel: Mit Hilfe der nichtmetrischen mehrdimensionalen Skalierung soll wie bei der multiplen Korrespondenzanalyse die Zusammenhangsmatrix \mathbf{G} zwischen der Schulbildung der Mütter und jener ihrer Partner untersucht werden (siehe Tabelle 3.13 auf Seite 60). Da die Zusammenhänge zwischen den Ausprägungen innerhalb einer nominalen Variablen rein rechnerische Größen darstellen, werden sie als fehlende Werte behandelt. Es wird eine unvollständige Ähnlichkeitsmatrix mit 49 validen Werten untersucht. Für die Analyse wurde die Ähnlichkeitsmatrix durch eine Multiplikation mit -1 in eine Unähnlichkeitsmatrix transformiert. In Abhängigkeit von der Dimensionszahl ergeben sich die in Tabelle 4.6 ausgewiesenen Werte für das Verhältnis der Freiheitsgrade.

Die maximale Dimensionszahl ist also 1, da bei zwei Dimensionen das Verhältnis der Freiheitsgrade bereits unter dem Schwellenwert von 2,5 für fehlerfreie Daten liegt. Man wird daher in einem ersten Schritt eine eindimensionale Lösung suchen. Für diese ergibt sich bei fünf unterschiedlichen Versuchen (Startkonfigurationen) ein minimaler STRESS von 0,061, also eine befriedigende Modellanpassung. Die eindimensionale Lösung mit dem STRESS-Wert von 0,061 ist inhaltlich aber nur schwer zu interpretieren. Eine Interpretation als soziale Schichtungsdimension ist nicht möglich. In einem zweiten Schritt wird man daher versuchen, ob – wie bei der multiplen Korrespondenzanalyse – eine zweidimensionale Lösung zu inhaltlich besser interpretierbaren Ergebnissen führt. Das Ergebnis einer ersten Analyse ist in der Abbildung 4.3 dargestellt.

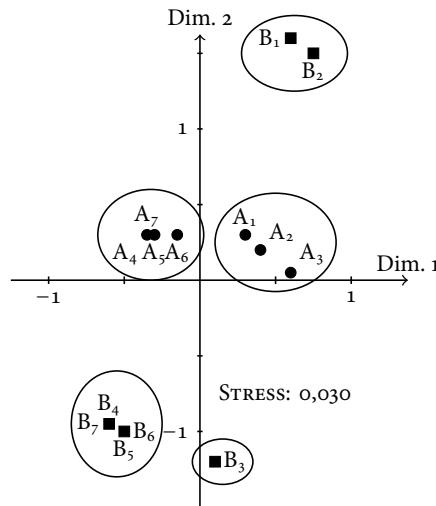


Abb. 4.3: Zweidimensionales Ergebnis der nichtmetrischen mehrdimensionalen Skalierung für die Zusammenhangsmatrix \mathbf{G} (STRESS = 0,030)

In der Lösung lassen sich fünf Cluster erkennen:

Cluster 1 (B_1, B_2): »untere« Bildungsschicht der Partner

Cluster 2 (A_1, A_2, A_3): »untere« und »mittlere« Bildungsschicht der Mütter

Cluster 3 (A_4, A_5, A_6, A_7): »obere« Bildungsschicht der Mütter

Cluster 4 (B_3) »mittlere« Bildungsschicht der Partner

Cluster 5 (B_4, B_5, B_6, B_7): »obere« Bildungsschicht der Partner

Eine inhaltliche Interpretation der Cluster ist möglich. Problematisch an dieser Interpretation ist jedoch, dass die Bildungsschichten der Frauen und Männer getrennt zugeordnet werden. Homogame Heiratsmuster hinsichtlich der Schulbildung sind nicht zu erkennen. Berechnet man die Distanzen (nicht durchgezogene Linien) zwischen den Bildungsschichten der Frauen und ihren Partnern, so ergeben sich nur schwach ausgeprägte Heiratspräferenzen. (Eine geringere Distanz drückt eine höhere Heiratspräferenz aus.) Inhaltlich würde dieses Ergebnis bedeuten, dass eine soziale Abschließung durch homogame Heiratsmuster hinsichtlich der Bildungsschichten nicht oder nur sehr schwach stattfindet. Dies würde eindeutig der impliziten Annahme der Analyse widersprechen, dass eine soziale Schichtung durch soziale Abschließung gekennzeichnet ist. Bevor diese inhaltliche Schlussfolgerung gezogen wird, sollten daher, entsprechend der in Kapitel 1 (Abschnitt 1.9) angeführten Vorgehensweise, andere Fehlerquellen geprüft werden. So ist zum Beispiel zu prüfen, ob die berechneten Ergebnisse nicht ein methodisches Artefakt darstellen und daher inhaltliche Schlussfolgerungen nicht zulässig sind. Die Ergebnisse würden wir dann als methodisches Artefakt betrachten, wenn sie instabil sind, wenn

also unterschiedliche Startkonfigurationen zu unterschiedlichen Punktekonfigurationen mit ähnlichen STRESS-Werten führen. Dies ist tatsächlich der Fall, wie man sich anhand Abbildung 4.4 überzeugen kann. Aus der Stabilitätsanalyse wird man daher folgern, dass im konkreten Anwendungsfall der Einsatz der nichtmetrischen mehrdimensionalen Skalierung nicht sinnvoll ist, da sie zu instabilen Ergebnissen führt.

Bei einer größeren Dimensionszahl kann – wie bei der Faktorenanalyse – ein Scree-Test zur Bestimmung der Dimensionszahl durchgeführt werden.² Dazu werden die STRESS-Koeffizienten in Abhängigkeit von der Dimensionszahl graphisch dargestellt. Wurde der richtige Metrikparameter gewählt, zeigt die Verlaufskurve einen ellenbogenförmigen Knick (siehe Abbildung 4.5 auf Seite 92 und Sixtl 1982, S. 348). Der Knick ist um so stärker ausgeprägt, je geringer das Fehlerausmaß ist. In dem fiktiven Beispiel der Abbildung 4.5 auf Seite 92 ist die »richtige« Dimensionszahl gleich 3, da hier ein ellenbogenförmiger Knick auftritt.

4.4 Unbekannter Metrikparameter p

Die nichtmetrische mehrdimensionale Skalierung ist im Unterschied zu anderen mehrdimensionalen Skalierungsverfahren nicht auf einen euklidischen Merkmalsraum begrenzt. Mit ihr kann ein Minkowski-Raum aufgespannt werden, indem der *Metrikparameter p* variiert wird. Ob eine empirische Bestimmung des Metrikparameters sinnvoll ist, hängt von dem Erhebungsdesign ab. In Abschnitt 4.1 wurden drei Erhebungsdesigns erwähnt:

- Direkte Methode durch die *Befragung von Ähnlichkeitsurteilen*,
- Indirekte Methode durch *Stimulusskalierung*,
- Indirekte Methode durch *Responseskalierung*.

Nur im *ersten Fall* ist eine empirische Bestimmung des Metrikparameters sinnvoll, da dadurch Anhaltspunkte über das Urteilsverhalten der Personen gewonnen werden können. Ein größerer Metrikparameter p bedeutet, dass eine Person (oder eine Personengruppe) größeren Unterschieden auf den Urteilsdimensionen ein größeres Gewicht beimisst als kleineren Unterschieden. Geht der Metrikparameter gegen unendlich ($p = \infty$), wird beim Gesamturteil überhaupt nur jene Dimension mit dem größten Unterschied zwischen zwei Objekten als Urteilsmaßstab verwendet.

Bei einer indirekten Zugangsweise (die Ähnlichkeitsmatrix wird berechnet) ist eine empirische Bestimmung des Metrikparameters wenig sinnvoll, da aus den Daten eine (Un-)Ähnlichkeitsmatrix berechnet wird. Man wird sich hier für einen Metrikparameter

² Siehe dazu Abschnitt 5.3.1 sowie die dort angeführte Literatur.

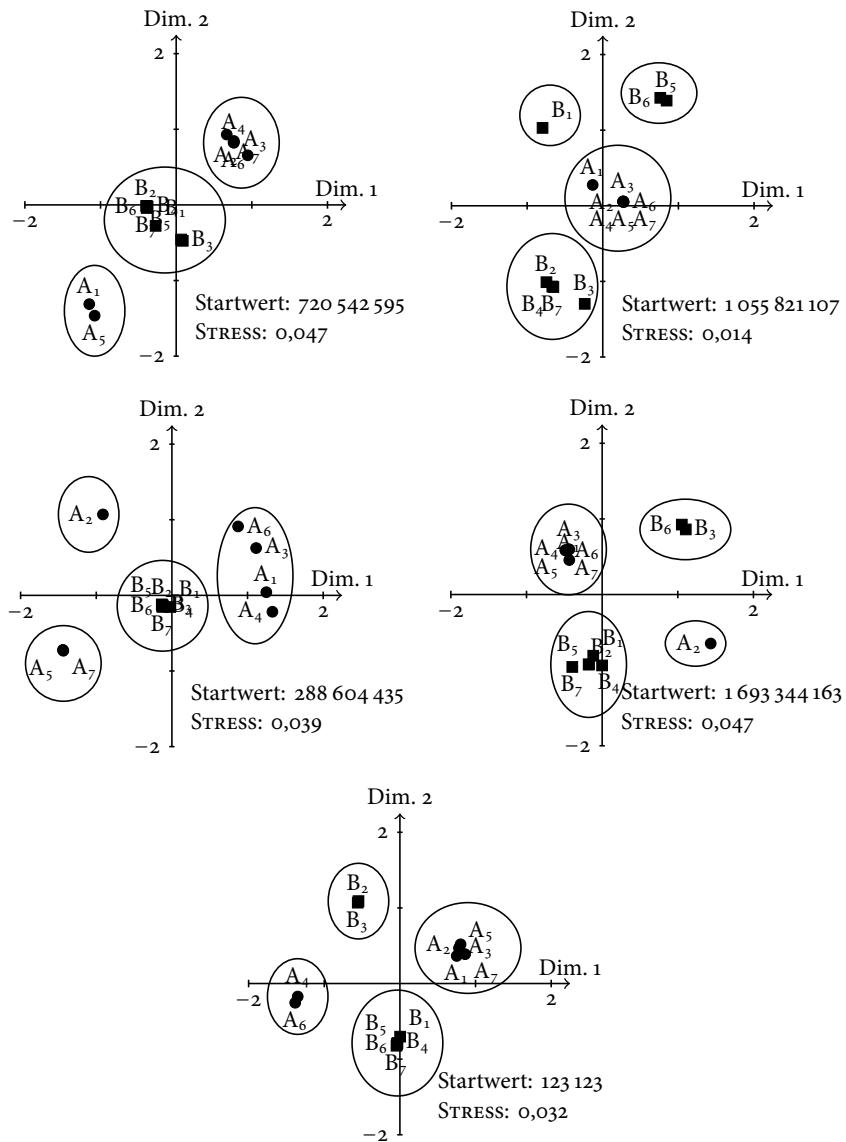


Abb. 4.4: Ergebnisse der multidimensionalen Skalierung für die Zusammenhangsmatrix G bei unterschiedlichen Startwerten für den Zufallszahlengenerator

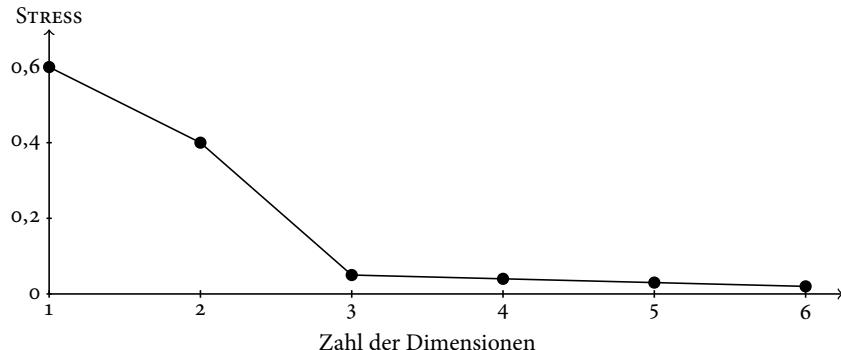


Abb. 4.5: Scree-Test zur Bestimmung der Dimensionszahl

entscheiden, der inhaltlich leicht interpretierbar ist oder bereits bei der Berechnung der empirischen (Un-)Ähnlichkeiten verwendet wurde. Wird bei Berechnung der empirischen Unähnlichkeiten beispielsweise die euklidische Distanz verwendet, wird man sich bei der nichtmetrischen mehrdimensionalen Skalierung auch für die euklidische Distanz ($p = 2$) entscheiden. Besteht dagegen für das verwendete (Un-)Ähnlichkeitsmaß kein eindeutig mathematisch nachweisbarer Bezug zu einem Metrikparameter, wird die Entscheidung entweder auf die City-Block-Metrik (Metrikparameter $p = 1$) oder die euklidische Distanz (Metrikparameter $p = 2$) fallen, da beide Distanzmaße leicht zu interpretieren sind. Formal kann eine *Bevorzugung der euklidischen Distanz* dadurch begründet werden, dass nur bei ihr eine rechtwinklige Rotation der Dimensionen zu keiner Änderung der berechneten Distanzen führt (Coxon 1982, S. 93).

4.5 Weitere Modellanpassungsgrößen

Neben dem STRESS-Koeffizienten wurde noch eine Reihe von weiteren Maßzahlen zur Beurteilung der Ergebnisse einer nichtmetrischen mehrdimensionalen Skalierung entwickelt. So kann man zum Beispiel zur Beurteilung des STRESS-Wertes von folgendem Nullmodell ausgehen: Es besteht kein monotoner Zusammenhang zwischen der Ordnungsrelation der empirischen Unähnlichkeiten und der Ordnungsrelation der berechneten Distanzen. Die erwarteten Distanzen sind daher gleich dem Durchschnitt (Mittelwert) aller berechneten Distanzen:

$$\bar{d}(q,p)_{i,j} = \frac{1}{m \cdot (m - 1)/2} \cdot \sum_i \sum_{j>i} d(q,p)_{i,j} .$$

Auf der Grundlage dieses Nullmodells wird der sogenannte *Anpassungsindex für den STRESS* (»Goodness-of-Fit Index for Stress«, GFIS-Index) konstruiert:

$$\text{GFIS}(q,p) = 1 - \frac{\sum_i \sum_{j>i} (\text{d}(q,p)_{i,j} - \hat{\text{d}}(q,p)_{i,j})^2}{\sum_i \sum_{j>i} (\text{d}(q,p)_{i,j} - \bar{\text{d}}(q,p)_{i,j})^2}.$$

Der Anpassungsindex liegt zwischen 0 und 1. Ein Wert von 0 bedeutet keine prozentuelle Verbesserung, ein Wert von 1 eine Verbesserung von 100 Prozent gegenüber dem Nullmodell.

Die Größe

$$\text{STRESS-2}(q,p) = \sqrt{\frac{\sum_i \sum_{j>i} (\text{d}(q,p)_{i,j} - \hat{\text{d}}(q,p)_{i,j})^2}{\sum_i \sum_{j>i} (\text{d}(q,p)_{i,j} - \bar{\text{d}}(q,p)_{i,j})^2}}$$

wird als STRESS-2-Koeffizient bezeichnet. Dieser ist in der Regel größer als der gewöhnliche STRESS-Koeffizient, da mit einer kleineren Zahl dividiert wird. Weitere Maßzahlen zur Messung der Übereinstimmung zwischen empirischen Unähnlichkeiten und berechneten Distanzen sind in Coxon (1982, S. 89–90) dargestellt. Ihnen liegt entweder die Berechnung von Produkten

$$\sum_i \sum_{j>i} \text{d}(q,p)_{i,j} \cdot \hat{\text{d}}(q,p)_{i,j}$$

oder von Distanzen

$$\sum_i \sum_{j>i} (\text{d}(q,p)_{i,j} - \hat{\text{d}}(q,p)_{i,j})^2$$

zugrunde. Der so genannte »Alienation«-Koeffizient (Coxon 1982, S. 89) beispielsweise wird als Produktmaß berechnet mit

$$A = \sqrt{1 - P^2},$$

wobei P das normierte Kreuzprodukt zwischen erwarteten und berechneten Distanzen ist. Der »Alienation«-Koeffizient nimmt einen Wert von 0 für eine perfekte Modellanpassung und einen Wert von 1 bei einer schlechten Modellanpassung an.

Allgemein empfehlen wir die Verwendung des gewöhnlichen STRESS-Koeffizienten, da für ihn Schwellenwerte zur Interpretation vorliegen.

Tab. 4.7: Freizeitaktivitäten von Kindern ($n = 2745$)

Ich habe/ich war (in den letzten vierzehn Tagen) ...	Kurzbezeichnung	Anteil in Prozent
mich ausgeruht, gefaulenzt	Ausruhen	36,7
mit Freunden und Freundinnen gespielt	Freunde	79,7
mit meiner Familie etwas unternommen	Familie	58,1
gebastelt, gemalt, gezeichnet	Basteln	41,6
Comichefte angeschaut	Comics	46,3
musiziert, Theater gespielt	Musizieren	26,5
mit Haustieren gespielt	Haustiere	55,3
im Kino gewesen	Kino	10,3
in Konzert, Theater, Ausstellung gewesen	Konzert	16,4
Musik gehört	Musikhören	72,3
Sonntags in die Kirche gegangen	Kirche	47,5
Fernsehen oder Videos angeschaut	Fernsehen	73,5
Computerspiele gespielt	Computersp.	39,1
Bücher gelesen	Buch	69,4
bei Vereinsveranstaltungen	Vereinsv.	13,7
radfahren	Radfahren	85,5
spazieren	Spazieren	60,2
etwas alleine gespielt	alleine Sp.	62,1
auf Partys/Festen	Partys	18,9
Sport betrieben	Sport	63,8

4.6 Freizeitverhalten von Kindern

Als ein konkretes Anwendungsbeispiel soll die dimensionale Struktur von Freizeitaktivitäten untersucht werden. Es soll also eine variablenorientierte Analyse durchgeführt werden. Die Daten sind dem bereits zitierten Forschungsprojekt zur Erfassung der Lebensbedingungen von Kindern entnommen (Wilk und Bacher 1993).³ Die Freizeitaktivitäten wurden in Form einer Liste erhoben. Jedes Kind sollte angeben, welche Freizeitaktivitäten es in den letzten vierzehn Tagen ausgeübt hatte. Die eindimensionale Verteilung der Antworten ist in der Tabelle 4.7 wiedergegeben.

Insgesamt wurden 20 Freizeitaktivitäten erfasst. Als Maß zur Messung der Ähnlichkeit der Freizeitaktivitäten soll der Φ -Koeffizient verwendet werden (zur Berechnung siehe Abschnitt 8.2). Da der Φ -Koeffizient ein Ähnlichkeitsmaß ist, wurden die Werte mit -1 multipliziert, um ein Unähnlichkeitsmaß zu erhalten. Die so gebildete Unähnlichkeitsmatrix ist eine vollständige Unähnlichkeitsmatrix.

³ Detaillierte Ergebnisse über das Freizeitverhalten der Kinder und dessen Determinanten sind in Kirchler und Nagl (1993) dargestellt.

Tab. 4.8: STRESS-Werte für unterschiedliche Dimensionszahlen

Startkonfiguration	Dim. 1	Dim. 2	Dim. 3	Dim. 4
1	0,458	0,249	0,279	0,195
2	0,474	0,216	0,158	0,134
3	0,410	0,311	0,230	0,202
4	0,540	0,219	0,154	0,115
5	0,384	0,356	0,267	0,215

Da 20 Variablen (Objekte) vorliegen, ist entsprechend den Richtwerten von Sixtl eine *Analyse mit maximal vier Dimensionen* sinnvoll. Bei vier Dimensionen ist das bei Sixtl angegebene Mindestverhältnis von 5 : 1 für vollständige Unähnlichkeitsmatrizen erfüllt. Das Verhältnis der Freiheitsgrade liegt bei vier Dimensionen aber unter dem Schwellenwert von 2,5 bei fehlerfreien Daten:

$$r_{df} = \frac{20 \cdot (20 - 1)/2}{4 \cdot (20 - 1) - 4 \cdot (4 - 1)/2} = \frac{110}{70} = 1,57 .$$

Bei der Interpretation der vierdimensionalen Lösung ist daher Vorsicht angebracht. Bei drei Dimensionen beträgt das Verhältnis der Freiheitsgrade

$$r_{df} = \frac{20 \cdot (20 - 1)/2}{3 \cdot (20 - 1) - 3 \cdot (3 - 1)/2} = \frac{110}{54} = 2,04 .$$

und liegt ebenfalls noch unter dem Schwellenwert von 2,5. Auch hier ist also bei der Interpretation noch Vorsicht angebracht. Bei drei und vier Dimensionen ist auf jeden Fall eine Stabilitätsprüfung durchzuführen.

Als Distanzmaß für die gesuchten Konfigurationen soll die euklidische Distanz verwendet werden. Es sollen vier Konfigurationen untersucht werden: (1) ein- (2) zwei- (3) drei- und (4) vierdimensionale Konfiguration für die euklidische Distanz. Für diese vier Konfigurationen ergeben sich die in der Tabelle 4.8 dargestellten STRESS-Werte, wenn fünf zufällige Startkonfigurationen untersucht werden.

Erst bei drei Dimensionen liegt ein STRESS-Wert unter 0,200. Bei vier Dimensionen ergibt sich für die beste Lösung (Startkonfiguration 4) ein STRESS von 0,115. Hier kann entsprechend den in der Tabelle 4.5 auf Seite 85 angegebenen Schwellenwerten von einer ausreichenden Modellanpassung gesprochen werden. Diese Lösung soll hier weiter untersucht werden. Die berechneten Koordinatenwerte sind in Tabelle 4.9 auf der nächsten Seite dargestellt.

Tab. 4.9: Koordinaten für die vierdimensionale Lösung

	Dim. 1	Dim. 2	Dim. 3	Dim. 4
Ausruhen	-0,5818	0,4437	0,4816	-0,4897
Freunde	0,5959	0,7955	0,0398	0,2131
Familie	0,4703	-0,3859	0,0282	-0,1375
Basteln	-0,2711	0,0975	0,6247	0,2057
Comics	-0,4710	0,5178	-0,1819	-0,5964
Musizieren	0,2902	-0,5778	0,7804	0,5761
Haustiere	0,7540	0,3067	0,4380	-0,7758
Kino	-0,8064	-0,7054	-0,2401	-0,6775
Konzert	-0,1440	-0,8393	0,5808	-0,4321
Musikhören	0,0931	0,5368	0,0674	0,1633
Kirche	1,0819	-0,5779	0,1916	0,7967
Fernsehen	-0,2982	0,6034	-0,6069	-0,1740
Computersp.	-0,4181	0,1877	-0,9663	-0,6323
Buch	0,2166	0,2652	0,4760	0,7229
Vereinsv.	0,1525	-0,7092	-0,6953	-0,2809
Radfahren	0,5119	0,3572	-0,5681	0,4379
Spazieren	-0,0620	0,1857	0,4310	0,4346
alleine Sp.	-0,3725	0,5326	0,0230	0,1340
Partys	-0,2670	-0,8832	-0,2975	-0,0113
Sport	-0,4742	-0,1514	-0,6066	-0,0748

4.6.1 Clusteranalytische Interpretation

Wir wollen zunächst untersuchen, ob sich in dem vierdimensionalen Raum eine Clusterstruktur erkennen lässt. Dazu gehen wir analog zur multiplen Korrespondenzanalyse vor (siehe Abschnitt 3.1.2). Die Koordinatenmatrix wird als neue Datenmatrix abgespeichert und in eine hierarchische Clusteranalyse einbezogen. Als Clusteranalyseverfahren wurde wie bei der multiplen Korrespondenzanalyse der Weighted-Average-Linkage verwendet. Die Ergebnisse sind in Abbildung 4.6 in Form eines Dendrogramms dargestellt. Es lassen sich fünf relativ gut interpretierbare Cluster erkennen:

Cluster 1: »Kulturell hoch bewertete« Freizeitaktivitäten (Besuch eines Kinos, Konzert-/Ausstellungs-/Theaterbesuch, Teilnahme an Vereinsveranstaltungen und Partys).

Cluster 2: Freizeitaktivitäten, die gemeinsam mit der Familie unternommen werden (mit der Familie etwas unternehmen, aber auch Spielen mit Haustieren).

Cluster 3: Vorwiegend aktive, häusliche oder außerhäusliche Freizeitaktivitäten (mit Freunden oder Freundinnen spielen, Basteln, Malen, Zeichnen, Musikhören, Bücher lesen, Radfahren, Spazierengehen, etwas alleine spielen, Sport betreiben)

Cluster 4: Eher passive, häusliche Freizeitaktivitäten (Ausruhen, Comics lesen, Fernsehen oder Videoschauen, Computerspielen)

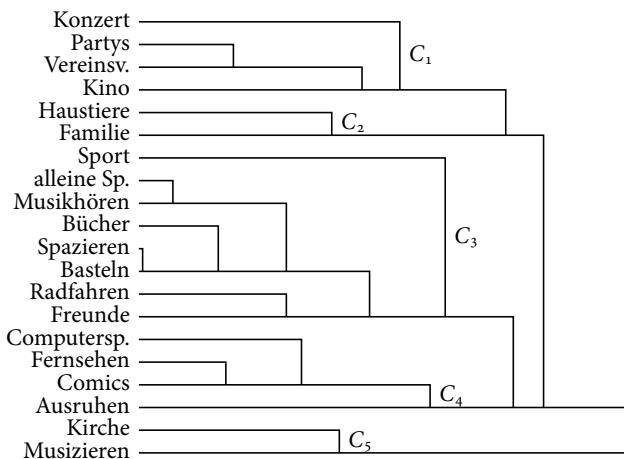


Abb. 4.6: Dendrogramm für eine hierarchische Clusteranalyse der vierdimensionalen Lösung der nichtmetrischen mehrdimensionalen Skalierung

Cluster 5: Aktivitäten »Musizieren, Theaterspielen« und Kirchenbesuch. Es kann vermutet werden, dass das Musizieren und das Theaterspielen in kirchlichen Organisationen stattfindet. Das Cluster soll daher als mit der Kirche verbundene Freizeitaktivität bezeichnet werden.

4.6.2 Faktorenanalytische Interpretation

Eine inhaltliche Interpretation der Dimensionen einer nichtmetrischen mehrdimensionalen Skalierung ist problematisch, da die Dimensionen nach keinem formalen Kriterium angeordnet werden. Wird die euklidische Distanz verwendet, kann durch eine rechtwinklige Rotation der Dimensionen untersucht werden, ob eine Einfachstruktur vorliegt, die dann inhaltlich interpretierbar ist. Wir wollen hier einen anderen Zugang aufzeigen. Dieser besteht darin, dass *inhaltlich begründete Startwerte* für die Dimensionen eingegeben werden. Den untersuchten Freizeitaktivitäten können möglicherweise folgende vier Dimensionen zugrunde liegen:⁴

Dimension 1: Aktive versus passive (konsumptive) Freizeitaktivitäten

Dimension 2: Häusliche versus außerhäusliche Freizeitaktivitäten

Dimension 3: Familiale versus außefamiliale Freizeitaktivitäten

Dimension 4: »Kulturell hoch bewertete« versus »kulturell weniger hoch bewertete« Freizeitaktivitäten

⁴ Siehe dazu auch die wesentlich umfassendere Typologie von Giegler (1985).

Tab. 4.10: Einschätzung der Freizeitaktivitäten auf vier inhaltlichen Dimensionen

Freizeitaktivität	aktiv vs. passiv	häuslich vs. außerhäuslich	familial vs. nichtfamilial	»kult. hoch bewertet« vs. »weniger hoch bewertet«
Ausruhen	-1	1	0	0
Freunde	1	0	0	0
Familie	1	0	1	1
Basteln	1	1	0	1
Comics	1	1	0	-1
Musizieren	1	0	0	1
Haustiere	1	0	1	0
Kino	1	-1	0	0
Konzert	1	-1	0	1
Musikhören	1	1	0	1
Kirche	0	-1	0	1
Fernsehen	-1	1	0	-1
Computersp.	0	1	0	-1
Buch	1	1	0	1
Vereinsv.	1	-1	-1	1
Radfahren	1	-1	0	0
Spazieren	1	-1	1	0
alleine Sp.	1	0	-1	0
Partys	1	-1	0	1
Sport	1	-1	0	0

Kodierung der Dimensionen: »1«: erste Bezeichnung trifft zu, »0«: beide Bezeichnungen treffen zu, »-1«: zweite Bezeichnung trifft zu.

Lesehilfe: »Ausruhen« ist eine passive, häusliche Freizeitaktivität, die entweder gemeinsam mit oder ohne Familie stattfindet und weder kulturell hoch noch gering bewertet wird.

Skaliert man die vorgegebenen Freizeitaktivitäten anhand dieser vier Dimensionen, könnte sich die in der Tabelle 4.10 dargestellte Skalierung für die Startwerte ergeben.

Das Rating wurde in dem Beispiel von den Autoren alleine durchgeführt. In der Praxis ist das Vorliegen mehrerer Ratings wünschenswert. Inwiefern diese Interpretation zutrifft, kann nun überprüft werden, indem die Tabelle 4.10 als Startkonfiguration vorgegeben wird und zur Modellprüfung die Korrelationen der berechneten Dimensionen mit den Startdimensionen berechnet werden.⁵ Die diesbezüglichen Ergebnisse sind in der Tabelle 4.11 dargestellt.

⁵ Dieses Vorgehen lässt sich in Richtung einer konfirmatorischen nichtmetrischen mehrdimensionalen Skalierung verallgemeinern, indem für die Analyse bestimmte Werte fixiert oder indem lineare Restriktionen definiert werden. Die konfirmatorische mehrdimensionale Skalierung ist in Borg (1981, S. 293–338) beschrieben.

Tab. 4.11: Berechnete Koordinaten bei vorgegebenen Startwerten

	Dim. 1	Dim. 2	Dim. 3	Dim. 4
Ausruhen	-0,6899	0,4646	-0,5299	0,0323
Freunde	0,8746	0,4635	0,2223	-0,1806
Familie	0,2458	-0,1434	0,7143	0,2698
Basteln	0,3106	0,5339	-0,1486	0,4557
Comics	-0,1532	0,4983	-0,4953	-0,6600
Musizieren	0,0937	0,0044	-0,1599	1,0414
Haustiere	0,2523	0,0908	1,0618	-0,5096
Kino	0,2925	-1,2177	-0,1053	-0,3339
Konzert	0,4259	-0,6170	-0,4114	0,7815
Musikhören	0,0215	0,6372	0,0859	-0,0334
Kirche	-0,9568	-0,4442	0,4730	0,8132
Fernsehen	-0,4190	0,2683	-0,0966	-0,7202
Computersp.	-0,2612	-0,2936	-0,3886	-0,9575
Buch	0,0437	0,6748	0,3097	0,5468
Vereinsv.	0,0011	-0,6315	-0,9015	0,1544
Radfahren	-0,4392	-0,0661	0,6689	-0,4283
Spazieren	-0,2184	0,3817	0,3714	0,3655
alleine Sp.	-0,1114	0,6320	-0,2756	-0,1838
Partys	0,2904	-0,9062	0,0755	0,1665
Sport	0,3966	-0,3300	-0,4699	-0,6197
γ -Koeffizient	0,882	0,740	0,871	0,919
Prod.-Mom.-Korr.	0,649	0,701	0,731	0,835

grau hinterlegt: Koordinatenwerte mit einem Absolutbetrag größer/gleich 0,6

Die berechneten Koordinaten korrelieren relativ gut mit den Startwerten. Für die erste Dimension ergibt sich ein γ -Koeffizient⁶ von 0,882, für die zweite Dimension von 0,740, für die dritte von 0,871 und für die vierte von 0,919. Die Produkt-Moment-Korrelationskoeffizienten sind etwas geringer, insbesondere für die erste Dimension. Die bei der Wahl der Startwerte vorgenommene inhaltliche Interpretation korreliert somit relativ gut mit den berechneten Dimensionen. Betrachten wir die Koordinatenwerte der Objekte im Detail, so zeigt sich Folgendes: Besonders passive Freizeitaktivitäten sind der Kirchenbesuch und das Ausruhen (Faulenzen), besonders aktiv ist das Spielen mit Freunden. Häusliche Freizeitaktivitäten (Dimension 2) sind insbesondere alleine Spielen, ein Buch lesen und Musikhören, außerhäusliche Freizeitaktivitäten der Besuch von Partys, von Vereinsveranstaltungen, von Kinovorführungen, von Theater- und Konzertaufführungen und von Ausstellungen. Gemeinsame Aktivitäten mit der Familie bzw. mit einzelnen Familienmitgliedern sind gemeinsame Familienunternehmungen und das Spielen mit

⁶ Zur Berechnung des γ -Koeffizienten siehe Abschnitt 8.4.

Haustieren. Diesen stehen auf der anderen Seite Vereinsveranstaltungen gegenüber. Als »kulturell hoch bewertet« gelten schließlich ein Konzertbesuch, der Besuch der Kirche und das Musizieren. Diesem steht das Fernsehen, das Lesen von Comic-Heften, aber auch das Betreiben von Sport als »kulturell weniger hoch bewertete« bzw. »für pädagogisch weniger wertvoll erachtete« Freizeitaktivität gegenüber. Die Zuordnung des Sports zu den »kulturell weniger hoch bewerteten« Freizeitaktivitäten ist problematisch und muss zweifelsohne weiter untersucht werden.

Insgesamt zeigt aber das Beispiel, dass durch die Eingabe theoretisch begründeter Startwerte eine faktorenanalytische Interpretation gefunden werden kann. Zu beachten ist bei der Interpretation, dass die Skalenwerte auf einer Dimension nur *relativ* interpretiert werden können. Ein negativer Wert für eine Freizeitaktivität auf der Dimension 1 beispielsweise bedeutet nur, dass diese Aktivität in Relation zu den untersuchten Aktivitäten eher passiv ist.

4.6.3 Freizeitaktivitäten und Sozialstruktur

Wir wollen nun darstellen, wie sich innerhalb der nichtmetrischen mehrdimensionalen Skalierung der Zusammenhang zwischen Freizeitaktivitäten und Sozialstruktur untersuchen lässt. Das Grundprinzip besteht darin, dass weitere Informationen über die untersuchten Objekte in die berechnete Punktekonfiguration hineinprojiziert werden. Für die zu untersuchende Fragestellung können folgende Informationen ausgewählt werden:

- Die mittlere abgeschlossene Schulbildung der Mütter der einzelnen Freizeitaktivitäten. Ein höherer Mittelwert für eine Freizeitaktivität bedeutet, dass sie eher bei einer höheren Bildung der Mutter ausgeübt wird.
- Die mittlere abgeschlossene Schulbildung der Partner der Mütter der einzelnen Freizeitaktivitäten. Die Mittelwerte sind analog zu oben zu interpretieren.
- Die mittlere Gemeindegröße der einzelnen Freizeitaktivitäten. Ein höherer Mittelwert bedeutet, dass die Freizeitaktivität eher in Städten ausgeübt wird.
- Das mittlere Einkommen der Mütter der einzelnen Freizeitaktivitäten. Ein höherer Mittelwert bedeutet, dass die Freizeitaktivität eher bei einem höheren Einkommen der Mutter ausgeübt wird.
- Das mittlere Einkommen der Partner der einzelnen Freizeitaktivitäten. Die Mittelwerte sind analog zu oben zu interpretieren.
- Der Jungenanteil der einzelnen Freizeitaktivitäten. Ein höherer Anteil bedeutet, dass die Freizeitaktivität eher von Jungen ausgeübt wird.
- Freizeitpräferenzen von Kindern, deren Väter selbständige Landwirte sind.

Tab. 4.12: Sozialstrukturelle Informationen über die untersuchten Objekte (Freizeitaktivitäten)

	SchM	SchP	Gemgr	EinkM	EinkP	Geschl	Landw	Selbs	mittAng
Ausruhen	1,5799	1,5806	1,7578	1,4766	2,3549	0,4960	4,34	4,12	3,72
Freunde	1,5708	1,5793	1,7195	1,4856	2,3500	0,4837	6,91	7,87	8,30
Familie	1,5967	1,5856	1,7908	1,4916	2,3619	0,4933	3,99	5,72	6,24
Basteln	1,5982	1,5780	1,7525	1,4879	2,3284	0,4563	3,72	4,43	4,15
Comics	1,5950	1,5565	1,8075	1,5027	2,3298	0,5702	4,16	4,92	4,67
Musizieren	1,6755	1,7115	1,6696	1,4743	2,3846	0,3771	3,10	3,20	3,15
Haustiere	1,5285	1,5431	1,6569	1,4915	2,3186	0,4868	7,79	5,29	5,52
Kino	1,5684	1,5322	1,9290	1,6339	2,3239	0,6996	0,53	0,67	1,11
Konzert	1,6803	1,6637	1,6953	1,5015	2,3710	0,5133	1,33	1,35	1,98
Musikhören	1,5679	1,5554	1,7032	1,4751	2,3308	0,4782	7,35	6,89	7,32
Kirche	1,5298	1,5385	1,4947	1,4237	2,2834	0,4848	8,95	5,10	4,80
Ferns.	1,5544	1,5412	1,7308	1,4948	2,3224	0,5231	6,73	7,13	7,32
Computersp.	1,6219	1,5974	1,8078	1,5558	2,3843	0,6835	2,13	4,74	4,02
Buch	1,6004	1,6073	1,7140	1,4896	2,3519	0,4368	7,62	7,32	7,39
Vereinsv.	1,5421	1,5772	1,6007	1,5203	2,3722	0,6542	0,70	1,78	1,46
Radfahren	1,5540	1,5413	1,6576	1,4742	2,3105	0,5116	9,65	8,49	8,97
Spazieren	1,5614	1,5556	1,7048	1,4750	2,3221	0,4647	6,47	5,78	5,37
alleine Sp.	1,5681	1,5546	1,6887	1,4963	2,3375	0,5309	6,55	6,33	5,98
Partys	1,5662	1,5275	1,7216	1,5322	2,3368	0,5545	1,42	1,91	1,89
Sport	1,6053	1,5978	1,7139	1,5031	2,3507	0,5598	6,55	6,95	6,65

Abkürzungen: SchM, SchP: abgeschlossene Schulbildung der Mutter / des Partners (höherer Wert: höhere Schulbildung); EinkM, EinkP: Einkommen der Mutter / des Partners (höherer Wert: höheres Einkommen); Gemgr: Gemeindegröße (höherer Wert: größere Gemeinde); Geschl: Geschlechterproportion (höherer Wert: höherer Anteil an Jungen); Landw, Selbs, mittAng: Freizeitpräferenzen von Kindern, deren Väter selbständige Landwirte / selbständige Gewerbetreibende oder Unternehmer / mittlere Angestellte oder Beamte sind.

- Freizeitpräferenzen von Kindern, deren Väter selbständige Unternehmer sind.
- Freizeitpräferenzen von Kindern, deren Väter mittlere Angestellte oder Beamte sind.

Die Interpretation der in Tabelle 4.12 dargestellten Werte soll exemplarisch für die beiden Variablen »BildM« und »Landw« verdeutlicht werden. Die Freizeitaktivität Ausruhen hat in der Variablen »SchM« (abgeschlossene Schulbildung der Mutter) einen Mittelwert von 1,5799. Dieser Mittelwert unterscheidet sich nur geringfügig von jenem der nachfolgenden Freizeitaktivität des Spielens mit Freunden (Mittelwert von 1,5708). Der Besuch eines Konzertes, einer Ausstellung und/oder einer Theateraufführung (Konzert) hat mit 1,6803 den höchsten Mittelwert. Diese Freizeitaktivität wird also eher bei einer höheren Schulbildung der Mutter ausgeübt als jene des Ausruhens. Die Werte für die Variablen »SchP« bis »Geschl« sind analog zu interpretieren.

Die Freizeitpräferenzen sind dagegen anders zu interpretieren. Sie sind gleich den Spaltenprozenten der Tabelle mit den Freizeitaktivitäten in den Zeilen und den Berufen in

den Spalten: 4,34 Prozent aller Freizeitaktivitäten der Kinder mit einem selbständigen Landwirt als Vater entfallen auf das Ausruhen, 6,91 Prozent auf das Spielen mit Freunden usw.

Die in der Tabelle 4.12 auf der vorherigen Seite enthaltenen Variablen können nun in die berechnete Punktekonfiguration hineinprojiziert werden. Dazu stehen zwei Techniken zur Verfügung:

- *Vektorrepräsentation:*⁷ Die Variable soll als Vektor in die Punktekonfiguration hineinprojiziert werden. Die Richtung des Vektors wird dabei so bestimmt, dass die Projektionen der Objekte mit den Variablenwerten maximal korrelieren.⁸
- *Idealpunktrepräsentation:*⁹ Die Variable soll als Punkt in die Punktekonfiguration hineinprojiziert werden. Die Lage des Punktes wird so bestimmt, dass die Distanzen der Objekte zu dem Idealpunkt mit den Variablenwerten bestmöglich übereinstimmen.

Abbildung 4.7 veranschaulicht diese Zielsetzungen. Für die Variable V_1 wird ein Vektor gesucht. Die Objekte A bis D sollen in V_1 die Werte 1, 2, 3, 4 haben. Wenn wir die rechtwinkligen Projektionen auf V_1 betrachten, so wird diese Ordnungsrelation abgebildet. Die Vektorrepräsentation V_1^* erfüllt dagegen nicht diese Voraussetzungen. Bei der Idealpunktrepräsentation wird nicht nach einem Vektor gesucht, sondern nach einer Darstellung der Variablen als Punkt. Wir wollen annehmen, dass in V_2 die Präferenzen einer Person für die vier Objekte stehen. Diese sind $B > A > D > C$ (A wird B bevorzugt usw.). Die Idealpunktrepräsentation V_2 erfüllt diese Bedingung, die Idealpunktrepräsentation V_2^* dagegen nicht. Bei wenigen Objekten gibt es für beide Modelle keine eindeutige Lösung. So bilden zum Beispiel alle in der grauen Fläche liegenden Vektoren die vorgegebene empirische Ordnungsrelation von V_1 ab. Auch alle in der grauen Fläche liegenden Idealpunkte bilden die empirische Präferenz von V_2 ab. Diese Nichteindeutigkeit nimmt mit der Anzahl der Objekte ab. Bei der Vektorrepräsentation wird ferner die Zahl möglicher Lösungen dadurch eingegrenzt, dass nicht nur die ordinale Information der Variablen, sondern ihre metrische Information abgebildet werden soll.

Formal kann die Aufgabenstellung der Vektorrepräsentation wie folgt ausgedrückt werden: Für eine Variable V ist ein Richtungsvektor y_V mit den Koordinaten y_{Vh} in den Dimensionen h gesucht. Dieser soll so bestimmt werden, dass die rechtwinkligen Projek-

⁷ Das Vektormodell wurde ursprünglich von Tucker und Messick (1963) zur Analyse von Präferenzen entwickelt.

⁸ Siehe dazu auch Holtmann (1975).

⁹ Dieses Modell geht auf das Unfoldingmodell von Coombs (1964) zurück. Die Beziehung zwischen Idealpunktmodell und Vektormodell wird in Coombs (1975) analysiert.

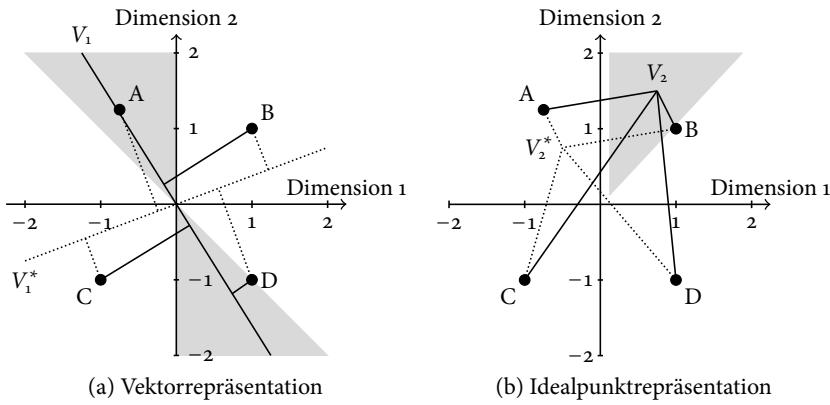


Abb. 4.7: Veranschaulichung der Vektor- und Idealpunktrepräsentation (schematische Darstellung)

tionen der Objekte mit den Variablenwerten maximal korrelieren. Die Maximierungsaufgabe ist also:

$$\text{KORR}(y_i, \hat{y}_i) = \frac{s(y_i, \hat{y}_i)}{\sqrt{s^2(\hat{y}_i)}} = \frac{\sum_i y_i \cdot \hat{y}_i}{\left[\sum_i \hat{y}_i^2 - m \cdot \bar{\hat{y}}^2 \right]^{1/2}} \rightarrow \text{Maximum ,} \quad (4.8)$$

wobei \hat{y}_i die Projektion des Objekts i auf den Vektor der Variablen ist. Die Projektionen berechnen sich wie folgt:

$$\hat{y}_i = \sum_h x_{ih} \cdot y_{Vh} .$$

Die unbekannten Größen in der Maximierungsaufgabe (Gleichung 4.8) sind die Koordinatenwerte y_{Vh} des Richtungsvektors. Sie lassen sich iterativ wiederum mit der Gradientenmethode bestimmen:

$$y_{Vh}^{\text{neu}} = y_{Vh}^{\text{alt}} + \alpha \cdot \text{grad}(y_{Vh}^{\text{alt}}) .$$

Neben der Maximierung der Produkt-Moment-Korrelation wurden noch eine Reihe weiterer Maximierungsfunktionen für die Projektionen in die nichtmetrische mehrdimensionale Skalierung eingeführt (Coxon 1982, S. 106–109, 172–175).

Bei der Idealpunktrepräsentation können die Koordinatenwerte x_{Ih} des Idealpunktes auf den Dimensionen analog den Koordinatenwerten der untersuchten Objekte berechnet werden, indem der STRESS

$$\text{STRESS}(q, p)_I = \sqrt{\frac{\sum_i (d(q, p)_{i,I} - \hat{d}(q, p)_{i,I})^2}{\sum_i (\hat{d}(q, p)_{i,I})^2}}$$

Tab. 4.13: Korrelationen für die Vektorrepräsentation und STRESS-Werte für die Idealpunktrepräsentation

Variable	Korrelation	Variable	STRESS
SchM	0,441	Landw	0,114
SchP	0,486	Selbs	0,151
Gemgr	0,680	mittAng	0,143
EinkM	0,839		
EinkP	0,406		
Geschl	0,741		

Abkürzungen: siehe Tabelle 4.12 auf Seite 101

mit Hilfe der Gradientenmethode minimiert wird (siehe Formel 4.6 auf Seite 85).

Ob für eine Variable eine Vektor- oder Idealpunktrepräsentation gesucht werden soll, hängt von der Interpretation der Variablen ab:

- Lässt sich die untersuchte Variable als »Variable« interpretieren, in der die untersuchten Objekte bestimmte Ausprägungen haben, ist eine Vektorrepräsentation sinnvoll.
- Lässt sich die untersuchte Variable als »neues Objekt« interpretieren, das zusätzlich zu den anderen Objekten räumlich dargestellt werden soll, ist eine Idealpunktrepräsentation sinnvoll. Dies ist insbesondere der Fall, wenn Präferenzen vorliegen.

In unserem Beispiel soll für die Variablen »SchM« bis »Geschl« eine Vektorrepräsentation gesucht werden, da hier eine Interpretation als Variable angemessener ist. Für die Freizeitpräferenzen soll dagegen eine Idealpunktrepräsentation gesucht werden. Es ergeben sich die in der Tabelle 4.13 dargestellten Korrelationen und STRESS-Werte. Relativ hohe Korrelationen (größer 0,60) ergeben sich für die Variablen »Geschl« (Geschlecht), »EinkM« (Einkommen der Mutter) und »Gemgr« (Gemeindegröße). Für die anderen Variablen (»SchM«, »SchP«, »EinkP«) ist eine Vektorrepräsentation weniger gut geeignet. Zwischen Freizeitaktivitäten und diesen Variablen besteht auch bivariat kein bzw. nur ein schwacher Zusammenhang.

Die Interpretation der Vektorrepräsentation soll hier nur exemplarisch für die Variable »Gemeindegröße« dargestellt werden (siehe Tabelle 4.14). Aufgrund der Projektionen ergibt sich folgende Ausdifferenzierung der Freizeitaktivitäten in Abhängigkeit von der Gemeindegröße: »Typische« Land-Freizeitaktivitäten sind das Musizieren, ein Kirchenbesuch und ein Spaziergang, »typische« Stadt-Freizeitaktivitäten das Lesen von Comics, ein Kinobesuch und Computerspiele.

Auch die Idealpunktrepräsentation soll exemplarisch für die Kinder, deren Väter als selbständige Landwirte erwerbstätig sind, diskutiert werden (siehe Tabelle 4.15 auf Seite 106). Ein größerer empirischer Zahlenwert bedeutet eine stärkere Präferenz. Folglich

Tab. 4.14: Projektionen der Objekte auf die Variable »Gemeindegröße«

Objekt	emp. Wert	Projektion	rechtwinklige Distanz
Ausruhen	1,758	0,294	0,957
Freunde	1,720	-0,234	1,009
Familie	1,791	-0,441	0,837
Basteln	1,753	-0,243	0,582
Comics	1,808	0,773	0,567
Musizieren	1,670	-0,728	0,849
Haustiere	1,657	-0,342	1,138
Kino	1,929	0,744	0,915
Konzert	1,695	-0,324	1,047
Musikhören	1,703	0,029	0,693
Kirche	1,495	-1,389	0,518
Fernsehen	1,731	0,844	0,403
Computersp.	1,808	1,054	0,388
Buch	1,714	-0,441	0,745
Vereinsv.	1,601	-0,191	1,034
Radfahren	1,658	0,292	0,877
Spazieren	1,705	-0,507	0,537
alleine Sp.	1,689	0,305	0,595
Partys	1,722	0,140	0,963
Sport	1,714	0,364	0,927

grau hinterlegt: Projektionen mit einem Absolutbetrag größer 0,500

bedeutet auch eine größere Distanz eine stärkere Präferenz. Wenn wir die erwarteten Distanzen betrachten, so können sechs Gruppen von Freizeitpräferenzen unterschieden werden. Die stärkste »Präferenz« liegt für das Radfahren und einen Kirchenbesuch vor. Daran anschließend folgt eine Reihe von Freizeitaktivitäten mit nur einer geringfügig geringeren Präferenz (erwartete Distanz = 1,573). Auch für das Basteln liegt noch annähernd die gleiche Präferenz vor. Mit deutlich geringerer Präferenz werden die anderen Freizeitaktivitäten ausgeübt.

Wie wir dem Beispiel entnehmen können, hängt die Interpretation der berechneten und erwarteten Distanzen von der Kodierung der Variablen ab. Bedeutet ein größerer empirischer Zahlenwert eine stärkere Präferenz, so bedeutet auch eine größere Distanz eine stärkere Präferenz. Bei einer graphischen Darstellung des Idealpunktes sind dann dessen Koordinaten mit -1 zu multiplizieren, damit der Idealpunkt nahe den Objekten mit der stärksten Präferenz liegt. Für die drei untersuchten Freizeitpräferenzen ergeben sich nach einer Multiplikation mit -1 die in der Tabelle 4.16 auf Seite 107 dargestellten Idealpunkte. Kinder, deren Väter selbstständig oder freiberuflich erwerbstätig sind, sind in einem größeren Ausmaß in Richtung aktiver, familialer und »kulturell höher bewerteter«

Tab. 4.15: *Idealpunktrepräsentation der Freizeitpräferenzen der Kinder mit Vätern, die als selbständige Landwirte tätig sind*

	empirische Präferenz ^{a)}	berechnete Distanz zu Idealpunkt	erwartete Distanz zu Idealpunkt
Kino	0,530	0,812	0,511
Vereinsv.	0,700	0,210	0,511
Konzert	1,330	0,916	0,667
Partys	1,420	0,418	0,667
Computersp.	2,130	1,339	1,268
Musizieren	3,100	1,198	1,268
Basteln	3,720	1,505	1,505
Familie	3,990	1,651	1,573
Comics	4,160	1,612	1,573
Ausruhen	4,340	1,882	1,573
Spazieren	6,470	1,574	1,573
alleine Sp.	6,550	1,638	1,573
Sport	6,550	1,494	1,573
Fernsehen	6,730	1,579	1,573
Freunde	6,910	1,452	1,573
Musikhören	7,350	1,565	1,573
Buch	7,620	1,535	1,573
Haustiere	7,790	1,316	1,573
Kirche	8,950	1,774	1,694
Radfahren	9,650	1,613	1,694

a) Ein größerer Zahlenwert bedeutet eine stärkere Präferenz.

Freizeitaktivitäten hin orientiert als Kinder von Landwirten oder mittleren Angestellten. Ihre Freizeitaktivitäten sind dagegen in einem geringeren Ausmaß häuslich orientiert. Die Unterschiede sind aber sehr gering, so dass sich im Wesentlichen dieselben Freizeitpräferenzen ergeben (siehe Tabelle 4.17 auf Seite 108). Allgemein sollten Unterschiede in den Idealpunkten nur dann interpretiert werden, wenn sich auch deutliche Unterschiede in den berechneten und erwarteten Distanzen ergeben. In unserem Beispiel ist dies nicht der Fall. Von einer Interpretation der Idealpunkte sollte daher Abstand genommen werden.¹⁰

¹⁰ Auch eine (nicht dargestellte) bivariate Korrespondenzanalyse erbringt das Ergebnis, dass Freizeitpräferenzen von Zehnjährigen nur schwach mit sozialstrukturellen Variablen korrelieren. Eine Ursache dafür ist in der sehr groben Messung der Freizeitaktivitäten zu sehen, da die Aktivitäten sehr allgemein formuliert waren und nur erfragt wurde, ob eine bestimmte Tätigkeit in den letzten vierzehn Tagen ausgeübt wurde, und nicht erfragt wurde, wie häufig dies geschah.

Tab. 4.16: *Idealpunkte der drei untersuchten Freizeitpräferenzen*

	Dim. 1 aktiv vs. passiv	Dim. 2 häuslich vs. außerhäuslich	Dim. 3 familial vs. nicht familial	Dim. 4 »kulturell« vs. »nicht kulturell«
Landw	0,23	0,94	0,08	0,26
Selbs	0,59	0,50	0,37	0,61
mittAng	0,04	1,00	0,06	0,32

Anmerkung: Ein größerer Zahlenwert bedeutet, dass die erste Dimensionsbezeichnung in einem größeren Ausmaß zutrifft.

4.7 Anwendungsempfehlungen

1. Die nichtmetrische mehrdimensionale Skalierung kann für direkt und indirekt erhobene Ähnlichkeitsurteile eingesetzt werden. Es wird nur die ordinale Information der Ähnlichkeiten zwischen den Objektpaaren verwendet (siehe Abschnitt 4.2).
2. Die Zahl der Dimensionen sollte nicht zu groß gewählt werden (siehe Abschnitt 4.3).
3. Der STRESS für eine ausgewählte Konfiguration sollte möglichst niedrig sein (siehe Tabelle 4.5 auf Seite 85).
4. Zur Stabilitätsprüfung und zur Vermeidung von lokalen Minima sollte mit unterschiedlichen zufälligen Startkonfigurationen gerechnet werden (siehe Abschnitt 4.3).
5. Zur inhaltlichen Interpretation können inhaltlich begründete Startwerte eingesetzt werden (siehe Abschnitt 4.6.2).
6. Um Zusammenhänge mit externen Variablen aufzuzeigen, können Vektor- oder Idealpunktrepräsentationen berechnet werden (siehe Abschnitt 4.6.3).

Tab. 4.17: Erwartete Präferenzen in Abhängigkeit von der Berufsposition des Vaters

Kinder von Landwirten	Kinder von Selbständigen	Kinder von mittleren Angestellten und Beamten
Kino, Vereinsv. (0,511)	Kino, Vereinsv. (0,494)	Kino, Konzert (0,654)
Konzert, Partys (0,667)	Partys (0,529)	Vereinsv., Partys (0,794)
Computersp., (1,268)	Musizieren Konzert (1,025)	Musizieren (0,977)
Basteln (1,505)	Musizieren (1,218)	
Familie, Comics, Ausruhen, Spazierengehen, alleine Spielen, Sport, Fernsehen, Radfahren, Freunde, Musikhören, Buch, Haustiere (1,573)	Comics, Ausruhen, Basteln, Spazierengehen, Kirche, Haustiere (1,5593)	Familie, Comics, Basteln, Ausruhen, Computersp., Spazierengehen, alleine Spielen, Sport, Fernsehen, Musikhören, Kirche, Buch, Haustiere (1,573)
Kirche, Radfahren (1,694)	Familie, Ausruhen, alleine Spielen, Sport, Fernsehen, Freunde, Musikhören, Buch, (1,603)	Freunde, Radfahren (1,655)
	Radfahren (1,716)	

Anmerkung: Zahlenwerte in Klammern entsprechen der erwarteten Distanz (größerer Zahlenwert: stärkere Präferenz).

5 Weitere räumliche Darstellungsverfahren

In diesem Kapitel werden drei Alternativen zu den beiden behandelten Verfahren dargestellt. Diese Alternativen werden allerdings weniger ausführlich besprochen, was bedeutet, dass auf ein Durchrechnen des Kalküls anhand eines Beispiels verzichtet wird.

5.1 Die bivariate Korrespondenzanalyse

Bei der *bivariaten Korrespondenzanalyse* wird eine räumliche Darstellung einer zweidimensionalen Tabelle gesucht, wobei die Zeilen und Spalten auch aus Variablengruppen bestehen können. Greifen wir die in den beiden vorausgehenden Kapiteln 3 und 4 behandelten Beispiele auf, lassen sich die in Tabelle 5.1 dargestellten Anwendungsbeispiele für die bivariate Korrespondenzanalyse anführen.

Das Vorgehen der bivariaten Korrespondenzanalyse ist ähnlich jenem der multiplen Korrespondenzanalyse. Im Unterschied zu dieser wird bei der bivariaten Korrespondenzanalyse aber nicht die gesamte Zusammenhangsmatrix \mathbf{G} einbezogen, sondern nur jener Teil, der von den Zeilen- und Spaltenvariablen gebildet wird (siehe Abbildung 5.1 auf der nächsten Seite). Diese in die Analyse einbezogene *Teilmatrix* der Zusammenhangsmatrix \mathbf{G} soll im Folgenden mit $\tilde{\mathbf{G}}$ bezeichnet werden. Die Zeilenvariable(n) werden mit i und ihre Ausprägungen mit k bezeichnet, die Spaltenvariable(n) mit j und ihre Ausprägungen mit l .

Tab. 5.1: Anwendungsbeispiele für die bivariate Korrespondenzanalyse

Beispiel	1. Variablengruppe (Zeile)	2. Variablengruppe (Spalte)
Sozialstrukturanalyse	abgeschl. Schulbildung der Mutter	abgeschl. Schulbildung des Vaters
Freizeitsoziologie	sozialstrukturelle Merkmale	Freizeitaktivitäten

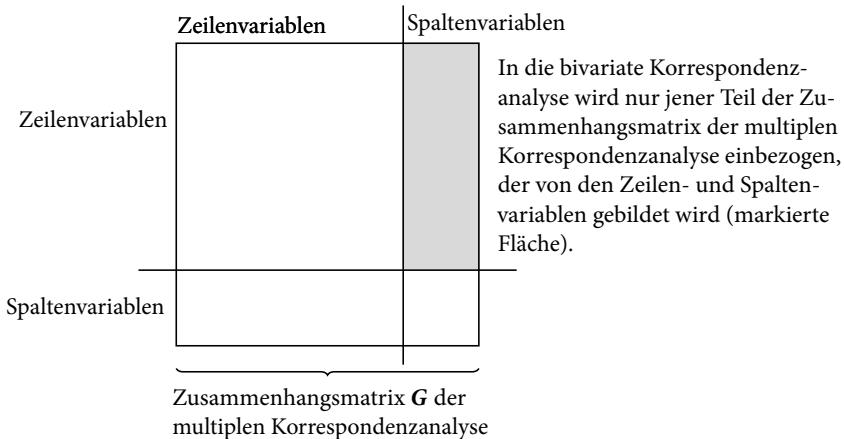


Abb. 5.1: Zusammenhangsmatrix der bivariaten Korrespondenzanalyse

gen mit k^* . Für die Matrix \tilde{G} wird analog zu der multiplen Korrespondenzanalyse eine Eigenwertzerlegung durchgeführt:

$$\tilde{G} = \tilde{V} \cdot \tilde{D} \cdot \tilde{U}^T,$$

wobei \tilde{V} die Matrix der Links-Eigenvektoren mit den Elementen $\tilde{v}_{i(k)h}$ (Wert der Ausprägung k der Zeilenvariablen i auf der Dimension h) ist. \tilde{D} ist die Diagonalmatrix der Eigenwerte mit den Eigenwerten \tilde{d}_h in der Diagonalen. \tilde{U} schließlich ist die Matrix der Rechts-Eigenvektoren mit den Elementen $\tilde{u}_{j(k^*)h}$. Die Unterscheidung in Links- und Rechts-Eigenvektoren ist erforderlich, da die untersuchte Matrix \tilde{G} nicht symmetrisch ist. Zur Berechnung der Koordinatenwerte für die räumliche Darstellung wurden unterschiedliche Reskalierungen der Eigenvektoren entwickelt. Die Skalierungsarten unterscheiden sich darin, wie die Eigenwerte in die Berechnung der Koordinatenwerte einbezogen werden (siehe Tabelle 5.2 sowie Carroll, Green u. a. 1986; SPSS Inc. 1990b, B21–B47).

Die allgemeine Formel zur Zeilenskalierung lautet:

$$\tilde{x}_{i(k)h} = \frac{1}{\sqrt{p_{i(k)}}} \cdot \tilde{v}_{i(k)h} \cdot \tilde{z}_h,$$

mit \tilde{z}_h als Skalierungsfaktor. Analog lautet die Formel zur Spaltenskalierung:

$$\tilde{y}_{j(k^*)h} = \frac{1}{\sqrt{p_{j(k^*)}}} \cdot \tilde{u}_{j(k^*)h} \cdot \tilde{s}_h,$$

mit \tilde{s}_h als Skalierungsfaktor. Die Reskalierungsansätze unterscheiden sich in der Wahl der Skalierungsfaktoren \tilde{z}_h und \tilde{s}_h .

Tab. 5.2: Reskalierungsansätze bei der bivariaten Korrespondenzanalyse

Name der Skalierung	Skalierung der Zeilenvariable(n) i Skalierungsfaktor \tilde{z}_h	Skalierung der Spaltenvariable(n) j Skalierungsfaktor \tilde{s}_h
Zeilenskalierung	\tilde{d}_h	1
Spaltenskalierung	1	\tilde{d}_h
kanonische Skalierung	$\sqrt{\tilde{d}_h}$	$\sqrt{\tilde{d}_h}$
Hauptkomponentenskalierung	\tilde{d}_h	\tilde{d}_h
multiple Korrespondenzanalyseskalierung	$\sqrt{\tilde{d}_h + 1}$	$\sqrt{\tilde{d}_h + 1}$

Die multiple Korrespondenzanalyseskalierung führt dazu, dass die Ergebnisse der bivariaten Korrespondenzanalyse mit jenen der multiplen Korrespondenzanalyse mit Ausnahme des konstanten Skalierungsfaktors $1/\sqrt{m}$ (m = Anzahl der Variablen) übereinstimmen. Es gilt (Andersen 1991, S. 390; Carroll, Green u. a. 1986, S. 275):

$$\tilde{x}_{i(k)h} = \sqrt{m} \cdot x_{i(k)h}$$

mit $\tilde{x}_{i(k)h}$ als Koordinatenwerte aus der bivariaten Korrespondenzanalyse (multiple Korrespondenzanalyseskalierung) und $x_{i(k)h}$ als Koordinatenwerte aus der multiplen Korrespondenzanalyse. Die einzelnen Skalierungsarten haben unterschiedliche Konsequenzen hinsichtlich der Interpretierbarkeit der Distanzen (Carroll, Green u. a. 1986; SPSS Inc. 1990b, B21–B47):

- Bei der Zeilenskalierung sind nur die Distanzen zwischen den Ausprägungen der Zeilenvariablen interpretierbar.
- Bei der Spaltenskalierung sind nur die Distanzen zwischen den Ausprägungen der Spaltenvariablen interpretierbar.
- Bei der kanonischen Skalierung sind nur die Distanzen zwischen den Ausprägungen der Zeilen- und Spaltenvariablen, nicht aber jene innerhalb der Zeilen- und Spaltenvariablen interpretierbar.
- Bei der Hauptkomponentenskalierung sind nur die Distanzen zwischen den Ausprägungen innerhalb der Zeilen- und Spaltenvariablen interpretierbar, nicht aber jene zwischen den Zeilen- und Spaltenvariablen.
- Bei der multiplen Korrespondenzanalyseskalierung sind alle Distanzen interpretierbar.

Nur die multiple Korrespondenzanalyseskalierung erlaubt also die Interpretation aller Distanzen. Soll eine clusteranalytische Interpretation der Zeilen- und Spaltenvariablen durchgeführt werden, ist sie daher zu wählen. Technisch bedeutet dies, dass anstelle einer bivariaten Korrespondenzanalyse eine multiple Korrespondenzanalyse durchgeführt

werden muss, da diese Skalierungsart in Programmen zu der bivariaten Korrespondenzanalyse in der Regel nicht zur Verfügung steht. Sie hat aber den Nachteil, dass die Dimensionszahl für eine faktoren- und clusteranalytische Interpretation nicht identisch ist. Dies ist dagegen bei den anderen Skalierungsarten der Fall.

Allgemein kann man sich bei der Entscheidung, ob eine multiple oder bivariate Korrespondenzanalyse durchgeführt werden soll, an folgenden Regeln orientieren:

1. Sollen alle Distanzen zwischen den Ausprägungen der untersuchten nominalen Variablen interpretiert werden, ist eine multiple Korrespondenzanalyse durchzuführen.
2. Ist eine Unterscheidung in zwei Variablengruppen nicht sinnvoll, ist ebenfalls eine multiple Korrespondenzanalyse durchzuführen.
3. Ist eine Unterscheidung in zwei Variablengruppen dagegen sinnvoll und sollen nicht alle Distanzen interpretiert werden, sollte eine bivariate Korrespondenzanalyse gerechnet werden.

Bei der Auswahl der geeigneten Skalierung können folgende Regeln hilfreich sein:

- Soll der Zusammenhang zwischen den Zeilen- und Spaltenvariablen durch deren räumliche Nähe interpretiert werden, wird man sich für die kanonische Skalierung entscheiden.
- Sollen die Ausprägungen der Zeilenvariablen geclustert werden, wird man eine Zeilenskalierung durchführen. Die Spaltenvariablen kann man dann zur Beschreibung der gebildeten Cluster verwenden.
- Sollen umgekehrt die Spaltenvariablen geclustert werden, wird man eine Spaltenskalierung durchführen. Die Zeilenvariablen können dann zur Beschreibung der gebildeten Cluster verwendet werden.
- Soll keine Clusterung von Spalten- und Zeilenvariablen durchgeführt werden, ist aber eine Unterscheidung in abhängige und unabhängige Variablen möglich, wird man in Richtung der abhängigen Variablen skalieren und die unabhängigen Variablen zur Namensgebung der Dimensionen verwenden. Bilden also die Spaltenvariablen die abhängigen Variablen, wird man eine Spaltenskalierung durchführen, andernfalls eine Zeilenskalierung.

Welche Interpretation den Daten angemessen ist, hängt neben der inhaltlichen Fragestellung von den Ergebnissen ab. In der Praxis bedeutet dies, dass man mehrere Skalierungsmöglichkeiten durchprobieren wird. Wir wollen das Vorgehen am Beispiel der Zusammenhangsmatrix zwischen der abgeschlossenen Schulbildung von Müttern und jener ihrer Partner veranschaulichen. Die abgeschlossene Schulbildung der Partner soll

dabei die Zeilenvariable bilden, jene der Mütter die Spaltenvariable. Die Ergebnisse einer kanonischen Skalierung und einer Spaltenskalierung sind in der Tabelle 5.3 auf der nächsten Seite dargestellt.

Die Eigenwerte lassen sich bei der bivariaten Korrespondenzanalyse – im Unterschied zu der multiplen Korrespondenzanalyse – direkt als erklärte χ^2 -Beiträge interpretieren, da hier die Summe aller Eigenwerte gleich dem mittleren χ^2 -Wert ist (Greenacre 1989, S. 86, 91):

$$\frac{\chi^2}{n} = \sum_h^{q_{\max}} \tilde{d}_h ,$$

wobei χ^2 der χ^2 -Wert der untersuchten Tabelle ist. Die Fallzahl wird mit n bezeichnet und q_{\max} stellt die maximale Dimensionszahl dar. Die maximale Dimensionszahl ist gleich dem Minimum aus der Zahl der Zeilen- und Spaltenausprägungen minus 1. Der mittlere χ^2 -Wert χ^2/n wird in der Korrespondenzanalyseliteratur als Trägheit (»inertia«) bezeichnet. In unserem Beispiel ist der χ^2 -Wert für die Zusammenhangstabelle der beiden Variablen gleich 1 055,4710 (siehe Tabelle 3.11 auf Seite 57). In die Berechnung gingen 1 823 Fälle ein. Die Trägheit beträgt daher 0,579. Die erste Dimension erklärt somit 54,1 Prozent des χ^2 -Wertes ($100 \cdot 0,313/0,579$), die zweite 26,4 Prozent ($100 \cdot 0,153/0,579$). Die beiden Dimensionen zusammen erklären also 80,5 Prozent (54,1 + 26,4 Prozent) des χ^2 -Wertes bzw. der Trägheit. Eine zweidimensionale Lösung erbringt eine sehr gute Modellanpassung.

Stellen wir die *kanonische Skalierung* graphisch dar, ergibt sich das in der Abbildung 5.2 auf der nächsten Seite verdeutlichte Bild. Die kanonische Skalierung ermöglicht eine Interpretation der Distanzen zwischen den Ausprägungen unterschiedlicher nominaler Variablen. Es können also die Distanzen zwischen A_1 und B_1 , zwischen A_1 und B_2 usw. interpretiert werden. Eine geringere Distanz zwischen zwei Ausprägungen bedeutet, dass Heiraten zwischen diesen Ausprägungen häufiger auftreten. Wird unter diesem Gesichtspunkt die Abbildung 5.2 auf der nächsten Seite gelesen, so zeigt sich unter anderem, dass Heiraten zwischen Personen mit einer höheren Schulbildung (A_7 und B_7) häufiger auftreten als zwischen B_7 und A_6 oder B_7 und A_5 usw. Wenn wir nur die Distanzen zwischen den jeweils gleichen Bildungsabschlüssen vergleichen, so liegt für die unteren Bildungsabschlüsse (A_2, B_2, A_3, B_3) eine größere Homogamie (geringere Distanz) als zwischen den anderen Schichten vor.

Während wir bei der kanonischen Skalierung an einer Interpretation der Distanzen der Ausprägungen zwischen den analysierten Variablen interessiert sind, liegt der *Spaltenskalierung* die Fragestellung zugrunde, ob sich die Bildungsabschlüsse der Mütter (Spaltenvariable) aufgrund ihres Heiratsverhaltens in homogene Schichten zusammenfassen lassen. Wenn wir die diesbezügliche Abbildung 5.3 auf Seite 115 betrachten, so ist eine Clusterstruktur für die einzelnen Bildungsabschlüsse (A_2 bis A_7) nicht zu erkennen.

Tab. 5.3: Ergebnisse der bivariaten Korrespondenzanalyse für die abgeschlossene Schulbildung der Mütter (SchulM) und der Partner (SchulP)

			Spaltenskal.		kanon. Skal.	
			Dim. 1	Dim. 2	Dim. 1	Dim. 2
<i>Spaltenvariable</i>						
SchulbM	k. A.	A ₁	0,88	2,18	1,19	3,49
SchulbM	Pfl. o. L.	A ₂	0,53	-0,09	0,71	-0,14
SchulbM	Pfl. m. L.	A ₃	0,24	-0,16	0,33	-0,26
SchulbM	BMS	A ₄	-0,18	-0,06	-0,25	-0,10
SchulbM	BHS	A ₅	-0,52	-0,08	-0,70	-0,13
SchulbM	AHS	A ₆	-0,83	0,10	-1,11	0,17
SchulbM	Uni	A ₇	-1,18	0,26	-1,57	0,42
<i>Zeilenvariable</i>						
SchulbP	k. A.	B ₁	1,46	4,33	1,09	2,71
SchulbP	Pfl. o. L.	B ₂	1,05	-0,15	0,78	-0,09
SchulbP	Pfl. m. L.	B ₃	0,49	-0,47	0,36	-0,29
SchulbP	BMS	B ₄	-0,36	-0,21	-0,27	-0,13
SchulbP	BHS	B ₅	-0,89	-0,10	-0,66	-0,06
SchulbP	AHS	B ₆	-1,13	0,05	-0,84	0,03
SchulbP	Uni	B ₇	-2,03	0,64	-1,52	0,40
Eigenwerte			0,313	0,153		
erklärter χ^2 -Beitrag (%)			54,1	26,4		

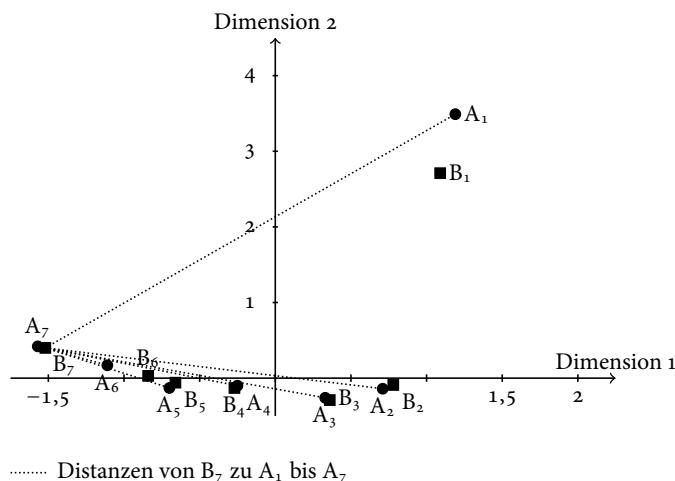


Abb. 5.2: Berechnete Punktekonfiguration bei kanonischer Skalierung (Abkürzungen siehe Tabelle 5.3)

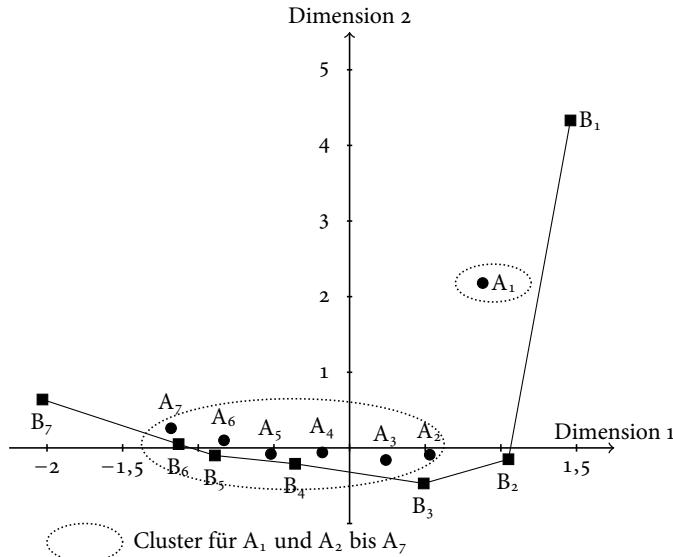


Abb. 5.3: Berechnete Punktekonfiguration bei einer Spaltenskalierung (Veranschaulichung der Interpretation der Zeilenvariablen; Abkürzungen siehe Tabelle 5.3)

Sie bilden ein einziges Cluster, das sich nur deutlich von dem Cluster mit fehlenden Werten unterscheidet. Eine *clusteranalytische Interpretation* der Ergebnisse ist somit nicht möglich.

Für eine *dimensionale Interpretation* müssen den Dimensionen Namen gegeben werden. Dazu können bei einer Spaltenskalierung die Ausprägungen der Zeilenvariablen durch eine Linie verbunden werden (Lebart, Morineau u. a. 1984, S. 57–58; siehe Abbildung 5.3). Verläuft diese Linie entlang einer Dimension, kann der Dimension der entsprechende Name der untersuchten Variablen gegeben werden. In unserem Beispiel ergibt sich folgendes Muster: Die Ausprägungen B₇ (SchulbP = Universität) bis B₂ (SchulbP = Pfl. o. L.) verlaufen entlang der ersten Dimension. Diese Ausprägungen bilden also die erste Dimension, die entsprechend den Ausprägungen als soziale Schichtungsdimension interpretiert werden kann. Die Ausprägung B₁ (SchulbP = k. A.) bildet dagegen die Dimension 2. Sie kann als Antwortverweigerungstendenz interpretiert werden, da nur jene Datensätze in die Analyse aufgenommen wurden, in denen die Mütter einen Partner haben. Fehlende Angaben bedeuten somit eine Antwortverweigerung und bilden eine eigenständige Dimension.

Vorausgesetzt wird bei dem eben dargestellten Vorgehen, dass die Variable zumindest teilweise ordinales Niveau hat. Bei nominalen Variablen kann für jede Ausprägung eine durch den Nullpunkt gehende Gerade gezeichnet werden. Verläuft diese Gerade entlang

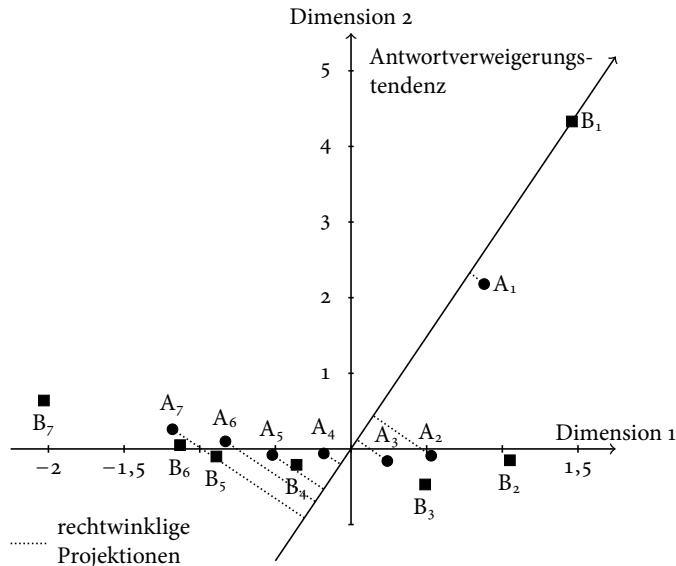


Abb. 5.4: Berechnete Punktekonfiguration bei einer Spaltenskalierung (Veranschaulichung der Vektorrepräsentation; Abkürzungen siehe Tabelle 5.3 auf Seite 114)

einer Dimension, kann der Dimension der Name der untersuchten Ausprägung gegeben werden. Ferner können auf diese Gerade die skalierten Ausprägungen (Ausprägungen, für die die Distanzen definiert sind) rechtwinklig projiziert werden (SPSS Inc. 1990b). Dadurch erhält man – ähnlich wie bei der Vektorrepräsentation – Anhaltspunkte, welche skalierten Ausprägungen auf dem Vektor der untersuchten Ausprägung ausdifferenziert werden. Dieses Vorgehen ist selbstverständlich auch bei ordinalen Variablen möglich. Legen wir beispielsweise durch die Ausprägung B_1 eine Gerade (siehe Abbildung 5.4), so sieht man, dass die Ausprägung A_1 von den anderen Ausprägungen ausdifferenziert wird. Weitere Interpretationstechniken sind zum Beispiel in den Arbeiten von Andersen (1991, S. 370–373), Greenacre (1989) oder Weller und Romney (1990) beschrieben. So zum Beispiel ist es in der »französischen Schule« der Korrespondenzanalyse üblich, in einem Diagramm für die Spaltenvariablen die Koordinatenwerte der Spaltenskalierung und für die Zeilenvariablen jene der Zeilenskalierung einzutragen. Der Winkel zwischen den Zeilen- und Spaltenvariablen gibt dann Auskunft über die Ähnlichkeit.

$U^2 =$	U_{11}^2	o	o	erste nominale Variable
	o	U_{22}^2	o	zweite nominale Variable
	o	o	U_{33}^2	dritte nominale Variable

Abb. 5.5: Aufbau der Residualmatrix bei der nominalen Faktorenanalyse nach McDonald

5.2 Nominale Faktorenanalyse nach McDonald

Bei der bivariaten Korrespondenzanalyse wird das Problem, dass die Zusammenhänge zwischen den Ausprägungen in den nominalen Variablen rein rechnerische Größen darstellen, dadurch gelöst, dass zwei Variablengruppen gebildet werden und nur die entsprechende Teilmatrix \tilde{G} in die Analyse eingeht. In der *nominalen Faktorenanalyse nach McDonald* (Arminger 1979, S. 162–171) wird ein anderer Zugang für die Lösung dieses Problems gewählt. Der Modellansatz besteht darin, dass für die Kovarianzmatrix K der in Dummies aufgelösten nominalen Variablen eine faktorenanalytische Zerlegung gesucht wird in:

$$K = L \cdot L^T + U^2 ,$$

wobei L die Matrix der Faktorladungen und U^2 die Residuenmatrix ist. K ist die Kovarianzmatrix der Dummies der nominalen Variablen mit den Elementen $k_{i(k)j(k^*)} = p_{i(k)j(k^*)} - p_{i(k)} \cdot p_{j(k^*)}$. Im Unterschied zu der Faktorenanalyse bei quantitativen Variablen mit einer Kommunalitätenschätzung (siehe Abschnitt 5.3.1) ist die Residualmatrix keine Diagonalmatrix, sondern eine Blockdiagonalmatrix, die als Residuen auch die rein rechnerisch bedingten Zusammenhänge zwischen den Ausprägungen innerhalb der nominalen Variablen enthält. Bei drei nominalen Variablen beispielsweise besitzt die Residualmatrix den in Abbildung 5.5 dargestellten Aufbau.

Die Schätzung der Faktorladungen und der Residuenmatrix erfolgt iterativ (Arminger 1979, S. 166–167). Die Faktorladungen werden dabei so normiert, dass die Summe der Faktorladungen der Ausprägungen einer nominalen Variablen in jedem Faktor h gleich o ist (Arminger 1979, S. 171):

$$\sum_k l_{i(k)h} = 0 \quad \forall i, h , \quad (5.1)$$

wobei $l_{i(k)h}$ die Faktorladung der Ausprägung k der nominalen Variablen i auf dem Faktor h ist. Für die berechneten Faktorladungen können alle im Rahmen der Faktoren-

analyse entwickelten Rotationsverfahren (siehe Abschnitt 5.3.1) eingesetzt werden, wobei die rotierten Faktorladungen entsprechend der Vorschrift (Gleichung 5.1 auf der vorherigen Seite) zu reskalieren sind. Die Unterschiede zu der multiplen Korrespondenzanalyse können wir wie folgt zusammenfassen:

1. Anstelle der Zusammenhangsmatrix \mathbf{G} wird die Kovarianzmatrix \mathbf{K} der Dummies untersucht. Zwischen der Zusammenhangsmatrix \mathbf{G} und der Kovarianzmatrix besteht folgende Beziehung:

$$\mathbf{G} = \mathbf{H}^{-1/2} \cdot \mathbf{K} \cdot \mathbf{H}^{-1/2},$$

wobei \mathbf{H} und \mathbf{G} entsprechend den Ausführungen zu der multiplen Korrespondenzanalyse definiert sind.

2. Bei beiden Verfahren wird zunächst eine Eigenwertzerlegung mit $\mathbf{G} = \mathbf{V} \cdot \mathbf{D} \cdot \mathbf{V}^T$ bzw. $\mathbf{K} = \mathbf{V}^* \cdot \mathbf{D}^* \cdot \mathbf{V}^{*T}$ der untersuchten Matrix durchgeführt. Da diese nicht skaleninvariant ist, unterscheiden sich die Eigenvektoren. Es gilt also: $\mathbf{V} \neq \mathbf{H}^{-1/2} \cdot \mathbf{V}^*$.
3. Im Unterschied zu der multiplen Korrespondenzanalyse wird nach einer ersten Berechnung der Eigenvektoren die Berechnung nicht beendet, sondern iterativ eine Kommunalitätenschätzung durchgeführt. Für die Bestimmung der Zahl der Faktoren bedeutet dies, dass alle Faktoren mit einem positiven Eigenwert als bedeutsam betrachtet werden können.
4. Eine Rotation der berechneten Faktoren ist bei der nominalen Faktorenanalyse nach McDonald möglich, bei der multiplen Korrespondenzanalyse dagegen nicht.
5. Eine Interpretation der rotierten oder unrotierten Faktorladungen als Mittelwerte der Ausprägungen auf den latenten Dimensionen ist nicht möglich. Die Faktorladungen der nominalen Faktorenanalyse sind als Kovarianzen der Ausprägungen mit den Faktoren zu interpretieren.

Wie stark sich die Ergebnisse einer multiplen Korrespondenzanalyse von jenen einer nominalen Faktorenanalyse unterscheiden können, kann für den Zusammenhang zwischen der abgeschlossenen Schulbildung der Mütter und jener ihrer Partner aufgezeigt werden. Bei der multiplen Korrespondenzanalyse werden zwei inhaltlich gut interpretierbare Dimensionen ermittelt. Bei einer nominalen Faktorenanalyse werden ebenfalls zwei Faktoren mit Eigenwerten größer 0 berechnet. Der dritte Eigenwert ist mit -0,0291 bereits negativ. Betrachtet man die berechneten Faktorladungen (Abbildung 5.6), so zeigt sich, dass nur die Ausprägungen A₂, A₃ sowie B₃ und B₇ hohe Faktorladungen in beiden Dimensionen haben. Alle anderen Ausprägungen liegen um den Nullpunkt. Eine inhaltliche Interpretation ist nicht möglich. Auch eine Rotation der Faktoren erbringt keine inhaltlich interpretierbaren Ergebnisse. Die rotierten Dimensionen aus einer schiefwinkligen Rotation sind in der Abbildung durch eine gestrichelte Linie gekennzeichnet.

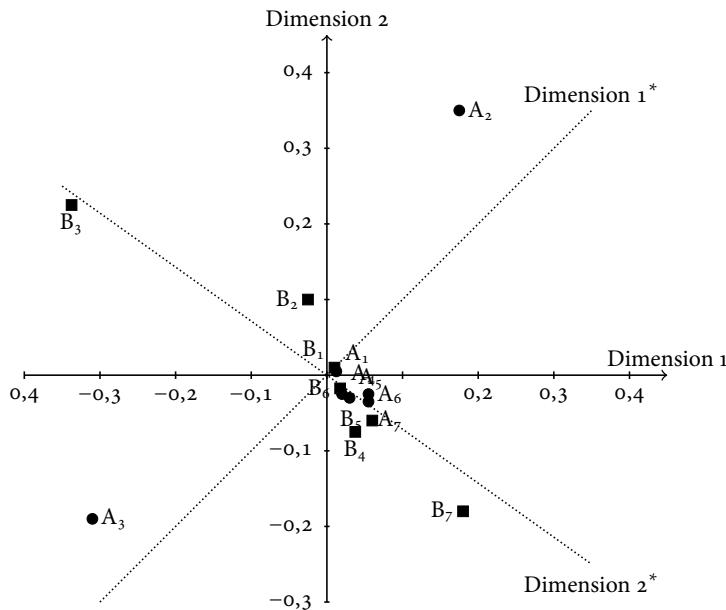


Abb. 5.6: Graphische Darstellung der Ergebnisse der nominalen Faktorenanalyse für eine Analyse des Zusammenhangs zwischen der abgeschlossenen Schulbildung von Müttern und jener ihrer Partner (Abkürzungen siehe Tabelle 5.3 auf Seite 114)

Die hier berichteten Unterschiede lassen sich vor allem auf die Verwendung unterschiedlicher Zusammenhangsmaße zurückführen. Die Verwendung der standardisierten Residuen bei der multiplen Korrespondenzanalyse führt dazu, dass Ausprägungen mit gleichen Kovarianzen in Abhängigkeit von den Erwartungswerten in einer multiplen Korrespondenzanalyse unterschiedliche Werte annehmen können. Eine Kovarianz von 0,20 zwischen zwei Ausprägungen kann beispielsweise dadurch zustande kommen, dass beide Ausprägungen einen Anteilswert von 0,80 haben und gemeinsam mit einer relativen Häufigkeit von 0,84 auftreten. Die Kovarianz ist in diesem Fall gleich 0,20 ($= 0,84 - 0,80 \cdot 0,80$). Für die standardisierten Residuen würde sich ein Wert von 0,25 ($= (0,84 - 0,80 \cdot 0,80) / \sqrt{0,80 \cdot 0,80}$) ergeben. Eine Kovarianz von 0,20 kann aber auch auftreten, wenn beide Ausprägungen einen Anteilswert von 0,20 und eine gemeinsame relative Auftrittshäufigkeit von 0,24 haben. Für die standardisierten Residuen würde sich in diesem Fall ein Wert von 1,00 ($= (0,24 - 0,20 \cdot 0,20) / \sqrt{0,20 \cdot 0,20}$) ergeben. Die Ausprägungen mit Anteilswerten von 0,20 und einer Kovarianz von 0,20 werden in der multiplen Korrespondenzanalyse also als ähnlicher betrachtet als die Ausprägungen mit Anteilswerten von 0,80 und einer Kovarianz von ebenfalls 0,20.

Die Frage, ob die multiple Korrespondenzanalyse oder nominale Faktorenanalyse angewendet werden soll, hängt somit von der Entscheidung ab, wie die Abweichungen der

relativen Auftrittshäufigkeiten $p_{i(k)j(k^*)}$ von ihren Erwartungswerten $(p_{i(j)} \cdot p_{i(k)})$ gewichtet werden sollen. Soll eine Gewichtung in Abhängigkeit von den Erwartungswerten durchgeführt werden, empfiehlt sich die Anwendung der multiplen Korrespondenzanalyse. Inhaltlich bedeutet dies, dass angenommen wird, dass dieselbe Abweichung (Kovarianz) zwischen der empirischen und erwarteten Auftrittshäufigkeit mehr über den Zusammenhang von zwei Ausprägungen aussagt, wenn diese mit einer geringeren erwarteten Häufigkeit auftreten. Ist eine derartige Gewichtung nicht erwünscht, wird man die nominale Faktorenanalyse anwenden. Ferner kann man sich bei der Entscheidung an folgenden Regeln orientieren:

- Interessiert primär eine graphische Darstellung in einem niedrigdimensionalen Raum, empfiehlt sich der Einsatz der multiplen Korrespondenzanalyse.
- Kann angenommen werden, dass den nominalen Variablen eine Einfachstruktur zugrunde liegt, empfiehlt sich die Verwendung der nominalen Faktorenanalyse nach McDonald, da bei ihr eine Rotation zum Auffinden der Einfachstruktur zulässig ist.

Als ein Beispiel soll mit Hilfe der nominalen Faktorenanalyse nach McDonald die dimensionale Struktur der Freizeitaktivitäten von Kindern untersucht werden. Betrachtet man die Eigenwerte, so zeigt sich ein deutlicher Eigenwertabfall zwischen dem ersten und zweiten Faktor (Eigenwert des ersten Faktors = 1,0283; Eigenwert des zweiten Faktors = 0,3827). Der erste Faktor lässt sich als durchschnittliches Freizeitprofil oder allgemeines Aktivitätsniveau interpretieren. Auf ihm laden alle Aktivitäten positiv. Insgesamt liegen sieben Faktoren mit Eigenwerten größer 0 vor. Eine Einfachstruktur lässt sich nicht erkennen (siehe Tabelle 5.4). Wegen der Normierungsvorschrift (Gleichung 5.1 auf Seite 117) sind die Faktorladungen des Ausübens und Nichtausübens einer Tätigkeit bis auf das Vorzeichen identisch, da ihre Summe gleich 0 sein muss. Eine Ausdifferenzierung von Freizeitaktivitäten der Art, dass bestimmte Tätigkeiten positiv und andere negativ laden, findet nicht statt. Auf dem ersten Faktor laden auch nach der Rotation alle Aktivitäten positiv. Er lässt sich als allgemeines Aktivitätsniveau interpretieren. Auf den folgenden vier Faktoren besitzt jeweils eine Aktivität eine hohe Faktorladung. Auf dem zweiten Faktor ist dies der Kirchenbesuch, auf dem dritten das Spielen mit Haustieren, auf dem vierten Faktor gemeinsame Familienaktivitäten und auf dem fünften Computerspiele. Auf dem sechsten Faktor besitzt unter anderem das Spazierengehen eine hohe Faktorladung. Der siebte Faktor wird von organisierten Freizeitaktivitäten gebildet.

Zusammenfassend zeigen die Ergebnisse, dass das Konzept der Einfachstruktur für die Freizeitaktivitäten nicht angemessen ist. Dies kann zwei Ursachen haben: Es liegt echte Mehrdimensionalität (jeder Freizeitaktivität liegen mehrere latente Dimensionen zugrunde) oder Strukturheterogenität (die Gesamtpopulation setzt sich aus K Subpopulationen mit unterschiedlichen faktorenanalytischen Strukturen zusammen) vor.

Tab. 5.4: Ergebnisse einer nominalen Faktorenanalyse von 20 Freizeitaktivitäten (schiefwinklige Quartimin-Rotation mit anschließender Normierung der Faktorladungen, achsparallel projizierte Faktorladungen)

Variable	Auspr.	Fakt. 1	Fakt. 2	Fakt. 3	Fakt. 4	Fakt. 5	Fakt. 6	Fakt. 7
Ausruhen	ja	0,1718	0,0419	0,0868	-0,1114	-0,1466	-0,1475	0,0728
Ausruhen	nein	-0,1718	-0,0419	-0,0868	0,1114	0,1466	0,1475	-0,0728
Freunde	ja	0,1236	0,0605	0,0629	-0,1069	-0,1143	-0,1285	0,0236
Freunde	nein	-0,1236	-0,0605	-0,0629	0,1069	0,1143	0,1285	-0,0236
Familie	ja	0,1817	0,0893	0,0840	-0,4102	-0,1772	-0,1869	0,0940
Familie	nein	-0,1817	-0,0893	-0,0840	0,4102	0,1772	0,1869	-0,0940
Basteln	ja	0,2221	0,0979	0,0965	-0,1561	-0,1683	-0,2208	0,1075
Basteln	nein	-0,2221	-0,0979	-0,0965	0,1561	0,1683	0,2208	-0,1075
Comics	ja	0,2018	0,0262	0,0718	-0,1204	-0,1921	-0,1525	0,0885
Comics	nein	-0,2018	-0,0262	-0,0718	0,1204	0,1921	0,1525	-0,0885
Musizieren	ja	0,1262	0,1261	0,0557	-0,1194	-0,1076	-0,1801	0,0908
Musizieren	nein	-0,1262	-0,1261	-0,0557	0,1194	0,1076	0,1801	-0,0908
Haustiere	ja	0,1281	0,0710	0,4767	-0,0972	-0,1201	-0,1242	0,0427
Haustiere	nein	-0,1281	-0,0710	-0,4767	0,0972	0,1201	0,1242	-0,0427
Kino	ja	0,0745	0,0043	0,0292	-0,0502	-0,0841	-0,0634	0,1349
Kino	nein	-0,0745	-0,0043	-0,0292	0,0502	0,0841	0,0634	-0,1349
Konzert	ja	0,0985	0,0614	0,0463	-0,0803	-0,0931	-0,1142	0,1397
Konzert	nein	-0,0985	-0,0614	-0,0463	0,0803	0,0931	0,1142	-0,1397
Musikhören	ja	0,2021	0,0787	0,0869	-0,1431	-0,1849	-0,2010	0,0459
Musikhören	nein	-0,2021	-0,0787	-0,0869	0,1431	0,1849	0,2010	-0,0459
Kirche	ja	0,0848	0,4464	0,0621	-0,0865	-0,0682	-0,1296	0,0089
Kirche	nein	-0,0848	-0,4464	-0,0621	0,0865	0,0682	0,1296	-0,0089
Fernsehen	ja	0,1788	0,0172	0,0625	-0,1139	-0,1942	-0,1407	0,0583
Fernsehen	nein	-0,1788	-0,0172	-0,0625	0,1139	0,1942	0,1407	-0,0583
Computer	ja	0,1567	-0,0171	0,0588	-0,0837	-0,2253	-0,0941	0,1550
Computer	nein	-0,1567	0,0171	-0,0588	0,0837	0,2253	0,0941	-0,1550
Buch	ja	0,1766	0,1346	0,0924	-0,1347	-0,1410	-0,1952	0,0071
Buch	nein	-0,1766	-0,1346	-0,0924	0,1347	0,1410	0,1952	-0,0071
Vereinsv.	ja	0,0953	0,0544	0,0278	-0,0798	-0,1135	-0,0996	0,1150
Vereinsv.	nein	-0,0953	-0,0544	-0,0278	0,0798	0,1135	0,0996	-0,1150
Radfahren	ja	0,1138	0,0844	0,0687	-0,0905	-0,1168	-0,1115	0,0020
Radfahren	nein	-0,1138	-0,0844	-0,0687	0,0905	0,1168	0,1115	-0,0020
Spazieren	ja	0,2186	0,1237	0,1027	-0,1814	-0,1717	-0,2191	0,0472
Spazieren	nein	-0,2186	-0,1237	-0,1027	0,1814	0,1717	0,2191	-0,0472
alleine Sp.	ja	0,2303	0,0712	0,1009	-0,1509	-0,2151	-0,2004	0,0444
alleine Sp.	nein	-0,2303	-0,0712	-0,1009	0,1509	0,2151	0,2004	-0,0444
Partys	ja	0,1166	0,0616	0,0427	-0,1007	-0,1191	-0,1201	0,1500
Partys	nein	-0,1166	-0,0616	-0,0427	0,1007	0,1191	0,1201	-0,1500
Sport	ja	0,1515	0,0737	0,0789	-0,1397	-0,1772	-0,1553	0,0820
Sport	nein	-0,1515	-0,0737	-0,0789	0,1397	0,1772	0,1553	-0,0820

grau hinterlegt: höchste Faktorladung jeder Ausprägung auf den Faktoren

5.3 Die Hauptkomponenten- und Faktorenanalyse

Die beiden bisher in diesem Kapitel behandelten Verfahren (bivariate Korrespondenzanalyse, nominale Faktorenanalyse nach McDonald) setzen nur nominalskalierte Variablen voraus. Sie können im Prinzip auch für ordinale und quantitative Variablen eingesetzt werden, wobei dann nur die nominale Information (Gleichheit oder Ungleichheit) der untersuchten Variablen in die Berechnung eingeht. Die *Hauptkomponentenanalyse* (»principle component analysis«, PCA) und *Faktorenanalyse* (»principle factor analysis«, PFA) dagegen wurden ursprünglich für quantitative Variablen entwickelt, eignen sich aber auch bei ordinalen und dichotomen Variablen (siehe dazu später). Die Hauptkomponentenanalyse ist ein räumliches Verfahren. Ihr Ziel ist die Darstellung von Variablen oder Objekten in einem dimensionalen Raum, der durch die sogenannten Hauptkomponenten aufgespannt wird. Der Faktorenanalyse liegt ein klares messtheoretisches Modell zugrunde, wobei im Idealfall jedes Item nur einen der gemeinsamen Faktoren misst. Es wird die Annahme getroffen, dass einer Menge von Variablen eine bestimmte Anzahl von gemeinsamen Faktoren zugrunde liegt. Zusätzlich besitzt jede Variable eine Eigenständigkeit und weist zufällige Messfehler auf. Die Faktorenanalyse greift somit auf ein Messmodell zurück; auf die Hauptkomponentenmethode trifft dies nicht zu. Ziel der Faktorenanalyse ist das Auffinden von gemeinsamen, inhaltlich benennbaren Dimensionen, während die Hauptkomponentenmethode »nur« die Darstellung in einem niedrigdimensionalen Raum anstrebt (Wolff und Bacher 2010). Die Faktorenanalyse kann sowohl zur Untersuchung von Variablen (*R-Faktorenanalyse*) als auch Personen (*Q-Faktorenanalyse*) durchgeführt werden. Der erstgenannte Fall stellt dabei die häufigste Anwendungssituation dar. Beide Möglichkeiten werden im weiteren Verlauf dargestellt.

5.3.1 Hauptkomponenten- und R-Faktorenanalyse

Zur *R-Faktorenanalyse* und zur *Hauptkomponentenmethode* liegt umfangreiche Literatur vor (siehe zum Beispiel Arminger 1979; Holm 1976; Ost 1984; u. a.). Wir wollen hier daher nur die wesentlichen Schritte darstellen:

1. Für die untersuchten Variablen wird eine Korrelationsmatrix berechnet. Sie soll im Folgenden mit R bezeichnet werden.
2. Für R soll geprüft werden, ob eine Faktorisierung sinnvoll ist (Dziuban und Shirley 1974). Hierzu stehen mehrere Testverfahren zur Verfügung. Die bekanntesten sind der *Bartlett-Test auf Sphärizität* (Bartlett 1950) und das *Kaiser-Mayer-Olkin-Kriterium* (Kaiser 1970; Kaiser und Rice 1974). Das Kaiser-Mayer-Olkin-Kriterium sollte größer 0,60 sein (Wolff und Bacher 2010).

3. Für \mathbf{R} wird eine Eigenwertzerlegung durchgeführt:

$$\mathbf{R} = \mathbf{V} \cdot \mathbf{D} \cdot \mathbf{V}^T$$

Bei der Faktorenanalyse wird nicht die Matrix \mathbf{R} faktorisiert, sondern die Matrix \mathbf{R}_h . Die Matrix \mathbf{R}_h unterscheidet sich von \mathbf{R} dadurch, dass in der Diagonale nicht Einsen stehen, sondern die sogenannten *Kommunalitäten* r_{ih}^2 der Variablen. Dies sind die Eigenkorrelationen der Variablen bzw. formal betrachtet, die durch die gemeinsamen Faktoren erklärten Varianzen in den Variablen. Die Größen sind unbekannt und müssen geschätzt werden (siehe dazu später).

4. Für die weitere Analyse wird eine bestimmte Faktoreanzahl ausgewählt. Dabei kann nach dem *Kaiserkriterium* (Guttman 1954a; Kaiser und Dickman 1959), dem *Scree-Test* (Cattell 1966; Horn 1965) und/oder dem *Eigenwertabfall* vorgegangen werden. Entsprechend dem Kaiserkriterium werden alle Faktoren mit einem Eigenwert größer 1 ausgewählt. Beim Scree-Test wird ein Scree-Diagramm mit der Faktoreanzahl in der Horizontalen und den entsprechenden Eigenwerten in der Vertikalen gezeichnet. Die Faktoreanzahl wird anhand des Knickpunktes festgelegt (Ost 1984, S. 603–604; u. a.):

$$\text{Faktoreanzahl} = \text{Knickpunkt} - 1 .$$

Bei dem Kriterium des Eigenwertabfalls wird die Faktoreanzahl dort festgelegt, wo ein deutlicher Abfall des Eigenwertes folgt. Besitzt der erste Faktor beispielsweise einen Eigenwert von 6,4, der zweite dagegen nur mehr einen Eigenwert von 3,4 (siehe Abbildung 5.7 auf Seite 125), wird mitunter entsprechend dem Eigenwertabfall die Faktoreanzahl gleich 1 gesetzt. Das Scree-Diagramm kann dagegen erst bei drei Faktoren einen signifikanten Knick aufweisen und es können fünf Faktoren mit Eigenwerten größer 1 vorliegen. Die drei Kriterien können somit zu einer unterschiedlichen Faktoreanzahl führen. In der Forschungspraxis wird man dann alle möglichen Lösungen weiter untersuchen.

5. Für die ausgewählten q Faktoren werden die Faktorladungen mit

$$\mathbf{F} = \mathbf{V}_q \cdot \mathbf{D}_q^{1/2}$$

berechnet, wobei \mathbf{V}_q die Matrix mit den ersten q Eigenvektoren ist. Bei der Faktorenanalyse wird hier zumeist ein Zwischenschritt eingeführt.

6. Die Hauptkomponentenanalyse ist damit in der Regel abgeschlossen. Die Ergebnisse können graphisch dargestellt werden, sofern die Zahl der Faktoren (Hauptkomponenten) kleiner 4 ist. Bei der Faktorenanalyse wird in der Regel – aber nicht immer! – ein Zwischenschritt eingeschaltet. Auf der Basis der q ausgewählten Faktoren werden die Kommunalitäten geschätzt mit $r_{ih}^2 = \sum_j f_{ij}^2$. Diese Schätzwerte werden in die Diagonale der Matrix \mathbf{R} eingetragen und die so entstehende Matrix \mathbf{R}_h wird

erneut faktorisiert. Die dabei berechneten Eigenvektoren und Eigenwerte werden in der Folge verwendet. Dieses Vorgehen ist beispielsweise in IBM-SPSS für die Hauptachsenmethode implementiert. Es führt dazu, dass sich die Zahl der Faktoren bei der Hauptkomponentenmethode und der Faktorenanalyse nicht unterscheidet.¹ Die Faktorladungen und damit die Eigenwerte und erklärten Varianzen fallen bei der Faktorenanalyse allerdings geringer aus. Wird auf den Zwischenschritt verzichtet, dann fallen die Faktorladungen und Eigenwerte größer aus. Im Folgenden soll von einer Faktorenanalyse gesprochen werden, wenn das Ziel das Auffinden von zugrunde liegenden gemeinsamen Faktoren ist und die nachfolgenden Schritte durchgeführt werden, unabhängig davon, ob eine Komunalitätenschätzung stattfindet oder nicht. Die Bezeichnung Hauptkomponentenanalyse soll dann verwendet werden, wenn nur eine graphische Darstellung in einem niedrigdimensionalen Raum gesucht wird.

7. Zur inhaltlichen Interpretation wird eine Rotation der Faktoren durchgeführt. Ziel der Rotation ist, eine neue Faktorladungsmatrix L zu erhalten, die eine Einfachstruktur aufweist: Jede untersuchte Variable soll nur auf einem Faktor eine hohe Ladung besitzen. Inhaltlich wird also angenommen, dass jede Variable nur eine Zieldimension misst. Formal lässt sich die Rotation als Transformation der ursprünglichen Faktorladungsmatrix darstellen als:

$$L = F \cdot T$$

wobei L die Matrix der rotierten Faktorladungen ist und T die $(q \times q)$ -Transformationsmatrix. Bei einer *schiefwinkligen Rotation*² können die Faktorladungen auf zwei Arten definiert werden: als Korrelationen der Variablen mit den Faktoren und als (standardisierte) Regressionskoeffizienten, die sich bei einer Regression mit den Faktoren als unabhängige Variablen und den untersuchten Variablen als abhängige Variablen ergeben. Die Faktorladungsmatrix mit den Korrelationen wird in der Literatur als Strukturmatrix oder Matrix der rechtwinklig projizierten Faktorladungen bezeichnet. Die Faktorladungsmatrix der (standardisierten) Regressionskoeffizienten wird als (eigentliche) Ladungsmatrix (»pattern matrix«) oder als Matrix der achsparallel projizierten Faktorladungen bezeichnet (Arminger 1979, S. 17; Holm 1976, S. 45). Diese werden mit dem Quartimin-Rotationsverfahren berechnet (Holm 1976, S. 109–118).³ In Abschnitt 5.3.2 werden wir die rechtwinklige Varimax-Rotation verwenden. Sie führt dazu, dass die Faktoren unkorreliert sind. Sie ist ebenfalls in Holm (1976, S. 108–109) beschrieben.

1 Es gibt auch andere Schätzverfahren für die Komunalitäten, die zu einer abweichenden Faktorenzahl führen können (siehe zum Beispiel Arminger 1979, S. 40–48).

2 »Schiefwinkelig rotieren« bedeutet, dass die berechneten Faktoren korrelieren können.

3 IBM-SPSS stellt als schiefwinklige Rotationsverfahren die Oblimin- und Promax-Rotation bereit (Wolff und Bacher 2010).

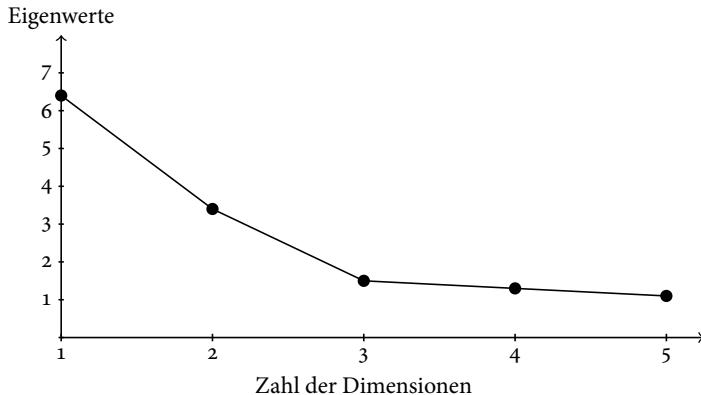


Abb. 5.7: Scree-Plot für fünf fiktive Eigenwerte

8. Ist die berechnete Faktorstruktur inhaltlich interpretierbar, können für die weitere Analyse Faktorwerte für die Personen berechnet werden mit (Arminger 1979, S. 114–116; Ost 1984, S. 625–626)

$$Y = Z \cdot R^{-1} \cdot L,$$

wobei Y die $(n \times q)$ -Matrix der Faktorwerte ist. Z ist die standardisierte $(n \times m)$ -Datenmatrix, R^{-1} die Inverse der Korrelationsmatrix und L die $(m \times q)$ -Faktorladungsmatrix.

Allgemein empfehlen wir eine vorsichtige Interpretation der Ergebnisse der Faktorenanalyse, wenn die Ergebnisse stark von einer Einfachstruktur abweichen. Hier ist durch weitere Analysen zu klären, ob echte Mehrdimensionalität (alle Items laden tatsächlich auf mehreren Dimensionen; Wolff und Bacher 2008) oder Strukturheterogenität (die Gesamtpopulation setzt sich aus K Subpopulationen mit unterschiedlichen faktorenanalytischen Strukturen zusammen) vorliegt.

Ist die Vorstellung der Einfachstruktur erfüllt, sollten – sofern dies möglich ist⁴ – anstelle von Faktorwerten *mittlere Gesamtpunktwerte* für die Faktoren berechnet werden, wenn die Faktoren in eine anschließende Clusteranalyse eingehen sollen. Dies hat den Vorteil, dass die mittleren Gesamtpunktwerte in der ursprünglichen Skala kodiert und interpretiert werden können und folglich Vergleiche zwischen Gesamtpunktwerten möglich sind (siehe Abschnitt 7.3). Der mittlere Gesamtpunktwert einer Person g in dem Faktor h ist

⁴ Dies ist dann möglich, wenn die Variablen vergleichbar sind oder theoretisch standardisiert werden können (siehe dazu Abschnitt 8.7).

Tab. 5.5: Berechnung der mittleren Gesamtpunktwerte

g	x_{g1}	x_{g2}	x_{g3}	w_{g1}	w_{g2}	w_{g3}	$\sum w_{gi} \cdot x_{gi}$	$\sum w_{gi}$	y_{gi}
1	1	1	KW	1	1	0	2	2	1
2	2	3	1	1	1	1	6	3	2
3	4	KW	KW	1	0	0	4	1	KW

Abkürzungen: kw: fehlender (nicht valider) Wert

als Mittelwert der Ausprägungen der Person g in den Variablen, die den Faktor h messen, definiert:

$$y_{gh} = \frac{1}{m_h} \sum_{i \in h} x_{gi}, \quad (5.2)$$

wobei x_{gi} die Ausprägung der Person g in der Variablen i ist. Die Zahl der Variablen i , die den Faktor h bilden, ist gleich m_h . Bei fehlenden Werten kann der mittlere Gesamtpunktwert als mittlere valide Summe mit

$$y_{gh} = \frac{\sum_{i \in h} w_{gi} \cdot x_{gi}}{\sum_{i \in h} w_{gi}} \quad (5.3)$$

berechnet werden, wobei w_{gi} eine Dummy-Variable ist, die den Wert 1 besitzt, wenn Person g in der Variablen i einen validen Wert hat, andernfalls ist w_{gi} gleich 0. Sind alle Ausprägungen valide, ergibt sich Gleichung 5.2. Um Divisionen mit 0 zu vermeiden und stabile Schätzwerte zu erhalten, ist die Definition eines Schwellenwertes für die Zahl valider Werte erforderlich. Tabelle 5.5 veranschaulicht die Berechnung des mittleren Gesamtpunktwertes bei fehlenden Werten. Die Variablen X_1 bis X_3 bilden den Faktor h . Die Person 1 hat in der dritten Variablen einen fehlenden Wert. Die Dummy-Variable w_{13} ist daher gleich 0. Für die Zahl der validen Werte ($\sum w_{gi}$) ergibt sich ein Wert von 2. Die valide Summe ($\sum w_{gi} \cdot x_{gi}$) ist gleich 2. Für den mittleren Gesamtpunktwert für die Person 1 erhält man somit nach Gleichung 5.3 einen Wert von 1 (= $2/2$). Die Person 2 erhält einen mittleren Gesamtpunktwert von 2. Für die dritte Person wurde der mittlere Gesamtpunktwert als nicht valide (kw) betrachtet, da bei der Berechnung gefordert wurde, dass mindestens 40 Prozent valide Werte vorliegen müssen. Die Person 3 besitzt aber nur zu 33 Prozent valide Werte.

Bevor wir ein Beispiel behandeln, soll noch auf einen für die Clusteranalyse wichtigen Sachverhalt hingewiesen werden: Das hier dargestellte Verfahren trifft keine Annahmen hinsichtlich der Verteilungsgestalt der Personen bzw. allgemein der untersuchten Objekte auf den Faktoren. Auf den zugrunde liegenden Dimensionen kann also eine homogene Verteilung (zum Beispiel eine Normalverteilung) ohne Clusterstruktur oder eine heterogene Verteilung (zum Beispiel eine Mischung von drei Normalverteilungen)

mit einer Clusterstruktur vorliegen. Dies ermöglicht nachfolgendes Vorgehen bei einer objektorientierten Clusteranalyse:

1. Liegt Einfachstruktur vor, werden mittlere Gesamtpunktwerte für die Faktoren berechnet. Ist die Annahme der Einfachstruktur verletzt, wird man zunächst deren Ursachen (echte Mehrdimensionalität oder Strukturheterogenität) überprüfen.
2. Die durch die mittleren Gesamtpunktwerte gemessenen Variablen werden anstelle der ursprünglichen Variablen in die Analyse einbezogen. Ist eine Berechnung von Gesamtpunktwerten nicht möglich, können Faktorwerte verwendet werden.

Dieses Vorgehen hat den Vorteil, dass in einer objektorientierten Clusteranalyse mit einer kleineren Variablenzahl gerechnet werden kann und somit der Interpretationsaufwand reduziert wird. Hinzu kommt, dass durch die Mittelwertbildung der Einfluss zufälliger Messfehler und damit von irrelevanten Variablen (siehe Abschnitt 6.7) reduziert wird.

Als ein Beispiel für eine R-Faktorenanalyse sollen wiederum die bereits mehrfach untersuchten Freizeitaktivitäten von Kindern analysiert werden. Dabei handelt es sich um dichotome Variablen. Streng genommen ist die Anwendung der Faktorenanalyse nicht zulässig, da diese quantitative Variablen voraussetzt. In der Forschungspraxis und in zahlreichen Simulationsstudien hat sich aber gezeigt, dass die Faktorenanalyse – wie die anderen multivariaten Verfahren – auch für ordinale und dichotome Variablen eingesetzt werden kann (siehe dazu zusammenfassend Bacher 1986). Nur in »extremen« Anwendungssituationen ist die Faktorenanalyse ungeeignet, die zugrunde liegende dimensionale Struktur zu reproduzieren. Eine derart »extreme« Anwendungssituation wäre zum Beispiel gegeben, wenn dichotome Variablen mit sehr unterschiedlichen Schwierigkeitsgraden⁵ vorliegen und diese fehlerfrei eine Dimension messen (siehe dazu auch das Beispiel in Denz 1982, S. 15–17). Formal ausgedrückt könnte eine Spezifikation der eben genannten »extremen« Anwendungssituation wie folgt aussehen:

1. Es liegen elf dichotome Variablen vor, denen ein gemeinsamer Faktor zugrunde liegt.
2. Die Personen besitzen auf dem gemeinsamen Faktor eine Standardnormalverteilung.
3. Die Variablen laden perfekt auf dem gemeinsamen Faktor. Sie besitzen also eine Faktorladung von 1 und eine Fehlervarianz von 0. Die nicht gemessenen quantitativen Variablen x_i besitzen daher ebenfalls eine Standardnormalverteilung.
4. Die Schwellenwerte, die zu einer Dichotomisierung der Antworten (0 = nein; 1 = ja) führen, sollen die in Tabelle 5.6 auf der nächsten Seite dargestellten sein.

⁵ Unter dem Schwierigkeitsgrad wird (bei dichotomen Variablen) der Anteil der Antworten mit 1 verstanden. In Tabelle 5.6 auf der nächsten Seite liegt er beispielsweise für das erste Item bei 95 Prozent, das heißt, dass 95 Prozent der Befragten diesem Item zugestimmt und mit 1 geantwortet haben. In der Tabelle variieren die Schwierigkeitsgrade zwischen 5 und 95 Prozent.

Tab. 5.6: Schwellenwerte für Simulationsexperiment

Item	Schwellenwert	Schwierigkeitsgrad (%)	Item	Schwellenwert	Schwierigkeitsgrad (%)
1	-1,6449	95	7	-0,2533	40
2	-1,2816	90	8	-0,5233	30
3	-0,8416	80	9	-0,8416	20
4	-0,5233	70	10	-1,2816	10
5	-0,2533	60	11	-1,6449	5
6	0,0000	50			

Die Annahmen 1 bis 4 entsprechen dem Modell einer perfekten Guttmanskala (Guttman 1950a).⁶

Bei einer Faktorenanalyse für 500 zufällig erzeugte Antwortmuster ergeben sich drei Faktoren mit Eigenwerten größer 1,0 (siehe Tabelle 5.7). Der erste Faktor misst die Zieldimension. Der zweite Faktor trennt die Variablen mit einem Schwierigkeitsgrad kleiner 50 Prozent von den Variablen mit einem Schwierigkeitsgrad größer 50 Prozent. Die Anordnung der Variablen auf diesem Faktor stimmt mit Ausnahme der beiden extremen Variablen (V_1 und V_{11}) mit jener der Schwierigkeitsgrade überein. Auf dem dritten Faktor werden die sehr leichten und sehr schwierigen Variablen (V_1 , V_2 und V_{10} , V_{11}) von den anderen Variablen getrennt. Das hier auftretende Muster wurde bereits Anfang der 1950er Jahre von Guttman (1950b, 1954b) untersucht.⁷ Der zweite Faktor wurde von Guttman als metrische Komponente der untersuchten Variablen bezeichnet, da er die Anordnung der Schwierigkeitsgrade auf einer quantitativen Skala abbildet. Der dritte Faktor wurde von Guttman als Intensitätsdimension interpretiert, wobei angenommen wird, dass extremere Variablen (sehr leichte oder sehr schwierige Variablen) mit einer intensiveren Einstellung verbunden sind.

Wichtig an dem Simulationsbeispiel ist, dass in »extremen« Anwendungssituationen bei dichotomen Variablen mehr »bedeutsame« Dimensionen (Faktoren mit Eigenwerten größer 1,0) berechnet werden können, als theoretisch vorhanden sind. Betrachtet man allerdings

⁶ Die Guttmanskala trifft keine Verteilungsannahme. Die Verteilungsannahme wurde hier nur zu Simulationszwecken spezifiziert.

⁷ Guttman analysierte allerdings nicht die Korrelationsmatrix (Φ -Korrelationen) der Variablen, sondern die Zusammenhangsmatrix mit den Elementen

$$\frac{P_{i(k)j(k^*)}}{\sqrt{P_{i(k)} \cdot P_{j(k^*)}}} .$$

Das Vorgehen entsprach somit jener einer multiplen Korrespondenzanalyse. Bildet man die nach Abzug des ersten Faktors verbleibende Residualmatrix, so ist diese Residualmatrix gleich der in der multiplen Korrespondenzanalyse untersuchten Zusammenhangsmatrix G .

Tab. 5.7: Ergebnisse der Faktorenanalyse bei dichotomen Variablen mit sehr unterschiedlichen Schwierigkeitsgraden, die perfekt einen Faktor messen (Ergebnisse einer Simulationsrechnung)

	Fakt. 1	Fakt. 2	Fakt. 3		Fakt. 1	Fakt. 2	Fakt. 3
V_1	0,3959	0,5071	-0,5105	V_7	0,8197	-0,1991	0,3014
V_2	0,5542	0,5759	-0,4024	V_8	0,7655	-0,4061	0,1166
V_3	0,6921	0,5129	-0,1116	V_9	0,6946	-0,5153	-0,0782
V_4	0,7805	0,3767	0,1423	V_{10}	0,5388	-0,5831	-0,4054
V_5	0,8202	0,2239	0,2829	V_{11}	0,3876	-0,5154	-0,5129
V_6	0,8355	-0,0079	0,3526				

Eigenwerte: 5,1045 (Faktor 1), 2,1202 (Faktor 2), 1,1980 (Faktor 3)

den Eigenwertabfall zwischen dem ersten und zweiten Faktor, so ergibt sich ein Verhältnis von 1 zu 2,41 (= 5,1045 / 2,1202). Der erste Faktor ist also beinahe zweieinhalbmal so groß wie der zweite Faktor. Aufgrund des Eigenwertabfalls könnte man sich daher nur für den ersten Faktor entscheiden.

Für die Forschungspraxis ist aber nicht das Verhalten der Faktorenanalyse beim Vorliegen einer perfekten Guttmankalierung entscheidend, sondern das *Verhalten bei Messfehlern*. Allgemein gilt hier, dass sich die Schwierigkeitsgrade der Variablen beim Auftreten zufälliger Messfehler in Richtung einer Gleichverteilung ändern.⁸ *Der Einfluss der Schwierigkeitsgrade, der bei einer Faktorenanalyse zu einem zweiten, dritten oder unter Umständen sogar zu einem vierten Faktor führt, geht somit verloren.* Tabelle 5.8 auf der nächsten Seite zeigt das Verhalten der Faktorenanalyse für unterschiedliche Faktorladungen und unterschiedliche Schwierigkeitsgrade der dichotomen Variablen. Ihr ist zu entnehmen, dass bei gleich schwierigen Items die einfaktorielle Struktur reproduziert wird. Auch bei nicht allzu unterschiedlichen Schwierigkeitsgraden (25 Prozent bis 75 Prozent) ist dies der Fall, wenn keine perfekte Guttmankala vorliegt. In den anderen Konstellationen tritt eine Überschätzung der Dimensionszahl auf, allerdings mit einem deutlichen Eigenwertabfall. *Da in der Forschungspraxis nicht von dem Vorliegen einer perfekten Guttmankala ausgegangen werden kann, kann die Faktorenanalyse auch bei dichotomen Variablen verwendet werden, wenn die Variablen nicht »extrem« unterschiedliche Schwierigkeitsgrade besitzen.* Dies gilt auch bei ordinalen Variablen. Hier nimmt der Einfluss von unterschiedlichen Schwierigkeitsgraden mit der Zahl der Kategorien und – selbstverständlich wie bei dichotomen Variablen – mit dem Vorhandensein von Messfehlern ab. Durch die Verwendung spezieller Korrelationskoeffizienten, zum Beispiel *tetrachorische Korrelationen* bei dichotomen Variablen oder *polychorische Korrelationskoeffizienten* bei ordinalen Variablen

⁸ Bei dichotomen Items nähern sich die Anteilswerte der beiden Ausprägungen dem Wert 0,5 an, bei trichotomen dem Wert 0,333 usw.

Tab. 5.8: Verhalten der Faktorenanalyse bei unterschiedlichen Faktorladungen und Schwierigkeitsgraden (Ergebnisse von Simulationsrechnungen)

Faktorladung ^{a)}	Variationsbereich der Schwierigkeitsgrade			
	5 bis 95 %	15 bis 85 %	25 bis 75 %	50 %
1,0	3 (2,41)	2 (2,55)	2 (4,20)	1
0,8	2 (2,73)	2 (3,50)	1	1
0,6	2 (2,52)	2 (2,87)	1	1

a) eine Faktorladung von 0,8 bedeutet in diesem Beispiel einen Fehlervarianzanteil von 36 Prozent ($100 \cdot (1 - 0,8 \cdot 0,8)$).

Anmerkung: Wird mehr als ein Faktor mit einem Eigenwert größer 1,0 berechnet, steht in Klammern der Eigenwertabfall zwischen dem ersten und zweiten Faktor, was dem Verhältnis zwischen dem ersten und dem zweiten Eigenwert entspricht.

(Arminger 1979, S. 160–162; Bollen 1989, S. 328–330), können höhere Faktorladungen erzielt werden.⁹

Betrachten wir nun die »Schwierigkeitsgrade« der im Abschnitt 4.6 untersuchten Freizeitaktivitäten der Kinder, so variieren diese zwischen 10,3 und 85,5 Prozent. Dies entspricht ungefähr der zweiten in den Simulationsrechnungen untersuchten Modellkonstellation (zweite Spalte mit einem Variationsbereich der Schwierigkeitsgrade von 15 bis 85 Prozent). Hier werden bei unterschiedlichen Faktorladungen durchgehend zwei Faktoren ausgewiesen. Bei einer Faktorenanalyse kann daher nicht ausgeschlossen werden, dass die Schwierigkeitsgrade die Ergebnisse beeinflussen. Inwiefern dies der Fall ist, kann durch eine Inspektion der unrotierten Faktorladungen überprüft werden. Liegt eine Verzerrung durch den Schwierigkeitsgrad vor, müssten 1) die Faktorladungen auf dem ersten Faktor alle positiv sein und 2) die Faktorladungen des zweiten Faktors mit den Schwierigkeitsgraden korrelieren. Bedingung 1) ist in diesem Beispiel erfüllt (siehe Tabelle 5.9), Bedingung 2) ist nicht erfüllt. Der zweite Faktor trennt – im Unterschied zu dem erwarteten Muster beim Vorliegen einer perfekten Guttman-Skala – nicht die Items nach ihren Schwierigkeitsgraden. So zum Beispiel besitzen sowohl das sehr »schwierige« Item »Kinobesuch« als auch das sehr »leichte« Item »Fernsehen« eine positive Faktorladung. Eine Verzerrung durch unterschiedlichen Schwierigkeitsgrade erscheint somit aufgrund des Musters der Faktorladungen wenig plausibel. Für die weitere Analyse sind zwei Entscheidungen möglich:

9 Diese Verfahren sind aber nicht bei einer Kombination mit clusteranalytischen Verfahren geeignet, da hier eine homogene Normalverteilung auf den zugrunde liegenden quantitativen Variablen angenommen wird.

Tab. 5.9: Unrotierte Faktorladungen einer Faktorenanalyse von dichotomen Freizeitaktivitäten

	Schwierigkeitsgrad (%)	Fakt. 1	Fakt. 2	Fakt. 3	Fakt. 4
Ausruhen	36,7	0,4029	0,0632	-0,1497	-0,4048
Freunde	79,7	0,3864	-0,1305	-0,1191	0,1378
Familie	58,1	0,4325	-0,0455	0,0970	0,0391
Basteln	41,6	0,4947	-0,1709	0,0353	-0,4075
Comics	46,3	0,4343	0,2459	-0,2362	-0,2419
Musizieren	26,5	0,3517	-0,3416	0,3530	-0,0052
Haustiere	55,3	0,3072	-0,0865	-0,0382	0,0822
Kino	10,3	0,2775	0,4315	0,3628	-0,1862
Konzert	16,4	0,3224	0,0283	0,4965	-0,1726
Musikhören	72,3	0,5146	-0,0811	-0,1910	-0,0354
Kirche	47,5	0,2269	-0,4513	0,2244	0,3056
Fernsehen	73,5	0,4465	0,3184	-0,3114	0,0935
Computersp.	39,1	0,3227	0,6057	-0,0846	0,2287
Buch	69,4	0,4506	-0,4025	-0,1045	0,0190
Vereinsv.	13,7	0,3384	0,2208	0,3690	0,2422
Radfahren	85,5	0,4104	-0,1180	-0,1894	0,4449
Spazieren	60,2	0,5136	-0,2707	-0,0843	-0,1869
alleine Sp.	62,1	0,5176	0,0409	-0,3027	-0,0349
Parties	18,9	0,3709	0,1810	0,4482	-0,0314
Sport	63,8	0,3998	0,1314	0,0596	0,4237
Eigenwerte		3,2666	1,4524	1,2847	1,1098

Abkürzungen: Zur genauen Bezeichnung der Freizeitaktivitäten siehe Abschnitt 4.6.

Anmerkung: Der Schwierigkeitsgrad entspricht dem Anteil der Kinder, die die untersuchte Freizeitaktivität ausführen.

1. Es wird – wie bei der nominalen Faktorenanalyse – von einer einfaktoriellen Lösung gesprochen, da zwischen dem ersten und zweiten Faktor ein starker Eigenwertabfall festgestellt werden kann. Der erste Faktor kann als allgemeines Aktivitätsniveau interpretiert werden. Die einfaktorielle Lösung erklärt 16,3 Prozent der Gesamtvarianz.
2. Die vierfaktorielle Lösung wird weiter untersucht. Sie erklärt insgesamt 35,6 Prozent (Summe der vier Eigenwerte dividiert durch Variablenzahl mal 100) der Gesamtvarianz.

Führt man für die vierfaktorielle Lösung eine Rotation durch, ergibt sich das in Tabelle 5.10 auf Seite 133 dargestellte Bild. Der *erste Faktor* wird von den Variablen »Ausruhen«, »Basteln«, »Comics lesen«, »Musikhören«, »Buchlektüre«, »Spazierengehen« und »alleine Spielen« gebildet. Er kann als *eher passiv, entspannende Freizeitaktivitätsdimension* bezeichnet werden. Der *zweite Faktor* wird auf dem positiven Pol durch die Variablen »Musizieren« und »Kirchenbesuch« gebildet, auf dem negativen Pol durch »Computer-

spielen«. Auch die Lektüre von Comics und Fernsehen besitzen auf dem zweiten Faktor relativ hohe negative Ladungen, während eine Buchlektüre eine relativ hohe positive Faktorladung besitzt. Die zugrunde liegende Dimension kann man als *pädagogisch orientierte Freizeitdimension* bezeichnen. Der *dritte Faktor* wird von Variablen gebildet, bei denen die Freizeitaktivitäten in »öffentlichen« Räumen stattfinden. Die Dimension könnte als *organisierte oder verwaltete Freizeitaktivitätsdimension* bezeichnet werden. Auf der *vierten Dimension* schließlich besitzen *sportliche Aktivitäten* (Sport betreiben und Radfahren) eine hohe Faktorladung. Sie kann als aktive Freizeitdimension bezeichnet werden. Ihr werden auch noch die verbleibenden Tätigkeiten zugeordnet. Diese besitzen aber deutlich geringere Faktorladungen und laden zudem auf anderen Faktoren relativ hoch. Insgesamt kann somit für die vierfaktorielle Lösung eine inhaltliche Interpretation gefunden werden. Gemeinsame Familienunternehmungen und das Spielen mit Haustieren lassen sich nur schlecht in dieser vierdimensionalen Struktur verorten, da sie in allen vier Dimensionen Faktorladungen kleiner 0,25 besitzen. Durch weiterführende Analysen sind die Ursachen hierfür zu ermitteln. Ferner laden einzelne Variablen auf mehreren Dimensionen, wie zum Beispiel das Fernsehen. Das Fernsehen wird formal der Gruppe der aktiven Freizeitaktivitäten (Faktor 4) zugeordnet, besitzt aber auch auf dem zweiten Faktor (pädagogisch »weniger wertvolle« Freizeitaktivitäten) und auf dem ersten Faktor der eher passiven, entspannenden Freizeitaktivitäten ebenso hohe Faktorladungen. Dies kann durchaus inhaltlich interpretiert werden. Ob die Abweichungen von der Einfachstruktur tatsächlich als »echte« Mehrdimensionalität interpretiert werden können, ist empirisch zu prüfen (Bacher 1992).

Vergleicht man die bisherigen Analysen der Freizeitaktivitäten (nichtmetrische mehrdimensionale Skalierung, nominale Faktorenanalyse und Faktorenanalyse) können folgende alternative Beschreibungsmodelle gegeben werden:

1. Den Freizeitaktivitäten liegt eine Dimension zugrunde, die sich als allgemeine Aktivitätsdimension interpretieren lässt.
2. Den Freizeitaktivitäten liegen mehrere Dimensionen zugrunde, wobei jede Freizeitaktivität in jeder Dimension einen Skalenwert besitzt. Die Annahme einer Einfachstruktur ist daher wenig zielführend. Die Rotationsverfahren der Faktorenanalyse tragen dieser Mehrdimensionalität keine Rechnung, da ihr Ziel das Auffinden einer Einfachstruktur ist. Einer theoretischen Mehrdimensionalität kann – unter den behandelten Verfahren – mit der nichtmetrischen mehrdimensionalen Skalierung durch die Eingabe von inhaltlichen Startwerten am besten Rechnung getragen werden (Wolff und Bacher 2008). Bezuglich der Faktorenanalyse stehen allerdings sogenannte konfirmatorische Verfahren zur Verfügung (zum Beispiel LISREL; Jöreskog und Sörbom 1984; Reinecke 2005), die ebenfalls – unter bestimmten Voraussetzungen – eine Modellierung einer Mehrdimensionalität erlauben.

Tab. 5.10: Rotierte Faktorladungsmatrix der Freizeitaktivitäten (schiefwinklige Quartimin-Rotation, rechtwinklige Projektionen)

	Fakt. 1	Fakt. 2	Fakt. 3	Fakt. 4
Ausruhen	0,7090	-0,1794	0,0793	-0,2804
Basteln	0,7221	0,1157	0,1608	-0,3016
Comics	0,5543	-0,3488	0,0697	-0,0396
Musikhören	0,4128	0,0061	-0,0557	0,2135
Buch	0,3394	0,3386	-0,1570	0,2132
Spazieren	0,5637	0,1941	-0,0244	-0,0056
alleine Sp.	0,4408	-0,1523	-0,0970	0,2485
Musizieren	0,1012	0,4780	0,2549	0,0397
Kirche	-0,2264	0,5713	0,0009	0,3857
Computersp.	-0,1466	-0,5194	0,2624	0,4672
Kino	0,1138	-0,2455	0,6363	-0,1667
Konzert	0,1371	0,1786	0,5765	-0,1860
Vereinsv.	-0,2824	0,0289	0,4878	0,3622
Parties	0,0038	0,0498	0,5951	0,0249
Freunde	0,1337	0,1084	-0,0805	0,3490
Familie	0,1553	0,1122	0,1815	0,2041
Haustiere	0,1080	0,0907	-0,0010	0,2324
Fernsehen	0,2173	-0,3831	-0,0163	0,3920
Radfahren	-0,1574	0,1250	-0,1823	0,7467
Sport	-0,2943	0,0084	0,1572	0,6736

Korrelationsmatrix der Faktoren				
Faktor 1	1,0000			
Faktor 2	0,1607	1,0000		
Faktor 3	0,2749	0,0686	1,0000	
Faktor 4	0,6195	0,0505	0,3126	1,0000

grau hinterlegt: höchste Faktorladung jeder Ausprägung auf den Faktoren

3. Es gibt unterschiedliche Freizeitprofile von Kindern, die die Ursache für die Mehrdimensionalität sind. Es liegt also Strukturheterogenität vor, während in Punkt 2 echte Mehrdimensionalität angenommen wird.
4. Schließlich kann auch ein g-Faktorenmodell gerechnet werden, bei dem ein genereller Faktor (allgemeines Aktivitätsniveau) und inhaltlich spezifische Faktoren angenommen werden (Wolff und Bacher 2008).

Des Weiteren sei noch auf einen oft in der Praxis wenig beachteten Sachverhalt hingewiesen: *Die Faktorenanalyse eignet sich nicht für Präferenzdaten mit fester Auswahl* (Bacher 1987; Wolff und Bacher 2010). Präferenzdaten sind dadurch gekennzeichnet, dass der Befragte aus einer Menge von Items eine bestimmte Anzahl auswählt (»wähle p von m Items«) oder alle Items (»reihe alle m Items«) oder eine Teilmenge davon (»reihe die

Tab. 5.11: Faktorenanalyse von Likert- und Präferenzdaten

Item	Ratingskalen		Rangskalen (3 Fakt.)			Rangskalen (2 Fakt.)	
	1	2	1	2	3	1	2
Ordnung und Ruhe	-0,112	0,788	-0,726	,166	0,070	-0,730	0,028
Mitsprache Regierung	0,716	-0,056	0,593	0,388	0,114	0,511	0,504
Inflationsbekämpfung	0,339	0,694	0,091	-0,990	-0,039	0,176	-0,838
freie Meinungsausübung	0,721	0,017	0,333	0,226	0,752	0,109	0,629
wirtschaftliches Wachstum	-0,104	0,798	-0,728	0,013	-0,063	-0,686	-0,167
Mitsprache Arbeitsplatz	0,771	0,037	0,508	0,206	-0,705	0,655	-0,071
Eigenwerte	1,811	1,686	1,846	1,276	1,000	1,846	1,276

*p wichtigsten«) in eine Rangreihe bringt. Durch die Auswahl bzw. Reihung entstehen lineare Abhängigkeiten, wenn die Zahl der auszuwählenden oder zu reihenden Items fest vorgegeben ist und nicht durch den Befragten frei gewählt werden kann. Müssen zum Beispiel drei von sechs Items ausgewählt werden, so sind nach der Auswahl der drei Items die Variablenwerte der verbleibenden drei Items unmittelbar bekannt. Wird »ausgewählt« mit »1« und »nicht ausgewählt« mit »0« kodiert, erhalten die drei ausgewählten Items den Wert »1«. Die verbleibenden Items haben dann automatisch den Wert »0«. Gleiches gilt, wenn die Items in eine Rangreihe gebracht werden müssen. Sind fünf von sechs Items gereiht, ist der Rangplatz des sechsten Items automatisch fixiert. Wurde also zum Beispiel Item *a* auf Platz 1 gereiht, Item *b* auf Platz 2, Item *c* auf Platz 3, Item *d* auf Platz 4 und Item *e* auf Platz 5, nimmt Item *f* automatisch den Platz 6 ein. Wenn die Zahl der auszuwählenden oder zu reihenden Items fest vorgegeben ist, soll von *Präferenzdaten mit fester Auswahl* gesprochen werden. Solche Daten werden auch als *ipsative Daten* (Horst 1965, S. 291–295) bezeichnet. Für diese Datenkonstellation eignet sich die Faktorenanalyse nicht. Dies soll anhand eines Beispiels verdeutlicht werden (siehe Tabelle 5.11).*

Ausgewählt werden sechs der zwölf Inglehart-Items von Denz (1989). Für jedes Item kann angeführt werden, ob es 1 »sehr wichtig«, 2 »wichtig«, 3 »weniger wichtig«, 4 »eher unwichtig« oder 5 »vollkommen unwichtig« ist. Es liegt eine Rating- oder Likertskala vor. Die Items sind strenggenommen ordinal, können aber entsprechend der Forschungspraxis als quantitativ (intervallskaliert) behandelt werden. Der Einsatz der Faktorenanalyse ist daher unproblematisch. Diese (Hauptkomponentenanalyse mit anschließender Varimax-Rotation) ermittelt auch sinnvolle Ergebnisse (siehe Tabelle 5.11): Der erste Faktor wird von den postmaterialistischen Items gebildet, der zweite von den materialistischen. In einem weiteren Schritt werden die Ratingdaten in eine Rangreihe transferiert. Für das Antwortmuster (3; 1; 3; 2; 2; 2) beispielsweise werden die Rangplätze (5,5; 1; 5,5; 3; 3; 3) ermittelt. Das Item *b* wird an erster Stelle gereiht, da es das einzige Item ist, das vom Befragten für »sehr wichtig« (1) gehalten wird. Die Items *d*, *e* und *f* nehmen mit einem

Wert von »wichtig« (2) die Rangplätze 2, 3 und 4 ein. Der mittlere Rangplatz ist 3 und wird den drei Items zugeordnet. Die letzten beiden Rangplätze teilen sich die Items *a* und *c*. Sie erhalten daher den mittleren Rangplatz 5,5 (= (5 + 6)/2). Für die rangtransformierten Daten wird erneut eine Faktorenanalyse durchgeführt, die drei Eigenwerte größer 1,0 ermittelt. Da der dritte Eigenwert nur sehr knapp über 1,0 liegt (Eigenwert 1,0001), stellt Tabelle 5.11 die zwei- und dreifaktorielle Lösung dar. In keiner der beiden Lösungen ist die ursprüngliche Struktur erkennbar. Der erste Faktor ist in der zwei- und dreifaktoriellen Lösungen bipolar und bildet die Entscheidung zwischen den materialistischen Items »Aufrechterhaltung der Ordnung und Ruhe«, »Wirtschaftswachstum« und den postmaterialistischen Items »Mitsprache bei wichtigen Regierungsentscheidungen« und »Mitsprache am Arbeitsplatz« ab. In der zweidimensionalen Lösung weist auch der zweite Faktor diese bipolare Struktur (»Inflationsbekämpfung« versus »freie Meinungsäußerung«) auf. In der dreidimensionalen bildet das Item »Inflationsbekämpfung« den zweiten Faktor und das Item »freie Meinungsäußerung« den dritten Faktor. Beide Lösungen legen eine andere Gruppierung von Variablen als die ursprüngliche Faktorenanalyse mit den Likertskalen in vier Gruppen nahe:

Gruppe 1: Aufrechterhaltung der Ordnung und Ruhe, Wirtschaftswachstum

Gruppe 2: Mitsprache bei wichtigen Regierungsentscheidungen, Mitsprache am Arbeitsplatz

Gruppe 3: Inflationsbekämpfung

Gruppe 4: freie Meinungsäußerung.

Das Beispiel zeigt deutlich, dass eine Faktorenanalyse von Präferenzdaten nicht sinnvoll ist. Gleiches gilt für die Korrespondenzanalyse. Geeignete Verfahren für Präferenzdaten mit fester Auswahl sind das eindimensionale und mehrdimensionale *Unfolding* (Sixtl 1982, S. 414–431; Coombs 1964, S. 61–201).

Eine weitere Datenkonstellation, für welche die Faktorenanalyse nicht geeignet ist, stellt das Vorhandensein mehrdimensionaler Items dar. Mehrdimensionale Items sind dadurch gekennzeichnet, dass sie auf mehreren gemeinsamen Faktoren laden. Freizeitaktivitäten (siehe Abschnitt 4.6) sind ein gutes Beispiel hierfür. Zur Reproduktion der dimensionalen Struktur ist die Faktorenanalyse nur bedingt geeignet, da die Rotationsverfahren Einfachstruktur erfordern, die nicht gegeben ist. Die Faktorenanalyse ermöglicht allerdings das Erkennen der Variablengruppen (Wolff und Bacher 2008). In diesem Fall sind (konfirmatorische) *mehrdimensionale Skalierungsverfahren* besser geeignet.

Für dichotome Items, die unterschiedlich schwierig sind, ist die Faktorenanalyse nur bedingt geeignet (siehe die vorausgehenden Simulationsrechnungen). Mit *Item-Response-Modellen*, wie der klassischen *Guttmanskala* (Guttman 1950a; Sixtl 1982, S. 387–400), der

Mokkenskalierung (Gerich 2001; Mokken 1971), der *Raschskalierung* und ihren Verallgemeinerungen (Rasch 1960; Sixtl 1982, S. 400–414), sind besser geeignete Verfahren verfügbar. Die genannten Verfahren nehmen eine hierarchische Datenstruktur an: Wird ein »schwierigeres« Item, zum Beispiel eine schwierige mathematische Aufgabe, mit einer bestimmten Wahrscheinlichkeit gelöst, dann wird ein »leichteres« Item mit einer größeren Wahrscheinlichkeit gelöst (Rost 2004, S. 90–96, 152–154).

5.3.2 Die Q-Faktorenanalyse

Die Faktorenanalyse wird üblicherweise zur Analyse des Zusammenhangs von Variablen (Spalten der Datenmatrix) eingesetzt. Im Prinzip kann die Aufgabenstellung auch umgedreht und eine *Q-Faktorenanalyse* für die Personen (Objekte, Zeilen der Datenmatrix) durchgeführt werden. Obwohl auf die Möglichkeit einer Q-Analyse in der Methodenliteratur häufig hingewiesen wird (so zum Beispiel Denz 1989; Kriz 1978, S. 260, 261), gibt es – im Unterschied zur Psychologie und Politikwissenschaft (McKeown und Thomas 1988; Sixtl 1982, S. 360–366) – nur wenige sozialwissenschaftliche Forschungsarbeiten, die diese Technik verwenden. Dies wird verständlich, wenn wir uns den Entstehungshintergrund und die Zielsetzung der Q-Analyse vor Augen führen. Ziel der Q-Analyse ist eine Analyse von subjektiven Einstellungsmustern und den dabei auftretenden interindividuellen Unterschieden (Stephenson 1959; Rinn 1961; McKeown und Thomas 1988). Entscheidend dabei ist nicht, wie viele Personen ein bestimmtes Einstellungsmuster haben, sondern wie Einstellungen individuell unterschiedlich strukturiert sein können. Wegen dieser Zielsetzung wird in der Regel nicht eine große Population (zum Beispiel $n = 1\,000$ oder $n = 2\,000$) befragt, sondern eine kleine Stichprobe (zum Beispiel $n = 100$) mit einem umfangreichen Fragebogen, wobei die Items aufgrund eines experimentellen Designs (»q-sort«) ausgewählt werden (Rinn 1961).

Bei der Q-Analyse wird dann davon ausgegangen, dass strukturelle Unterschiede zwischen den Objekten vorliegen und durch die Q-Analyse abgebildet werden können, wobei die Objekte nicht nur Individuen, sondern auch Aggregate (Organisationen, Nationen, Parteien usw.) sein können. So kann zum Beispiel gefragt werden, ob sich durch eine Q-Analyse unterschiedliche Entwicklungsmuster von Entwicklungsländern auffinden lassen. Zur Beantwortung dieser Frage soll für die Länder Mittel- und Südamerikas eine Q-Analyse durchgeführt werden. Als Variablen werden die in Tabelle 5.12 dargestellten Indikatoren der demographischen, wirtschaftlichen und sozialen Entwicklung verwendet (Nohlen 1984, S. 630–634). Die untersuchten Länder besitzen die in der Tabelle 5.13 auf Seite 139 dargestellten Werte in diesen Indikatoren.

Tab. 5.12: Entwicklungsindikatoren

	Indikator
V_3	Bruttonsozialprodukt pro Einwohner (BruSozPr)
V_4	Wirtschaftswachstum in den 1960er Jahren (Wachst6o)
V_5	Wirtschaftswachstum in den 1980er Jahren (Wachst8o)
V_6	Industrialisierungsquote in den 1960er Jahren (Indust6o)
V_7	Industrialisierungsquote in den 1980er Jahren (Indust8o)
V_8	Exportanteil der Rohstoffe (Export)
V_9	Schuldenbelastung (Schulden)
V_{10}	Import von Nahrungsmitteln (ImpNahr)
V_{11}	Kalorienversorgung (Kalorien)
V_{12}	Lebenserwartung (LebErw)
V_{13}	Kindersterblichkeit (Kindsterb)
V_{14}	Alphabetisierungsquote in den 1960er Jahren (Alpha6o)
V_{15}	Alphabetisierungsquote in den 1980er Jahren (Alpha8o)
V_{16}	Einschulungsquote (Einschulung)
V_{17}	Einwohner pro Arzt (EinwproArzt)
V_{18}	Bevölkerungswachstum in den 1970er Jahren (BevZu7o)
V_{19}	Bevölkerungswachstum in den 1980er Jahren (BevZu8o)
V_{20}	erwartetes Bevölkerungswachstum bis zum Jahr 2000 (BevZuoo)

Das Vorgehen der Q-Analyse besteht nun aus folgenden Schritten:

1. Die Variablen werden zunächst mit

$$z_{gi} = \frac{x_{gi} - \bar{x}_i}{s_i}$$

standardisiert, wobei \bar{x}_i der Mittelwert der Variablen i und s_i ihre Standardabweichung ist. Die dabei entstehende Datenmatrix soll mit Z bezeichnet werden. In den Spalten der standardisierten Datenmatrix stehen die Variablen, in den Zeilen die Objekte (Länder).

2. Da bei der Q-Analyse die Objekte faktorisiert werden sollen, muss die Datenmatrix transponiert werden. Die transponierte Datenmatrix soll mit ZQ symbolisiert werden. In den Zeilen stehen nun die Variablen, in den Spalten die Objekte. Die Objekte bilden die »neuen« Variablen. Formal ist die transponierte Datenmatrix wie folgt definiert:

$$ZQ = Z^T$$

mit zq_{ig} als Wert der Person g in den standardisierten Variablen i .

3. Für die Objekte wird eine Korrelationsmatrix berechnet. Sie soll mit \mathbf{Q} symbolisiert werden. Die Elemente der Matrix \mathbf{Q} bei Daten ohne fehlende Werte sind:¹⁰

$$q_{g,g^*} = \frac{\sum_i ((zq_{ig} - \bar{zq}_g) \cdot (zq_{ig^*} - \bar{zq}_{g^*}))}{[\sum_i (zq_{ig} - \bar{zq}_g)^2 \cdot \sum_i (zq_{ig^*} - \bar{zq}_{g^*})^2]^{1/2}}.$$

4. Für die Korrelationsmatrix \mathbf{Q} wird eine Faktorenanalyse durchgeführt und die Faktorladungen der Objekte berechnet.
5. Die Faktorladungen werden interpretiert. Wie bei der R-Analyse kann dazu eine Rotation der Faktoren durchgeführt werden. Aufgrund der Faktorladungen erkennt man, welche Objekte welchen Faktor bilden.
6. Zur inhaltlichen Interpretation können Faktorwerte der Variablen berechnet und/oder Cluster der untersuchten Objekte gebildet werden.

Führt man in unserem Beispiel eine Faktorenanalyse für die Korrelationsmatrix der Objekte (Länder) durch, so werden insgesamt sechs Eigenwerte größer 1,0 berechnet, die 85,8 Prozent der Gesamtvarianz erklären. Die rotierten Faktorladungen (Varimax-Rotation) der Objekte sind in Tabelle 5.14 auf Seite 140 wiedergegeben.

Bei der Zuordnung der Objekte zu den Faktoren ist es sinnvoll, nur signifikante Faktorladungen zu betrachten. Unter der Nullhypothese, dass eine Faktorladung von 0 vorliegt, berechnet sich die Standardabweichung der Faktorladungen mit $1/\sqrt{m}$ (m = Zahl der Variablen; McKeown und Thomas 1988, S. 50), da m Variablen als Fälle in die Analyse eingehen. Legt man ein Signifikanzniveau von 95 Prozent zugrunde, können jene Faktorladungen als signifikant bezeichnet werden, die größer $1,96 \cdot 1/\sqrt{m}$ sind. In unserem Beispiel ergibt sich ein Schwellenwert von $1,96 \cdot 1/\sqrt{18} = 0,462$. Faktorladungen mit einem Absolutbetrag größer 0,462 können als »signifikant« betrachtet werden.¹¹

Auf dem *ersten Faktor* werden somit die in Tabelle 5.15a auf Seite 141 dargestellten Länder ausdifferenziert, der *zweite Faktor* differenziert die in Tabelle 5.15b auf Seite 141 aufge-listeten Länder aus. Der *dritte Faktor* besitzt nur auf dem negativen Pol signifikante Faktorladungen für Bolivien und Peru. Signifikante Faktorladungen auf dem *vierten Faktor* haben Brasilien, die Dominikanische Republik und Mexiko. Auf der *fünften Dimension* kann wiederum eine Ausdifferenzierung festgestellt werden. Der positive Pol wird von Ecuador und Venezuela gebildet, der negative von Jamaika. Der *sechste Faktor*

¹⁰ Bei Daten mit fehlenden Angaben kann nach der Methode des paarweisen Ausscheidens vorgegangen werden (siehe dazu Abschnitt 8.8).

¹¹ Für eine variablenorientierte Analyse ist die hier dargestellte Signifikanzberechnung für die Faktorladungen nicht sinnvoll. Da die Fallzahl deutlich größer ist, würden sehr kleine Schwellenwerte berechnet werden. Eine Stichprobe von $n = 1\,000$ beispielsweise würde zu einer Schwelle von $1,96 \cdot 1/\sqrt{1\,000} = 0,0632$ führen.

Tab. 5.13: Demographische, wirtschaftliche und soziale Indikatoren für die Länder Mittel- und Südamerikas

	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉	V ₁₀	V ₁₁
Argentinien	2390	4,2	2,9	38	-9	76	166	6	125
Bolivien	570	5,2	4,8	25	29	97	259	10	87
Brasilien	2050	5,3	9,8	35	37	61	340	10	109
Chile	2150	4,5	0,1	51	37	80	229	14	114
Costa Rica	1730	6,5	5,7	20	29	75	164	9	116
Dom. Rep.	1160	4,4	9,1	23	27	74	215	17	105
Ecuador	1270	-9	9,2	19	38	97	140	8	88
El Salvador	660	5,9	4,1	19	21	76	35	18	99
Guatemala	1080	5,6	6,0	-9	-9	79	35	8	93
Haiti	270	-2,0	0,4	-9	-9	61	42	26	96
Honduras	560	5,3	3,6	19	25	90	99	10	96
Jamaika	1040	4,4	-1,1	36	37	48	128	20	119
Kolumbien	1180	5,1	6,4	26	30	78	96	12	108
Kuba	910	1,1	2,9	-9	-9	99	-9	-9	122
Mexiko	2090	7,3	5,0	29	38	61	319	8	121
Nicaragua	740	7,3	9	21	31	88	145	15	99
Panama	1730	7,8	3,5	21	-9	-9	184	10	103
Paraguay	1300	4,3	7,2	20	25	89	113	13	134
Peru	930	4,9	3,0	33	45	89	313	20	99
Trinidad	4370	3,9	3,4	46	62	94	22	11	113
Uruguay	2810	1,2	1,6	28	33	56	118	8	110
Venezuela	3630	5,9	5,7	22	47	99	132	15	112

(a) Teil 1

	V ₁₂	V ₁₃	V ₁₄	V ₁₅	V ₁₆	V ₁₇	V ₁₈	V ₁₉	V ₂₀
Argentinien	71	2	91	93	110	530	14	16	11
Bolivien	51	23	39	63	82	1850	23	25	24
Brasilien	64	7	61	86	89	1700	29	22	20
Chile	68	2	84	88	119	1920	21	17	14
Costa Rica	73	1	-9	90	107	1470	34	25	20
Dom. Rep.	62	5	65	67	96	4020	27	30	25
Ecuador	62	8	68	81	107	1620	30	30	27
El Salvador	63	7	49	62	82	3040	29	29	27
Guatemala	59	5	32	46	69	8600	30	30	30
Haiti	54	17	15	23	62	8200	15	17	20
Honduras	59	9	45	60	89	3120	31	34	30
Jamaika	71	3	82	90	99	2830	14	15	20
Kolumbien	63	4	63	81	128	1920	30	23	20
Kuba	73	1	-9	95	112	700	20	13	12
Mexiko	66	4	65	83	124	1260	33	31	25
Nicaragua	57	10	-9	90	85	1800	26	34	29
Panama	71	1	73	85	115	980	29	23	21
Paraguay	65	2	75	84	102	1710	25	32	24
Peru	58	9	61	80	112	1390	28	26	23
Trinidad	72	1	93	95	96	1490	20	13	15
Uruguay	71	2	-9	94	105	540	11	3	10
Venezuela	68	2	63	82	110	950	34	33	23

Abkürzung: »-9«: fehlender Wert; zu den Variablen siehe Tabelle 5.12.

Anmerkung: Die Werte der Variablen V₁₈ bis V₂₀ sind mit 10 multipliziert. Der Wert von 14 für Argentinien in V₁₈ entspricht also einem Bevölkerungswachstum von 1,4 Prozent.

(b) Teil 2

Tab. 5.14: Rotierte Faktorladungen der Objekte aus der Q-Analyse (Ergebnisse der Varimax-Rotation der ersten sechs Faktoren)

	Fakt. 1	Fakt. 2	Fakt. 3	Fakt. 4	Fakt. 5	Fakt. 6
Argentinien	0,8760	-0,1760	0,2195	0,0104	-0,2228	0,0417
Bolivien	-0,3794	0,3598	-0,6015	0,1465	0,3718	-0,1586
Brasilien	0,1201	-0,1163	-0,0179	0,9246	0,0916	-0,1375
Chile	0,7798	-0,0786	-0,2109	-0,1049	-0,4413	-0,0973
Costa Rica	0,0134	-0,7804	0,3956	0,0911	0,0534	0,3826
Dom. Rep.	-0,5970	0,2880	0,0634	0,4808	0,0809	0,3835
Ecuador	-0,3334	-0,1290	-0,1436	0,0393	0,8294	0,1100
El Salvador	-0,9312	0,1240	0,1187	-0,1948	-0,0675	0,0799
Guatemala	-0,7568	0,3128	0,3132	0,0869	0,2254	-0,2158
Haiti	-0,4299	0,8076	-0,0610	-0,0026	-0,2768	-0,1468
Honduras	-0,8580	0,0647	-0,1431	-0,1611	0,3420	0,0732
Jamaika	0,3061	0,0395	0,0709	-0,1258	-0,8862	-0,1252
Kolumbien	0,0368	-0,3355	-0,0365	0,0251	0,2057	0,6452
Kuba	0,8181	0,0480	0,1007	-0,3324	-0,0722	0,4072
Mexiko	-0,0308	-0,7722	-0,0960	0,4846	-0,1185	0,0591
Nicaragua	-0,6682	-0,1038	-0,4464	-0,3693	0,0245	-0,1106
Panama	0,0697	-0,9347	0,0444	-0,1426	0,0875	0,0130
Paraguay	-0,0744	-0,0605	0,3203	-0,1284	0,0149	0,8134
Peru	-0,0443	0,0932	-0,8896	0,0136	0,0171	-0,1596
Trinidad	0,7501	0,0321	0,3482	-0,2974	0,0977	-0,4001
Uruguay	0,8149	-0,0123	0,2713	0,0991	-0,2792	-0,1306
Venezuela	0,0479	-0,4055	0,2917	-0,3619	0,5360	-0,0751

grau hinterlegt: signifikante Faktorladungen

weist nur signifikante positive Faktorladungen für Kolumbien und Paraguay aus. Für die weitere Interpretation sind zwei Zugänge möglich:

1. Bildung von Clustern und deren Beschreibung
2. Berechnung von Faktorwerten für die Variablen

Bildung von Clustern: Liegen nur zwei Faktoren vor, ist eine graphische Darstellung der Objekte möglich. Eine Clusterbildung kann – wie bei den bisherigen Verfahren – aufgrund einer graphischen Inspektion der räumlichen Darstellung vorgenommen werden. Bei mehr als zwei Dimensionen ist folgendes Vorgehen möglich:

1. Die signifikanten Faktoren werden auf 1 bzw. -1 gesetzt, alle nicht signifikanten auf 0. Dadurch erhält man eine neue Faktorladungsmatrix.
2. Objekte mit gleichen Zeilenmustern werden zu einem Cluster zusammengefasst.
3. Für die so gebildeten Cluster werden zur Beschreibung die Mittelwerte in den Variablen gebildet.

Tab. 5.15: Positiver und negativer Pol des ersten und zweiten Faktors

positiver Pol des 1. Faktors	negativer Pol des 1. Faktors
Argentinien	Dominikanische Republik
Chile	El Salvador
Kuba	Guatemala
Trinidad	Honduras
Uruguay	Nicaragua

(a) erster Faktor

positiver Pol des 2. Faktors	negativer Pol des 2. Faktors
Haiti	Costa Rica
	Mexiko
	Panama

(b) zweiter Faktor

Alternativ ist es natürlich möglich, analog zu der multiplen Korrespondenzanalyse oder zu der nichtmetrischen mehrdimensionalen Skalierung, für die Koordinatenmatrix eine hierarchische Clusteranalyse zu rechnen.

Faktorwerte für die Variablen: Zur Berechnung von Faktorwerten für die Variablen existieren unterschiedliche Zugänge. Ein Zugang besteht darin, die Faktorwerte für die Variablen analog der R-Analyse zu berechnen. Wir wollen hier nicht diesen Zugang, sondern den Ansatz von Brown (zitiert in McKeown und Thomas 1988) darstellen. Die Formel zur Berechnung des Faktorwertes einer Variablen i in einem Faktor h besteht aus folgenden Schritten:

1. Ziehe in die Berechnung nur jene Objekte ein, die auf dem untersuchten Faktor eine signifikante Faktorladung besitzen.
2. Berechne für jedes Objekt g mit signifikanter Faktorladung f das Kreuzprodukt:

$$zq_{ig} \cdot fq_{gh} \cdot$$

3. Berechne einen gewichteten Mittelwert über die Kreuzprodukte mit

$$\bar{fq}_{ih} = \frac{\sum zq_{ig} \cdot fq_{gh} \cdot w_{gh}}{\sum w_{gh}} .$$

Dies ist der gesuchte Faktorwert der Variablen i auf dem Faktor h , wobei die Gewichte w definiert sind mit

$$w_{gh} = \frac{1}{1 - fq_{gh}^2} .$$

Tab. 5.16: Faktorwerte der in die Analyse einbezogenen Variablen

	Fakt. 1	Fakt. 2	Fakt. 3	Fakt. 4	Fakt. 5	Fakt. 6
BruSozPr	0,6909	-0,3474	0,5773	0,3223	0,3157	-0,2314
Wachst60	-0,5635	-1,3552	-0,0795	0,2620	0,1946	-0,0782
Wachst80	-0,2682	-0,1609	0,2683	1,4004	1,3488	0,6877
Indust60	0,8823	0,5448	-0,3146	0,4822	-0,7149	-0,4925
Indust80	0,8627	0,1087	-0,6218	0,1236	0,0981	-0,6361
Export	-0,0210	0,0115	-0,6121	-0,9381	1,3938	0,3134
Schulden	0,4012	-0,5465	-1,3042	1,4937	0,0646	-0,4034
ImpNahr	-0,4378	0,8581	-0,8906	-0,3770	-0,8849	-0,0098
Kalorien	0,6819	-0,1832	0,7279	0,1326	-0,8542	1,1296
LebErw	0,6466	-0,9204	1,0255	-0,0727	-0,5282	-0,0231
Kindsterb	-0,4182	0,8547	-0,8817	0,1318	0,2865	-0,4235
Alpha60	0,8391	-0,7308	0,2218	-0,0297	-0,3539	0,3315
Alpha80	0,7261	-0,7627	0,0537	0,2790	-0,2364	0,2153
Einschulung	0,6514	-0,9823	-0,3227	-0,3625	0,1959	0,4596
EinwproArzt	-0,5492	0,8306	0,3440	-0,1899	-0,2572	-0,2018
BevZu70	-0,7525	-0,7920	-0,2342	0,4781	1,0480	0,1602
BevZu80	-0,9055	-0,2119	-0,2114	-0,0302	0,7857	0,5063
BevZuoo	-1,1259	-0,0516	-0,2544	-0,0721	0,4034	0,1773

Abkürzungen: siehe Tabelle 5.12 auf Seite 137

Sie führen dazu, dass Objekte mit hohen Faktorladungen in der Berechnung ein höheres Gewicht erhalten.

Für unser Beispiel ergeben sich die in der Tabelle 5.16 dargestellten Faktorwerte. Die Interpretation der Ergebnisse soll exemplarisch für den ersten Faktor verdeutlicht werden (siehe Tabelle 5.17): Der positive Pol des Faktors ist durch einen relativ hohen Industrialisierungsgrad in den 1960er und 1980er Jahren sowie durch eine hohe Alphabetisierungsquote in diesen Jahren gekennzeichnet. Ferner charakterisieren ein relativ hohes Bruttonsozialprodukt, eine relativ gute Kalorienversorgung, eine relativ hohe Lebenserwartung und eine relativ hohe Einschulungsquote den positiven Faktor. Betrachten wir die Objekte mit signifikanten positiven Faktorladungen so zeigt sich, dass dieses Muster auf Argentinien, Chile, Kuba, Trinidad und Uruguay zutrifft. Der negative Pol ist dagegen vor allem durch ein hohes Bevölkerungswachstum gekennzeichnet, das sich bis zum Jahr 2000 (BevZuoo) fortsetzen wird. Durch dieses Muster sind die Dominikanische Republik, El Salvador, Guatemala, Honduras und Nicaragua charakterisiert.

Auf der Grundlage des Q-Analyse-Ansatzes haben Shepard und Arabie (1979) ein additatives Clustermodell entwickelt, beim dem für eine Ähnlichkeitsmatrix \ddot{A} (zum Beispiel eine Q-Korrelationsmatrix) eine clusteranalytische Darstellung mit

$$\ddot{A} = C \cdot W \cdot C + U$$

Tab. 5.17: Beziehung zwischen Faktorwerten und Objekten

positiver Pol des 1. Faktors	negativer Pol des 1. Faktors
<i>Variablen (Faktorwerte)</i>	
Indust60 (0,88)	BevZu00 (-1,13)
Indust80 (0,87)	BevZu80 (-0,90)
Alpha60 (0,84)	BevZu70 (-0,75)
Alpha80 (0,73)	Wachst60 (-0,56)
BruSozPr (0,69)	EinwproArzt (-0,54)
Kalorien (0,68)	
LebErw (0,65)	
Einschulung (0,65)	
<i>Objekte</i>	
Argentinien	Dominikanische Republik
Chile	El Salvador
Kuba	Guatemala
Trinidad	Honduras
Uruguay	Nicaragua

Abkürzungen: siehe Tabelle 5.12 auf Seite 137

gesucht wird, wobei in der Matrix C die Klassifikation steht. W ist die Diagonalmatrix der Eigenwerte und U die Matrix der additiven Konstanten. Die Klassifikationsmatrix C enthält nur Werte von 0 oder 1. »0« in einer Spalte h (Cluster h) für ein Objekt bedeutet, dass das untersuchte Objekt nicht dem Cluster h angehört, »1« bedeutet dagegen, dass das Objekt dem Cluster h angehört. Das additative Clustermodell gehört der Gruppe der deterministischen Clusteranalyseverfahren an. Überlappungen sind erlaubt. DeSarbo (1982) hat diesen Ansatz für die Analyse einer nicht symmetrischen Ähnlichkeitsmatrix verallgemeinert.

5.4 Anwendungsempfehlungen

- Sollen gemeinsame zugrunde liegende Dimensionen (Faktoren) für *Variablen* aufgefunden werden, empfehlen wir den Einsatz der »gewöhnlichen« Faktorenanalyse für quantitative, ordinale und dichotome Variablen (siehe Abschnitt 5.3.1) und der nominalen Faktorenanalyse nach McDonald für nominalskalierte Variablen (siehe Abschnitt 5.2). Eine dimensionale Interpretation ist aber auch mit der multiplen Korrespondenzanalyse und der nichtmetrischen mehrdimensionalen Skalierung möglich (siehe Kapitel 3 und 4). Mit der multiplen Korrespondenzanalyse haben wir gute Ergebnisse erzielt. Die Entscheidung, ob mit der nominalen Faktorenanalyse

nach McDonald oder mit der multiplen Korrespondenzanalyse gerechnet wird, hängt davon ab, ob das gemeinsame Auftreten von zwei Merkmalen abhängig von ihren Randhäufigkeiten gewichtet werden soll (siehe Abschnitt 5.2).

2. Wird dagegen eine räumliche Darstellung in einem niedrigdimensionalen Raum gesucht, wird zum Einsatz der multiplen Korrespondenzanalyse (siehe Kapitel 3), der nichtmetrischen mehrdimensionalen Skalierung (siehe Kapitel 3) oder Hauptkomponentenmethode (siehe Abschnitt 5.3.1) geraten. Von den drei Verfahren empfehlen wir die multiple Korrespondenzanalyse oder die Hauptkomponentenmethode, da sie im Vergleich zur nichtmetrischen mehrdimensionalen Skalierung eine eindeutige Lösung besitzen. Ist eine Unterscheidung in zwei Variablengruppen sinnvoll, empfehlen wir die bivariate Korrespondenzanalyse (siehe Abschnitt 5.1).
3. Wurden die Ähnlichkeiten direkt erhoben oder indirekt mittels Stimulusskalierung berechnet, sollte die nichtmetrische mehrdimensionale Skalierung angewendet werden (siehe Kapitel 4). Alle anderen hier behandelten Verfahren setzen die Verwendung eines bestimmten Zusammenhangsmaßes voraus.
4. Liegen dagegen responseskalierte Daten vor, empfehlen wir für eine räumliche Darstellung bei nominalen Variablen die multiple Korrespondenzanalyse (siehe Kapitel 3) und bei quantitativen Variablen die Hauptkomponentenmethode (siehe Abschnitt 5.3.1). Bei ordinalen Variablen kommen beide Verfahren in Betracht, die multiple Korrespondenzanalyse hat den Vorteil, dass die Ordinalität der Antwortkategorien geprüft werden kann. Bei fehlenden Werten kann sie zur Schätzung von Skalenwerten für die fehlenden Werte verwendet werden.
5. Bei der Faktorenanalyse empfehlen wir die Anwendung eines schiefwinkligen Rotationsverfahrens, insbesondere dann, wenn ein erster Faktor mit einem hohen Eigenwert vorliegt, da dies oft ein Hinweis auf Gemeinsamkeiten zwischen den Faktoren ist.
6. Ist die Annahme der Einfachstruktur bei der Faktorenanalyse nicht erfüllt, sollte eine Analyse durchgeführt werden, ob echte Mehrdimensionalität oder Strukturheterogenität vorliegt (siehe Abschnitt 5.3.1).
7. Bei der Faktorenanalyse empfehlen wir die Verwendung von mittleren Gesamtpunktwerten, da sie eine Interpretation auf der ursprünglichen Skala ermöglichen und Vergleiche von Skalenwerten eines Objektes in den Faktoren möglich sind (siehe Abschnitt 5.3.1).

Teil II

Deterministische Clusteranalyseverfahren

6 Einleitende Übersicht

6.1 Überlappende und überlappungsfreie Clusterlösungen

Deterministische Clusteranalyseverfahren unterscheiden sich von den in Teil I behandelten unvollständigen Clusteranalyseverfahren unter anderem darin, dass Cluster berechnet und die Klassifikationsobjekte¹ diesen deterministisch zugeordnet werden. »Deterministisch« bedeutet, dass jedes Klassifikationsobjekt mit einer Wahrscheinlichkeit von 0 oder 1 einem oder mehreren Clustern angehört. Bezüglich der Zuordnung lassen sich die in diesem Kapitel dargestellten Verfahren weiter unterscheiden in:

- *Nichtüberlappende bzw. überlappungsfreie Clusteranalyseverfahren*: Die Cluster werden so gebildet, dass jedes Klassifikationsobjekt nur einem Cluster angehört. Diese Verfahren werden auch als disjunktive Clusteranalyseverfahren bezeichnet (Lohse, Ludwig u. a. 1982, S. 396; Sodeur 1974, S. 10).
- *Überlappende Clusteranalyseverfahren*: Die Cluster werden so gebildet, dass ein oder mehrere Klassifikationsobjekte mehreren Clustern angehören können.

Inhaltlich ist bei der Entscheidung für ein überlappendes Verfahren zu beachten, dass mit dem Anteil der Überlappungen die Heterogenität zwischen den Clustern abnimmt. Besteht zwischen zwei Clustern ein hoher Anteil an Überlappungen, sind die Cluster nicht deutlich voneinander getrennt. Die Grundforderung an eine Klassifikation, dass die Cluster voneinander gut getrennt sein sollen (siehe dazu Abschnitt 1.2), ist in einem geringen Ausmaß erfüllt. Umgekehrt kann das Zulassen von Überlappungen zu einer geringeren Clusteranzahl führen und den Daten »besser« angepasst sein. In der Forschungspraxis wird man daher mitunter – abhängig von der Fragestellung, den Daten und der zur Verfügung stehenden Software – ein überlappungsfreies und überlappendes Verfahren anwenden und aufgrund der Ergebnisse beurteilen, welches Modell den Daten angemessener ist.

¹ Zur Erinnerung: Klassifikationsobjekte können sowohl Objekte (Zeilen einer Datenmatrix) als auch Variablen (Spalten einer Datenmatrix) sein.

Folgende in diesem Kapitel behandelten Verfahren liefern überlappungsfreie Cluster:

- Complete- und Single-Linkage (siehe Abschnitte 9.1 und 9.2)
- Mittelwertverfahren (Average-Linkage, Weighted-Average-Linkage Within-Average-Linkage, siehe Abschnitt 9.5)
- Median-, Zentroid- und Ward-Verfahren (siehe Kapitel 11)
- verallgemeinerte Nächste-Nachbarn-Verfahren (siehe Abschnitt 9.4)
- K-Means-Verfahren (siehe Kapitel 12)

Überlappende Clusterstrukturen können mit dem

- Complete-Linkage für überlappende Cluster (siehe Abschnitt 9.3) und dem
- Repräsentanten-Verfahren (siehe Kapitel 10)

aufgefunden werden.

6.2 Grundvorstellungen über die zu bildenden Cluster

Den behandelten Verfahren, die nur eine Auswahl darstellen,² liegt im Wesentlichen eines der folgenden Prinzipien zur Bildung der Cluster zugrunde:

- Nächste-Nachbarn-Verfahren (Single-, Complete- und verallgemeinerte Nächste-Nachbarn-Verfahren)
- Mittelwertmodelle (Average-, Weighted-Average- und Within-Average-Linkage)
- Klassifikationsobjekte als Repräsentanten der Cluster (Repräsentanten-Verfahren)
- Clusterzentren als Repräsentanten (Verfahren zur Konstruktion von Clusterzentren, Median-, Zentroid-, Ward- und K-Means-Verfahren)

Nächste-Nachbarn-Verfahren: Die Cluster werden so gebildet, dass a) jedes Klassifikationsobjekt eine bestimmte Anzahl von nächsten Nachbarn in dem Cluster hat, dem es angehört, oder dass b) jedes Klassifikationsobjekt in dem Cluster zumindest einen B-ten nächsten Nachbarn (zum Beispiel einen dritt nächsten Nachbarn) hat. »Nachbar« bedeutet, dass sich zwei Objekte ähnlich sind. In dem Zugang a) wird ein Klassifikationsobjekt j als nächster Nachbar des Klassifikationsobjekts i bezeichnet, wenn es zu dem Klassifikationsobjekt i eine (Un-)Ähnlichkeit kleiner/gleich bzw. größer/gleich einem

² Bijnen (1973) führt in seiner Überblicksarbeit über 40 Clusteranalyseverfahren an, Bailey (1975) nennt über 20 Verfahren. Allen diesen Verfahren liegt aber eine der nachfolgend behandelten Grundvorstellungen zugrunde.

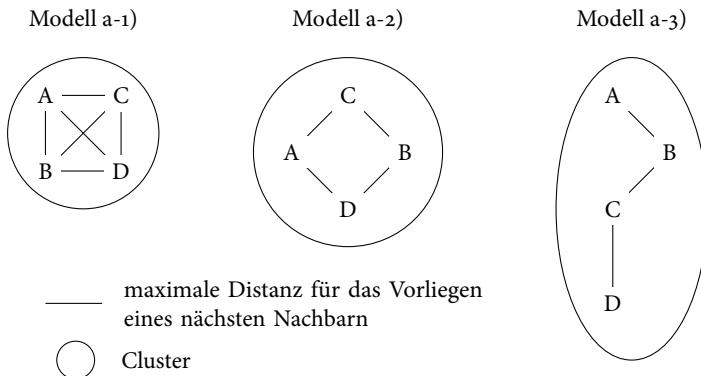


Abb. 6.1: Modellvorstellung der Nächste-Nachbarn-Verfahren

bestimmten Schwellenwert aufweist. Abbildung 6.1 veranschaulicht diese Modellvorstellung graphisch. Im Modell a-1) wird gefordert, dass alle Klassifikationsobjekte eines Clusters nächste Nachbarn zueinander sein müssen. Dieser Modellansatz entspricht dem *Complete-Linkage (Methode des weitest entfernten Nachbarn, Maximum-Methode)*. Es liegt eine sehr strenge Forderung hinsichtlich der Homogenität innerhalb eines Clusters vor. Da hier alle Klassifikationsobjekte eines Clusters nächste Nachbarn zueinander sein müssen, werden die Cluster – in Anlehnung an die soziometrische Literatur – auch als *Cliquen* bezeichnet. Homogenität in einem Cluster liegt dann vor, wenn alle Objekte eines Clusters zueinander nächste Nachbarn sind. Beim Modellansatz a-2) soll jedes Klassifikationsobjekt eines Clusters zumindest zwei nächste Nachbarn bzw. allgemein k nächste Nachbarn haben. Dies ist der Zugang der *verallgemeinerten Nächste-Nachbarn-Verfahren*. Homogenität in einem Cluster liegt somit dann vor, wenn alle Objekte zumindest k nächste Nachbarn im Cluster haben.³ Die Modellvorstellung a-3) liegt dem *Single-Linkage (Methode des nächsten Nachbarn, Minimum-Methode)* zugrunde. Es wird gefordert, dass jedes Klassifikationsobjekt zumindest einen nächsten Nachbarn im Cluster besitzen soll. Es liegt also eine sehr schwache Vorstellung der Homogenität innerhalb der Cluster vor. Die gebildeten Cluster werden daher auch – in Anlehnung an die Stichprobentheorie – als *Klumpen* bezeichnet. Homogenität in einem Cluster liegt dann vor, wenn alle Objekte zumindest einen nächsten Nachbarn im Cluster haben.

Mittelwertmodelle: Die Cluster werden durch die durchschnittliche paarweise Ähnlichkeit bzw. Unähnlichkeit der Klassifikationsobjekte innerhalb der Cluster und/oder zwis-

³ Eine andere Modellvorstellung der *verallgemeinerten Nächste-Nachbarn-Verfahren* geht von folgender Überlegung aus: Bei n Klassifikationsobjekten hat jedes Klassifikationsobjekt $n - 1$ Nachbarn, einen erst-nächsten Nachbarn, einen zweitnächsten Nachbarn usw. Bezuglich eines zu bildenden Clusters wird nun gefordert, dass jedes Klassifikationsobjekt mindestens einen B-ten Nachbarn im Cluster hat.

schen den Clustern charakterisiert. Dieser Gruppe von Modellen gehören an: *Average*-, *Weighted-Average-Linkage* und *Within-Average-Linkage*. Gefordert wird, dass die durchschnittlichen Ähnlichkeiten ein bestimmtes Ausmaß überschreiten.

Klassifikationsobjekte als Repräsentanten der Cluster (Repräsentanten-Verfahren): Jedes Cluster k soll durch ein typisches Klassifikationsobjekt (Repräsentant) charakterisiert sein. Bei der Clusterbildung wird also nach typischen Klassifikationsobjekten gesucht. Die verbleibenden, nicht typischen Klassifikationsobjekte werden jenem Repräsentanten zugeordnet, zu dem sie am ähnlichsten sind, oder sie bleiben unklassifiziert, da die Ähnlichkeit zum nächsten Repräsentanten zu klein ist. Homogenität in einem Cluster wird hier durch die Ähnlichkeit eines Objekts zum Repräsentanten des Clusters erfasst.

Clusterzentren als Repräsentanten (Verfahren zur Konstruktion von Clusterzentren): Hier wird angenommen, dass nicht ein einzelnes Klassifikationsobjekt Repräsentant eines Clusters ist, sondern dass ein Cluster durch seine Clusterzentren (Mittelwerte der in die Clusterbildung einbezogenen Variablen) charakterisiert werden kann. Von diesem Modellansatz gehen das *Median*-, *Zentroid*- und *Ward*-*Verfahren* sowie die *K-Means-Verfahren* aus. Die Cluster werden so bestimmt, dass a) die Clusterzentren maximal voneinander entfernt sind (*Median*- und *Zentroid*-Verfahren) oder dass b) die Streuung zwischen den Clusterzentren maximiert wird (*Ward*-Verfahren, *K-Means*-Verfahren). Homogenität wird hier durch die quadratischen Abweichungen vom Clustermittelwert bzw. -zentrum definiert.

6.3 Complete- und Single-Linkage als Basismodelle

Bereits aus der sehr groben Modellcharakterisierung des Abschnitts 6.2 lassen sich folgende Anwendungsvoraussetzungen für die einzelnen Verfahren ableiten:

- Sollen *Clusterzentren als Repräsentanten* bei der Clusterbildung verwendet werden, muss eine Datenmatrix mit quantitativen Variablen vorliegen, da strenggenommen nur für quantitative Variablen Mittelwertbildungen erlaubt sind. Ferner ist die Verwendung eines bestimmten Unähnlichkeitsmaßes, nämlich der quadrierten euklidischen Distanz, erforderlich. Wie wir später sehen werden, können Clusterzentren auch für nominale und ordinale Variablen berechnet werden (siehe Abschnitt 12.10). Auch andere Distanzmaße können eingesetzt werden.
- Bei den *Mittelwertverfahren* (*Average*-, *Weighted-Average* und *Within-Average-Linkage*) werden Mittelwerte (Durchschnitte) aus den (Un-)Ähnlichkeiten gebildet. Bei

der Clusterbildung wird also angenommen, dass die (Un-)Ähnlichkeiten metrischen Informationsgehalt aufweisen, dass also Aussagen der Art »Die Ähnlichkeit zwischen A und B ist doppelt so groß wie zwischen A und C« möglich sind.⁴

Wir können die behandelten Verfahren nach folgenden Kriterien charakterisieren:

1. *Ist das Vorliegen einer Datenmatrix erforderlich?* Ist dies der Fall, kann eine direkt erhobene (Un-)Ähnlichkeitsmatrix nicht untersucht werden.
2. *Müssen quantitative Variablen vorliegen?* Das Messniveau spielt allerdings eine untergeordnete Rolle (siehe dazu Abschnitt 7.5).
3. *Ist die Verwendung eines bestimmten (Un-)Ähnlichkeitsmaßes erforderlich?* Auch bezüglich dieses Kriteriums sind die Verfahren relativ robust (siehe dazu Abschnitt 8.1).
4. *Muss das verwendete (Un-)Ähnlichkeitsmaß metrischen Informationsgehalt besitzen?*

Hinsichtlich dieser vier Kriterien treffen der Complete- und Single-Linkage sowie der Complete-Linkage für überlappende Cluster und die Repräsentanten-Verfahren die »schwächsten« Modellvoraussetzungen, wobei *Complete- und Single-Linkage als Basismodelle betrachtet werden können*. »Schwach« bedeutet dabei »wenig« Voraussetzungen hinsichtlich des erforderlichen Datenmaterials. Diese sind für den Complete- und Single-Linkage:

1. *Das Vorliegen einer Datenmatrix ist nicht erforderlich.* Es kann auch eine direkt erhobene (Un-)Ähnlichkeitsmatrix untersucht werden. Diese kann nichtmetrisch sein.
2. *Sofern eine Datenmatrix vorliegt, können die Variablen, die zur Klassifikation verwendet werden, nichtmetrisch sein.*
3. *Die Verwendung eines bestimmten (Un-)Ähnlichkeitsmaßes ist nicht erforderlich.* Es können beliebige Distanzmaße eingesetzt werden.
4. *Bei der Clusterbildung wird nur die ordinale Information der (Un-)Ähnlichkeiten zwischen den Klassifikationsobjekten verwendet.* Das heißt nur die ordinale Information, dass das Objektpaar (g, g^*) ähnlicher ist als ein anderes Objektpaar (g^{**}, g^{***}) fließt in die Berechnung ein. Die Ergebnisse sind daher gegenüber monotonen Transformationen der berechneten oder erhobenen empirischen (Un-)Ähnlichkeiten invariant. Die Werte der (Un-)Ähnlichkeitsmatrix können also quadriert, linear transformiert, logarithmiert usw. werden, ohne dass sich die empirische Klassifikation ändert.
5. *Bei der Clusterbildung entsteht eine monoton hierarchische Struktur zwischen den Klassifikationsobjekten und Clustern, die die formale Eigenschaft einer Ultrametrik*

⁴ Strenggenommen genügt Intervallskalenqualität, das heißt, dass die Differenzen der Ähnlichkeiten definiert sind.

besitzt (Batagelj 1981; Jain und Dubes 1988, S. 68–70; Milligan 1979; u. a.). Diese Eigenschaften können für Modelltests genutzt werden.

Alle anderen Verfahren kann man sich als Modifikationen dieser beiden Verfahren vorstellen, wobei versucht wird, bestimmte Nachteile zu verbessern. Die Nachteile der beiden Verfahren sind: Das *Single-Linkage-Verfahren* führt oft zu so genannten Verkettungen, so dass Cluster, die voneinander »verschieden« sind, verschmolzen werden. Dieser Effekt wird als *Verkettungseffekt* oder *Kontraktionseffekt* bezeichnet (Steinhäusen und Langer 1977, S. 75): Der von den Variablen aufgespannte Klassifikationsraum wird zusammengezogen. Die Ursache dafür ist, dass der Single-Linkage von einer zu »schwachen« Vorstellung der Homogenität in den Clustern ausgeht. Dies lässt sich für die Analyse von Ausreißern nutzen (siehe Abschnitt 12.6). Der *Complete-Linkage* führt dagegen oft dazu, dass sehr viele Cluster gebildet werden, da er von einer sehr »strengen« Vorstellung hinsichtlich der Homogenität in den Clustern ausgeht. Dieser Effekt wird als *Dilatationseffekt* bezeichnet (Steinhäusen und Langer 1977, S. 75). Beide Verfahren sind nur für eine *kleine oder mittlere Klassifikationsobjektmenge* geeignet.⁵ Der Einsatz hierarchischer Verfahren ist bei großen Datensätzen – genaue Angaben, was »groß« ist, sind nicht möglich – nicht sinnvoll, da in frühen Phasen der Verschmelzung Bindungen auftreten (zwei oder mehrere Clusterpaare haben dieselbe Ähnlichkeit), die einen starken Einfluss auf die nachfolgenden Ergebnisse haben.

Die anderen Verfahren versuchen diese Nachteile zu kompensieren über:

- *Mittelung der Distanzen*: Dies ist die Strategie der Mittelwertverfahren (siehe Kapitel 9). Der »Preis« besteht darin, dass die Invarianzeigenschaft gegenüber monotonen Transformationen verloren geht – die (Un-)Ähnlichkeitsmatrix muss metrisch (intervallskaliert) sein.
- *Suche nach Clusterzentren*: Dies ist die Strategie der hierarchischen Verfahren zur Konstruktion von Clusterzentren (siehe Kapitel 11) und des K-Means-Verfahren (siehe Kapitel 12). Der »Preis« besteht hier darin, dass eine Datenmatrix mit – strenggenommen – quantitativen Variablen erforderlich ist. Zur Ähnlichkeitsmessung muss die quadrierte euklidische Distanz eingesetzt werden.
- *Partitionierung statt Hierarchie*: Dies ist die Strategie des K-Means-Verfahrens (siehe Kapitel 12). Das Verfahren ist für große Datensätze sehr gut geeignet. Der »Preis« bei dieser Methode ist, dass die Hierarchie-Eigenschaft verloren geht. Erforderlich sind – strenggenommen – quantitative Variablen und die quadrierte euklidische Distanz.

⁵ Früher war die Anwendung durch den zur Verfügung stehenden Arbeitsspeicher des Computers eingeschränkt, da die hierarchisch-agglomerativen Verfahren rechentechnisch voraussetzen, dass eine (Un-)Ähnlichkeitsmatrix im Arbeitsspeicher gehalten werden muss. Dieser Nachteil spielt heute angesichts leistungsfähiger Personal-Computer de facto keine Rolle mehr.

Allerdings gibt es für das K-Means-Verfahren Weiterentwicklungen für Variablen gemischten Messniveaus (siehe Abschnitt 12.10).

6.4 Auswahl eines geeigneten Verfahrens

Bei der Beantwortung der Frage, welches Verfahren für die Klassifizierungsaufgabe geeignet ist, sind zwei unterschiedliche Aufgabenstellungen zu unterscheiden:

1. *Auffinden einer hierarchischen Ähnlichkeitsstruktur:* Es wird untersucht, ob einer Ähnlichkeitsmatrix eine hierarchische Struktur in Form eines Baums zugrunde liegt. Hierfür eignen sich nur die hierarchisch-agglomerativen Verfahren.
2. *Auffinden einer empirischen Klassifikation:* Es wird untersucht, ob eine Klassifikation (Clusterstruktur) vorliegt und wie die Objekte diesen zugeordnet sind. Dies ist die primäre Zielsetzung von clusteranalytischen Verfahren. Hierfür können im Prinzip alle hier behandelten Verfahren eingesetzt werden.

Soll eine *hierarchische Ähnlichkeitsstruktur* gefunden werden, eignen sich nur die hierarchisch-agglomerativen Verfahren (Single- und Complete-Linkage, Mittelwertverfahren, Median-, Zentroid- und Ward-Verfahren). Man wird in Abhängigkeit von der Forderung an die gesuchte Hierarchie anwenden, wenn

1. *Invarianz gegenüber monotonen Transformationen erwünscht ist:* Complete- und Single-Linkage (siehe Abschnitte 9.1 und 9.2).
2. *Invarianz gegenüber monotonen Transformationen nicht erwünscht ist*, in Abhängigkeit vom Datenmaterial und der Richtung der Datenanalyse:
 - bei einer direkt erhobenen, nichtmetrischen (Un-)Ähnlichkeitsmatrix: Complete- und Single-Linkage (siehe Abschnitte 9.1 und 9.2).
 - bei einer direkt erhobenen, metrischen (Un-)Ähnlichkeitsmatrix: Mittelwertverfahren (Average-, Weighted-Average-, Within-Average-Linkage, siehe Abschnitt 9.5), da bei diesen Verfahren die zu schwachen bzw. zu strengen Homogenitätsvorstellungen des Single- und Complete-Linkage abgeschwächt werden.
 - beim Vorliegen einer Datenmatrix und einer variablenorientierten Datenanalyse: Mittelwertverfahren (Average-, Weighted-Average-, Within-Average-Linkage, siehe Abschnitt 9.5).
 - beim Vorliegen einer Datenmatrix und einer objektorientierten Datenanalyse: Mittelwertverfahren (Average-, Weighted-Average, Within-Average-Linkage, siehe Abschnitt 9.5) und Median-, Zentroid- und Ward-Verfahren (siehe Kapitel 11).

Soll eine *empirische Klassifikation* gefunden werden, wird man anwenden,

1. wenn die gesuchte Klassifikation invariant gegenüber monotonen Transformationen sein soll: Complete-, Single-Linkage (siehe Abschnitt 9.1), verallgemeinerte Nächste-Nachbarn-Verfahren (siehe Abschnitt 9.4) und Repräsentanten-Verfahren (siehe Kapitel 10; die Invarianzeigenschaft des Repräsentanten-Verfahrens gegenüber monotonen Transformationen gilt nur bedingt).
2. wenn Invarianz gegenüber monotonen Transformationen nicht erwünscht ist, in Abhängigkeit vom Datenmaterial und der Richtung der Datenanalyse:
 - bei einer direkt erhobenen, nichtmetrischen (Un-)Ähnlichkeitsmatrix: Complete- und Single-Linkage (siehe Abschnitt 9.1) sowie verallgemeinerte Nächste-Nachbarn- und Repräsentanten-Verfahren (siehe Abschnitt 9.4 und Kapitel 10).
 - bei einer direkt erhobenen, metrischen (Un-)Ähnlichkeitsmatrix: Mittelwertverfahren (Average-, Weighted-Average-, Within-Average-Linkage, siehe Abschnitt 9.5) sowie verallgemeinerte Nächste-Nachbarn- und Repräsentanten-Verfahren (siehe Abschnitt 9.4 und Kapitel 10).
 - beim Vorliegen einer Datenmatrix und einer variablenorientierten Datenanalyse: Mittelwertverfahren (Average-, Weighted-Average-, Within-Average-Linkage, siehe Abschnitt 9.5) sowie verallgemeinerte Nächste-Nachbarn- und Repräsentanten-Verfahren (siehe Abschnitt 9.4 und Kapitel 10).
 - beim Vorliegen einer Datenmatrix und einer objektorientierten Datenanalyse: Ward-Verfahren (siehe Abschnitt 11.1) und bei großen Datensätzen insbesondere die K-Means-Verfahren (siehe Kapitel 12), einsetzbar sind auch Mittelwertverfahren (Average-, Weighted-Average-, Within-Average-Linkage; siehe Abschnitt 9.5), Median- und Zentroid-Verfahren (siehe Kapitel 11), verallgemeinerte Nächste-Nachbarn-Verfahren (siehe Abschnitt 9.4) und Repräsentanten-Verfahren (siehe Kapitel 10).

Allgemein ist zu den Regeln anzumerken:

1. Das *Auffinden einer hierarchischen (Un-)Ähnlichkeitsbeziehung* ist nur sinnvoll, wenn die *analysierten Einheiten eine inhaltliche Bedeutung* haben. Dies ist immer der Fall, wenn Variablen untersucht werden. Werden dagegen Objekte (Zeilen der Datenmatrix) untersucht, trifft dies allgemein nicht zu. So besitzen zum Beispiel zwar politische Parteien oder Nationen eine inhaltliche Bedeutung, nicht aber die in einer Umfrage befragten Personen. Deshalb ist es nicht sinnvoll, eine hierarchische Ähnlichkeitsbeziehung zwischen den Befragten einer Umfrage zu bestimmen.

2. Zum *Auffinden einer hierarchischen (Un-)Ähnlichkeitsstruktur* eignen sich *nur die hierarchisch-agglomerativen Verfahren*.⁶
3. Wird *eine erhobene (Un-)Ähnlichkeitsmatrix* untersucht, *scheiden das Median-, Zentroid- und Ward-Verfahren sowie die K-Means-Verfahren aus*, da diese das Vorliegen einer Datenmatrix voraussetzen.
4. Bei einer *variablenorientierten Clusteranalyse* (Auffinden einer Klassifikation von Variablen) *ist die Anwendung des Median-, Zentroid- und Ward-Verfahrens sowie der K-Means-Verfahren nicht sinnvoll*, da hier nicht Clusterzentren gefunden werden sollen. Dagegen kann es sinnvoll sein, eine einzelne Variable als Repräsentanten auszuwählen und somit ein Repräsentanten-Verfahren einzusetzen.
5. Werden aus *einer Datenmatrix (Un-)Ähnlichkeitsmaße berechnet*, so sind diese in der Regel metrisch. Man wird hier daher in der Regel *nicht den Complete- oder Single-Linkage* wegen deren zu strengen bzw. zu schwachen Homogenitätsvorstellungen anwenden, außer diese oder Invarianz gegenüber monotonen Transformationen ist erwünscht.
6. Soll eine *überlappende Klassifikation* gefunden werden, können von den hier behandelten deterministischen Verfahren *nur der Complete-Linkage für überlappende Cluster und Repräsentanten-Verfahren* verwendet werden.⁷
7. Kann aufgrund von inhaltlichen Überlegungen *die quadrierte euklidische Distanz unter keinen Umständen verwendet werden, scheiden K-Means-Verfahren und das Median-, Zentroid- und Ward-Verfahren aus*. Allerdings sind die Verfahren relativ robust gegenüber der Wahl eines »falschen« Unähnlichkeitsmaßes (siehe dazu Abschnitt 6.7 und 8.1).
8. Eine metrische (Un-)Ähnlichkeitsmatrix kann durch ein vorausgehendes graphischen Verfahren ermittelt werden, zum Beispiel durch eine *mehrdimensionale Skalierung* oder eine *multiple Korrespondenzanalyse*. Damit können dann Mittelwertverfahren zum Einsatz kommen, die den Kontraktions- und Dilatationseffekt vermeiden.

⁶ Das bei den hierarchisch-agglomerativen Verfahren verwendete Modell der ultrametrischen Hierarchie trifft sehr strenge Anforderungen hinsichtlich der gesuchten Hierarchie. In den 1970er Jahren wurden daher Verfahren zur Konstruktion additiver Hierarchien mit weniger strengen Anforderungen entwickelt (Cunningham 1978; Sattath und Tversky 1977; u. a.).

⁷ Überlappende Klassifikationen können auch mit den im Teil III behandelten probabilistischen Verfahren gefunden werden. Zu beachten ist dabei allgemein, dass mit dem Grad der Überlappungen die Heterogenität zwischen den Clustern abnimmt (siehe dazu die Ausführungen in Abschnitt 6.1).

6.5 Lösungsschritte einer Klassifikationsaufgabe

Das Auffinden einer objektorientierten Klassifikation, also eine Clusterung der Zeilen einer Datenmatrix, ist das primäre Anwendungsfeld der Clusteranalyse. Zur Lösung dieser Zielsetzung sind folgende Analyseschritte durchzuführen:

1. *Auswahl der Variablen* (Spalten der Datenmatrix).
2. *Auswahl der Objekte* (Zeilen der Datenmatrix).
3. *Spezifikation der Eigenschaften, die die Klassifikation erfüllen soll*: Soll die gesuchte Klassifikation invariant gegenüber monotonen Transformationen sein? Soll die Klassifikation überlappungsfrei sein? usw.
4. *Auswahl eines Verfahrens*, das die spezifizierten Eigenschaften besitzt.
5. *Transformation und Gewichtung der Variablen* zur Beseitigung der Nichtvergleichbarkeit.
6. *Auswahl eines Ähnlichkeits- oder Unähnlichkeitsmaßes*, sofern das Verfahren nicht die Verwendung eines bestimmten (Un-)Ähnlichkeitsmaßes erfordert. Die Auswahl eines geeigneten (Un-)Ähnlichkeitsmaßes ist erforderlich bei: Complete- und Single-Linkage, Complete-Linkage für überlappende Cluster, verallgemeinerte Nächste-Nachbarn-Verfahren, Mittelwertverfahren und Repräsentanten-Verfahren.
7. *Durchführen der Clusteranalyse*.
8. *Bestimmung der Clusterzahl*.
9. *Prüfung der Modellanpassung*.
10. *Beschreibung und inhaltliche Interpretation* der Cluster.
11. *Inhaltliche Validitätsprüfung* für die gefundenen Clusterlösungen.
12. *Stabilitätstests* für die gefundenen Clusterlösungen.
13. *Formale Gültigkeitsprüfung* für die gefundenen Clusterlösungen.

Die Abfolge der Schritte nach der Prüfung der Modellanpassung können variiert werden. So zum Beispiel ist es möglich, vor der Beschreibung und inhaltlichen Interpretation der Cluster und der inhaltlichen Validitätsprüfung die Stabilitäts- und formale Gültigkeitsprüfung durchzuführen. Die Probleme und Fehlerquellen, die bei den einzelnen Schritten auftreten können, werden in den nachfolgenden beiden Abschnitten behandelt.

6.6 Ein Anwendungsbeispiel

Da im Unterschied zu Teil 1 nach der einleitenden Übersicht nicht unmittelbar ein Verfahren mit einem Anwendungsbeispiel behandelt wird, sollen die im vorausgehenden Abschnitt genannten Verfahrensschritte hier dargestellt werden. Ziel der Analyse

ist eine Klassifikation der Länder Mittel- und Südamerikas aufgrund ihrer sozialen, demographischen und wirtschaftlichen Entwicklung in den 1980er Jahren. Aus dieser Zielformulierung folgt:

Auswahl der Variablen: In die Analyse sollen Indikatoren der sozialen, demographischen und wirtschaftlichen Entwicklung für die 1980er Jahre eingehen (Nohlen 1984, S. 630–634). Diese sind (siehe dazu Tabelle 5.12 auf Seite 137): »Bruttonsozialprodukt pro Einwohner« (BruSozPr), »Wirtschaftswachstum in den 1980er Jahren« (Wachst80), »Industrialisierungsquote in den 1980er Jahren« (Indust80), »Exportanteil der Rohstoffe« (Export), »Schuldenbelastung« (Schulden), »Nahrungsmittelimport« (ImpNahr), »Kalorienversorgung« (Kalorien), »Lebenserwartung« (LebErw), »Kindersterblichkeit« (Kindsterb), »Alphabetisierungsquote in den 1980er Jahren« (Alpha80), »Einschulungsquote« (Einschulung), »Einwohner pro Arzt« (EinwproArzt) und »Bevölkerungswachstum in den 1980er Jahren« (BevZu80). Insgesamt werden also 13 Variablen in die Analyse einbezogen, sechs Variablen (BruSozPr bis ImpNahr) messen in einem weitesten Sinn die wirtschaftliche Entwicklung, die verbleibenden sieben Variablen (Kalorien bis BevZu80) die soziale und demographische Entwicklung.

Auswahl der Objekte: Alle Länder Süd- und Mittelamerikas sollen geclustert werden. Des Weiteren wurden für die Analyse folgende Entscheidungen getroffen:

Spezifikation der Eigenschaften, die die gesuchte Klassifikation erfüllen soll: Hinsichtlich der gesuchten Klassifikation nehmen wir an, dass sie nicht überlappend sein soll und die Cluster durch Clusterzentren (Mittelwerte) gebildet werden sollen.

Auswahl eines geeigneten Verfahrens: Das Ward-Verfahren erfüllt neben anderen Verfahren (Median-, Zentroid- und K-Means-Verfahren) diese Eigenschaften. Gegenüber den anderen hierarchischen Verfahren (Median- und Zentroid-Verfahren) hat es den Vorteil, dass das Verschmelzungsschema klar interpretierbar ist und Inversionen vermieden werden (siehe Kapitel 11). Im Unterschied zum K-Means-Verfahren eignet es sich auch für kleine Datensätze und liefert eine hierarchische Darstellung.

Transformation und Gewichtung der Variablen: Eine Transformation und Gewichtung der Variablen ist erforderlich, da diese nicht vergleichbar sind (siehe dazu Abschnitt 7.1). Nichtvergleichbarkeit ist aus zwei Gründen gegeben:

- Die Variablen sind in unterschiedlichen Skaleneinheiten gemessen. Das Bruttonsozialprodukt pro Kopf beispielsweise ist in Dollar gemessen, das Wirtschaftswachstum dagegen in Prozentpunkten.
- Die latenten Dimensionen soziale, demographische und wirtschaftliche Entwicklung sind über- bzw. unterrepräsentiert. Die soziale und demographische Entwicklung wird durch sieben Indikatoren erfasst, die wirtschaftliche durch sechs Indikatoren.

Tab. 6.1: Rotierte Faktorladungen der Faktorenanalyse (Ergebnisse der Quartimin-Rotation)

Variable	Faktor 1	Faktor 2	Faktor 3	Faktor 4
<i>Faktorladungen</i>				
Kalorien	0,8231	-0,2146	0,2931	0,5285
LebErw	0,9554	-0,1987	0,5912	0,5653
Kindsterb	-0,9245	0,0182	-0,4747	-0,4742
Wachst80	-0,0539	0,8053	-0,1849	0,1044
Export	-0,2439	0,6385	0,3099	0,0793
ImpNahr	-0,3971	-0,5487	-0,2362	-0,4276
BevZu80	-0,4383	0,7068	-0,4287	-0,1745
BruSozPr	0,6256	0,0241	0,8307	0,4799
Indust80	0,2768	-0,1337	0,9078	0,4445
Schulden	-0,0454	0,1383	-0,1007	0,6448
Alpha80	0,7188	0,0315	0,6869	0,8903
Einschulung	0,6311	0,1214	0,4177	0,8434
EinwproArzt	-0,4939	-0,1591	-0,6145	-0,9165
<i>Korrelation der Faktoren</i>				
Faktor 1	1,0000			
Faktor 2	-0,1568	1,0000		
Faktor 3	0,5092	-0,0670	1,0000	
Faktor 4	0,5781	0,1320	0,5298	1,0000

grau hinterlegt: höchste Faktorladung jeder Ausprägung auf den Faktoren

Die Nichtvergleichbarkeit kann in dem Beispiel durch eine vorausgehende R-Faktorenanalyse mit anschließender schiefwinkliger Rotation (siehe Abschnitt 5.3.1) gelöst werden. Die Ergebnisse sind in Tabelle 6.1 zusammengefasst: Es werden vier Faktoren mit Eigenwerten größer 1 berechnet. Entsprechend dem Kaiser-Kriterium (siehe Abschnitt 5.3.1) liegen somit vier bedeutsame Faktoren vor. Diese erklären insgesamt 78,1 Prozent der Gesamtvarianz. Wegen des starken Eigenwertabfalls zwischen dem ersten und zweiten Faktor (Eigenwert des ersten Faktors: 5,2756; Eigenwert des zweiten Faktors: 2,1598) könnte man sich auch für eine einfaktorielle Lösung entscheiden und den ersten Faktor als allgemeine Entwicklungsdimension interpretieren. Der dabei entstehende Informationsverlust (erklärte Varianz: 40,6 Prozent) erscheint aber zu groß, so dass die vierfaktorielle Lösung weiter untersucht wird. Die rotierte Lösung weist keine perfekte Einfachstruktur auf. So lädt zum Beispiel die Variable Import von Nahrungsmitteln (ImpNahr) neben dem zweiten Faktor auch noch auf dem ersten und vierten Faktor relativ stark. Dies gilt auch für das Pro-Kopf-Bruttosozialprodukt (BruSozPr) und für andere Variablen. Trotz dieser Einschränkungen lässt sich die vierfaktorielle Lösung gut interpretieren. Den Faktoren können folgende Namen gegeben werden:

Faktor 1: Befriedigung existentieller Grundbedürfnisse (ExGrundbed), wie zum Beispiel ausreichende Kalorienversorgung

Faktor 2: wirtschaftliches Wachstum (wirtWachst)

Faktor 3: Industrialisierung (Indust)

Faktor 4: erreichtes Bildungsniveau (Bildung)⁸

Die vier Entwicklungsdimensionen sind nicht unabhängig voneinander. Relativ hohe Korrelationen (größer 0,5) bestehen zwischen der Befriedigung existentieller Grundbedürfnisse und der erreichten Industrialisierung, zwischen der Befriedigung existentieller Grundbedürfnisse und dem erreichten Bildungsniveau sowie zwischen der Industrialisierung und dem Bildungsniveau. Das Wirtschaftswachstum korreliert dagegen nur schwach mit den anderen drei Dimensionen, was sich unter anderem dadurch erklären lässt, dass sich das Wirtschaftswachstum zu einem nicht unbeträchtlichen Anteil auf den Export von Rohstoffen zurückführen lässt, also auf einen Faktor, der nur bedingt mit innerstaatlichen Entwicklungen in den anderen Dimensionen zu tun hat. Für die vier Faktoren werden für die weitere Analyse Faktorwerte berechnet. Eine Berechnung von mittleren Gesamtpunktwerten ist nicht möglich, da die Variablen nicht theoretisch standardisiert werden können und keine perfekte Einfachstruktur vorliegt. Die berechneten Faktorwerte sind vergleichbar, da sie in der gleichen Skaleneinheit mit Mittelwert 0 und Varianz 1 gemessen werden. Ferner hat jeder Faktor in der nachfolgenden Analyse dasselbe Gewicht. Über- bzw. Unterrepräsentativität liegt nicht vor. Durch das Vorgehen sind somit beide Gründe der Nichtvergleichbarkeit gelöst.

Auswahl eines geeigneten Ähnlichkeits- oder Unähnlichkeitsmaßes: Der Schritt der Auswahl eines (Un-)Ähnlichkeitsmaßes entfällt, da das Ward-Verfahren die Verwendung der quadrierten euklidischen Distanzen voraussetzt (siehe dazu Kapitel 11). In Tabelle 6.2 auf der nächsten Seite ist die Unähnlichkeitsmatrix der quadrierten euklidischen Distanzen auszugsweise wiedergegeben. Ein größerer Zahlenwert bedeutet eine größere Unähnlichkeit. Die größte Ähnlichkeit (kleinste Unähnlichkeit: 0,0959) besteht zwischen Kuba und Argentinien, die geringste Ähnlichkeit bzw. größte Unähnlichkeit liegt für das Länderpaar Argentinien und Haiti vor. Die quadrierte euklidische Distanz beträgt 34,7585.

Durchführen der Clusteranalyse: Das Ward-Verfahren gehört der Gruppe der hierarchisch-agglomerativen Verfahren an. Das bedeutet, dass die Objekte schrittweise zu Clustern verschmolzen werden. Im ersten Schritt werden aus den 22 Ländern 21 Cluster gebildet, im zweiten Schritt werden durch eine Verschmelzung von zwei aus den 21 Clustern 20

⁸ Obwohl die Variable Schuldenanteil ebenfalls auf dieser Dimension lädt, wurde sie nicht zur Namensgebung verwendet, da die Variablen Alphabetisierungs- und Einschulungsquote (Alpha80 und Einschulung) sowie die Variable Einwohner pro Arzt (EinwproArzt), die im engen Zusammenhang mit Investitionen im Bildungssystem stehen, höhere Faktorladungen auf dieser Dimension aufweisen.

Tab. 6.2: Quadrierte euklidische Distanzen zwischen den Ländern Süd- und Mittelamerikas in den vier Faktoren

	Objektnr.	Argent.	Boliv.	Bras.	Chile	Costa-R.	...
Argentinien	1	0	17,7161	3,3735	1,0932	1,2722	...
Bolivien	2	17,7161	0	6,7816	13,8173	12,4229	...
Brasilien	3	3,3735	6,7816	0	3,0817	1,2278	...
Chile	4	1,0932	13,8173	3,0817	0	2,4961	...
Costa Rica	5	1,2722	12,4229	1,2278	2,4961	0	...
Domen. Rep.	6	8,2152	4,3893	1,9437	7,2463	3,5667	...
Ecuador	7	9,1097	4,8443	3,1958	9,4967	5,1790	...
El Salvador	8	10,6580	5,4083	5,0331	8,2511	6,2803	...
Guatemala	9	18,6462	7,3853	11,0745	16,7618	12,1745	...
Haiti	10	34,7585	17,8722	27,0206	26,3027	30,4075	...
Honduras	11	11,5803	2,6100	3,9648	9,9010	6,2923	...
Jamaika	12	6,3726	20,3509	9,4839	3,3320	8,1664	...
Kolumbien	13	2,7849	7,7095	0,4398	2,8743	0,5772	...
Kuba	14	0,0959	17,1245	3,4832	0,6858	1,7266	...
Mexiko	15	1,8682	11,5858	0,6894	2,4658	0,9147	...
Nicaragua	16	7,8833	2,3600	2,2901	5,2217	4,9257	...
Panama	17	0,6941	11,6139	1,1718	1,1773	0,4948	...
Paraguay	18	3,8277	9,6991	1,2467	5,1024	0,7005	...
Peru	19	6,4782	4,3141	2,2647	3,4813	5,4719	...
Trinidad	20	6,0309	24,7428	12,8183	6,6447	9,8310	...
Uruguay	21	1,5001	20,3113	6,3776	0,8518	3,9991	...
Venezuela	22	3,4000	12,0815	3,9316	4,4926	3,2963	...

grau hinterlegt: kleinste und größte Unähnlichkeit

Cluster gebildet, im dritten Schritt 19 usw. Die Ergebnisse der Verschmelzung werden in einem so genannten Verschmelzungsschema protokolliert (siehe Tabelle 6.3). In dem Verschmelzungsschema werden die Objekte fortlaufend mit 1 beginnend durchnummeriert. Die Ziffer 1 steht also für Argentinien, die Ziffer 2 für Bolivien, die Ziffer 3 für Brasilien usw. Das Schema ist wie folgt zu lesen: Im ersten Schritt werden die Objekte 1 (Argentinien) und 14 (Kuba) bei einem Distanzniveau von 0,096 verschmolzen. Dies führt zu 21 Clustern. Im zweiten Schritt werden die Objekte 3 (Brasilien) und 13 (Kolumbien) bei einem Distanzniveau von 0,440 verschmolzen. Es entstehen 20 Cluster. Der Zuwachs im Distanzniveau gegenüber der Lösung mit 21 Clustern beträgt 0,344 (= 0,440 – 0,096) usw. Tatsächlich werden in einem Verschmelzungsschritt nicht Objekte, sondern Cluster verschmolzen, wobei aus Gründen der Übersichtlichkeit jeweils nur das erste Objekt jedes Clusters angeführt wird. In den beiden ersten Schritten bestehen die Cluster jeweils nur aus einem Objekt, so dass die vorgenommene Interpretation zulässig ist. Der fünfzehnte Verschmelzungsschritt (7 Cluster) ist dagegen wie folgt zu interpretieren: Es werden die Cluster, die als erstes Objekt das Objekt 1 (Argentinien) bzw. 12 (Jamaika) enthalten, bei einem Distanzniveau von 6,577 verschmolzen. Aus welchen Objekten die beiden Cluster

Tab. 6.3: Verschmelzungsschema des Ward-Verfahrens für die Entwicklungsländerdaten

Schritt	Verschmelzung der Cluster mit den Objekten	Zahl der Cluster	Verschmelzungs- niveau	Zuwachs
1	1 – 14	21	0,096	0,000
2	3 – 13	20	0,440	0,344
3	5 – 17	19	0,495	0,055
4	6 – 11	18	0,764	0,269
5	4 – 21	17	0,852	0,088
6	3 – 18	16	1,011	0,159
7	5 – 15	15	1,018	0,007
8	16 – 19	14	1,387	0,369
9	6 – 8	13	1,484	0,098
10	1 – 4	12	1,766	0,281
11	3 – 5	11	1,975	0,209
12	7 – 22	10	3,281	1,306
13	6 – 9	9	3,965	0,684
14	2 – 16	8	3,987	0,022
15	1 – 12	7	6,577	2,590
16	3 – 7	6	8,070	1,493
17	2 – 6	5	11,137	3,067
18	1 – 20	4	11,229	0,092
19	2 – 10	3	25,548	14,319
20	1 – 3	2	25,621	0,073
21	1 – 2	1	67,502	41,881

grau hinterlegt: beträchtliche Zunahme im Verschmelzungsniveau

bestehen, kann aus dem Verschmelzungsschema rekonstruiert werden. Die Objekte 1 und 14 werden im ersten Schritt (21 Cluster) zu einem Cluster verschmolzen, die Objekte 4 und 21 im fünften Schritt (17 Cluster). Die beiden Cluster (1, 14) und (4, 21) werden im zehnten Schritt (12 Cluster) zu einem Cluster verschmolzen, so dass im fünfzehnten Schritt (7 Cluster) das Cluster mit dem Objekt 1 als erstem Objekt aus den Objekten 1, 14, 4 und 21 (Argentinien, Kuba, Chile, Uruguay) besteht. Das Cluster 2 besteht dagegen aus einem einzigen Objekt, nämlich Jamaika (Objekt 12), da es bis zum fünfzehnten Schritt im Verschmelzungsschema noch nicht auftritt.

Bestimmung der Clusterzahl: Das Ziel der Datenanalyse ist eine empirische Klassifikation der Länder Süd- und Mittelamerikas. Dies bedeutet zunächst, dass wir aus den bei der Verschmelzung gebildeten 21 Clusterlösungen brauchbare Clusterlösungen auswählen. Die Aufgabenstellung ist somit jener der Bestimmung der Zahl bedeutsamer Dimensionen bei den unvollständigen Verfahren ähnlich. Tatsächlich wird hier auch analog zu

dem Kriterium des Eigenwertabfalls vorgegangen: Das Verschmelzungsschema wird von oben nach unten gelesen und große Distanzzuwächse werden markiert. Diese treten in unserem Beispiel an folgenden Übergängen auf, die in der Tabelle 6.3 auf der vorherigen Seite grau hinterlegt sind:

1. von elf zu zehn Clustern (Distanzzuwachs: 1,306)
2. von acht zu sieben Clustern (Distanzzuwachs: 2,590)
3. von sechs zu fünf Clustern (Distanzzuwachs: 3,067)
4. von vier zu drei Clustern (Distanzzuwachs: 14,319)
5. von zwei Clustern zu einem Cluster (Distanzzuwachs: 41,881)

Es kommen somit fünf Clusterlösungen in Betracht: 11-, 8-, 6-, 4- und 2-Clusterlösung.⁹ Wir wollen hier zunächst die 8-Clusterlösung weiter untersuchen.

Prüfung der Modellanpassung: Zur Messung der Übereinstimmung einer Clusterlösung mit den empirischen Daten wurden eine Reihe von Verfahren und Teststatistiken entwickelt, die noch im Laufe dieses Teils dargestellt werden. Wir wollen hier nur einige dieser Maßzahlen betrachten. Als eine Maßzahl zur Messung der Übereinstimmung zwischen der berechneten Clusterlösung und den untersuchten Daten kann der Korrelationskoeffizient γ berechnet werden (Goodman und Kruskal 1954, siehe auch Tabelle 8.10 auf Seite 212). Er gibt an, in welchem Ausmaß die Distanzen (Unähnlichkeiten) innerhalb der Cluster kleiner sind als jene zwischen den Clustern. Für unser Beispiel ergibt sich ein γ -Koeffizient von 0,915. Dies bedeutet eine sehr gute Modellanpassung. Beim Ward-Verfahren kann ferner die durch die Cluster erklärte Varianz berechnet werden. Die Gesamtstreuungsquadratsumme ist in unserem Beispiel gleich 88,00. Die Streuungsquadratsumme in den acht Clustern ist gleich 11,25. Die durch die 8-Clusterlösung erklärte Streuung ist somit $1 - \frac{11,25}{88,00} = 0,872$ (87,2 Prozent). Auch dies ist ein sehr hoher Wert. Aufgrund dieser beiden Maßzahlen kann von einer sehr guten Modellanpassung gesprochen werden. Die 8-Clusterlösung ist geeignet, die empirischen Daten zu reproduzieren. Eine gute Modellanpassung ist eine notwendige, aber keinesfalls eine hinreichende Voraussetzung für das Vorliegen einer Clusterlösung. Dazu ist unter anderem die Frage der inhaltlichen Interpretierbarkeit zu klären.

Beschreibung und inhaltliche Interpretation: Aufgabe der clusteranalytischen Interpretation ist, für die einzelnen Cluster Namen zu finden. Als Ausgangspunkt dafür wird man die Clusterzentren und -streuungen betrachten, wie sie in Tabelle 6.4 dokumentiert sind.

⁹ Wie man die Signifikanz der Zuwächse prüfen kann, wird in Abschnitt 9.1.3 dargestellt. In den Abschnitten 8.6 und 9.2 wird gezeigt, wie man a priori das Vorhandensein einer Clusterstruktur prüfen kann. In Abschnitt 9.1.4 schließlich wird ein Testverfahren behandelt, mit dem man überprüfen kann, ob ein empirisches Verschmelzungsschema signifikant von dem Verschmelzungsschema bei Zufallsdaten abweicht.

Tab. 6.4: Mittelwerte (\bar{x}) und Standardabweichungen (s) der 8-Clusterlösung in den untersuchten Variablen (Faktorwerte)

Cluster	n	ExGrundbed		wirtWachst		Indust		Bildung	
		\bar{x}	s	\bar{x}	s	\bar{x}	s	\bar{x}	s
C_1	4	1,1	0,3	-0,8	0,5	0,7	0,1	0,8	0,2
C_2	3	-1,5	0,6	0,2	0,4	-0,3	0,4	0,1	0,5
C_3	6	0,6	0,3	0,4	0,2	-0,2	0,3	0,6	0,4
C_4	4	-0,6	0,2	0,5	0,4	-0,9	0,1	-1,2	0,8
C_5	2	-0,2	0,5	1,4	0,4	0,8	0,0	0,3	0,2
C_6	1	-2,0	0,0	-2,4	0,0	-1,3	0,0	-3,0	0,0
C_7	1	0,8	0,0	-2,4	0,0	-0,1	0,0	-0,0	0,0
C_8	1	-1,0	0,0	-0,2	0,0	2,9	0,0	0,1	0,0

Abkürzungen: siehe Text

Aufgrund der Clusterzentren können als Ausgangspunkt für weitere Analysen folgende Namen vergeben werden:

Cluster 1: Typus der relativ gut entwickelten Länder, aber mit unterdurchschnittlichem Wirtschaftswachstum. Diesem Typus gehören Argentinien, Chile, Kuba und Uruguay an.

Cluster 2: Typus der mittel entwickelten Länder, aber mit unterdurchschnittlicher Befriedigung existentieller Bedürfnisse. Diesem Cluster gehören Bolivien, Nicaragua und Peru an.

Cluster 3: Typus der mittel entwickelten Länder mit Brasilien, Costa Rica, Kolumbien, Mexiko, Panama und Paraguay.

Cluster 4: Typus der mittel entwickelten Länder, aber mit unterdurchschnittlichem Bildungsniveau. Dieses Cluster wird von der Dominikanischen Republik, von El Salvador, Guatemala und Honduras gebildet.

Cluster 5: Typus der mittel entwickelten Länder mit starkem Wirtschaftswachstum. Dies sind Ecuador und Venezuela.

Cluster 6: Typus des unterentwickelten Landes. Dies ist Haiti.

Cluster 7: Typus des mittel entwickelten Landes mit sehr unterdurchschnittlichem Wirtschaftswachstum. Dieser Typus wird von Jamaika gebildet.

Cluster 8: Typus des mittel entwickelten Landes mit sehr hoher Industrialisierung. Dieses Cluster wird von Trinidad gebildet.

Die hier durchgeführte Namensgebung ist als erster Interpretationsversuch zu verstehen. Er wirft eine Reihe von Fragestellungen auf, die zu beantworten sind, bevor die Interpretation abgeschlossen ist. So zum Beispiel kann gefragt werden, ob im Prinzip ein einheitlicher Entwicklungsprozeß vorliegt und sich die Abweichungen in den einzelnen

Tab. 6.5: Bevölkerungswachstum in den gebildeten Clustern zu unterschiedlichen Zeitpunkten (Clustermittelwerte)

Cluster	BevZu70	BevZu80	BevZu00
C_1	1,7	1,2	1,2
C_2	2,6	2,8	2,5
C_3	3,0	2,6	2,1
C_4	1,5	1,8	2,1
C_5	3,2	3,2	2,5
C_6	1,5	1,7	2,0
C_7	1,4	1,5	2,0
C_8	2,0	1,3	1,5

Abkürzungen: siehe Tabelle 5.12 auf Seite 137
 grau hinterlegt: zur inhaltlichen Validitätsprüfung ausgewähltes Cluster

Dimensionen auf spezifische politische, historische und/oder soziale Konstellationen zurückführen lassen.

Inhaltliche Validitätsprüfung: Die inhaltliche Validitätsprüfung hängt von der inhaltlichen Interpretation ab bzw. genauer davon, welche Hypothesen bei der Interpretation formuliert werden. Eine Anschlußhypothese für eine Validitätsprüfung könnte wie folgt aussehen: In dem Cluster 2 mit einer mittleren Entwicklung, aber einer unterdurchschnittlichen Befriedigung existentieller Bedürfnisse (ausreichende Nahrung, Gesundheit), ist die schlechte Versorgungslage mit Lebensmitteln auf ein hohes Bevölkerungswachstum zurückzuführen. Entsprechend dieser Hypothese wird man untersuchen, ob das Cluster 2 ein höheres Bevölkerungswachstum aufweist als die anderen Cluster. Es ergibt sich das Bild der Tabelle 6.5: Das Cluster C_2 weist gegenüber den anderen Clustern – mit Ausnahme des Clusters C_5 und des Clusters C_3 – in den 1980er Jahren ein höheres Bevölkerungswachstum auf. Auch für das Jahr 2000 wird ein relativ hohes Bevölkerungswachstum prognostiziert. Die in die Validitätsprüfung einbezogene Hypothese kann somit partiell bestätigt werden. Analog können weitere Hypothesen formuliert und geprüft werden.

Stabilitätsprüfung: Von der hier dargestellten inhaltlichen, kriterienbezogenen Validitätsprüfung ist die Untersuchung der Stabilität einer gefundenen Clusterlösung zu unterscheiden. Eine stabile Clusterlösung ist eine notwendige Voraussetzung für deren inhaltliche Gültigkeit. Bei der Stabilitätsanalyse werden die im Rahmen der Lösung einer Klassifikation getroffenen »unsicheren« Entscheidungen untersucht. Mit »unsicher« sind Entscheidungen gemeint, die nicht eindeutig inhaltlich und/oder empirisch begründet

werden konnten.¹⁰ So zum Beispiel wurde in der Analyse das Ward-Verfahren verwendet. Clusterzentren können aber auch mit den Median- oder Zentroid-Verfahren ermittelt werden. In einer Stabilitätsprüfung wird man daher untersuchen, ob sich die Ergebnisse ändern, wenn anstelle des Ward-Verfahrens das Median- und Zentroid-Verfahren verwendet werden. Ist dies der Fall, wird man die Ergebnisse hinsichtlich geeigneter Clusteranalyseverfahren als nicht stabil bezeichnen. Führt man für unser Beispiel die Stabilitätsprüfung durch, ergibt sich das in der Tabelle 6.6 auf Seite 167 dargestellte Bild: Die drei Verfahren führen zu teilweise unterschiedlichen Ergebnissen. Dennoch erkennen wir bestimmte stabile Clusterbildungen: Bei allen drei Verfahren werden Haiti, Jamaika und Trinidad als selbständige Cluster ausgewiesen. Beim Median- und Zentroid-Verfahren werden die Cluster C_1 und C_2 des Ward-Verfahrens zu einem einzigen Cluster zusammengefaßt. Eine Maßzahl zur Messung der Ähnlichkeit der einzelnen Lösungen ist der so genannte RAND-Index (siehe Abschnitt 9.5). Er gibt den Prozentsatz der übereinstimmenden Zuordnungen an. Die geringste Übereinstimmung liegt zwischen dem Zentroid- und dem Ward-Verfahren vor. Sie beträgt 81,8 Prozent. Das Median- und Ward-Verfahren stimmen relativ gut überein. Allgemein können Werte größer 0,7 bzw. 70 Prozent als ausreichende Übereinstimmung interpretiert werden (Dreger 1986; Fraboni und Saltstone 1992). Die 8-Clusterlösung kann somit als weitgehend stabil betrachtet werden. Analog wird man die anderen möglichen Clusterlösungen untersuchen und dann nach der in Kapitel 20 beschriebenen Methode der formalen Gültigkeitsprüfung vergleichend gegenüberstellen.

Abbildung 6.2 auf der nächsten Seite fasst die einzelnen Analyseschritte zur Lösung einer Klassifikationsaufgabe nochmals zusammen, wobei die in der Forschungspraxis ineinander übergehenden Schritte der Interpretation, der Stabilitäts- und inhaltlichen Validitätsprüfung entsprechend dem Anwendungsbeispiel angeordnet sind.

6.7 Fehlerquellen

Fehlerquellen, die bei einer Clusteranalyse auftreten und zu invaliden Ergebnissen führen können, sind:

1. Die Variablen sind so stark fehlerbehaftet, dass die zugrunde liegende Clusterstruktur nicht entdeckt werden kann.
2. Es wurden Variablen aufgenommen, die nicht zur Trennung der Cluster beitragen. In zahlreichen Simulationsstudien wurde nachgewiesen, dass irrelevante Variablen

¹⁰ Untersucht werden in einer Stabilitätsprüfung auch geringfügige Änderungen in den Daten (siehe dazu Abschnitt 12.7).

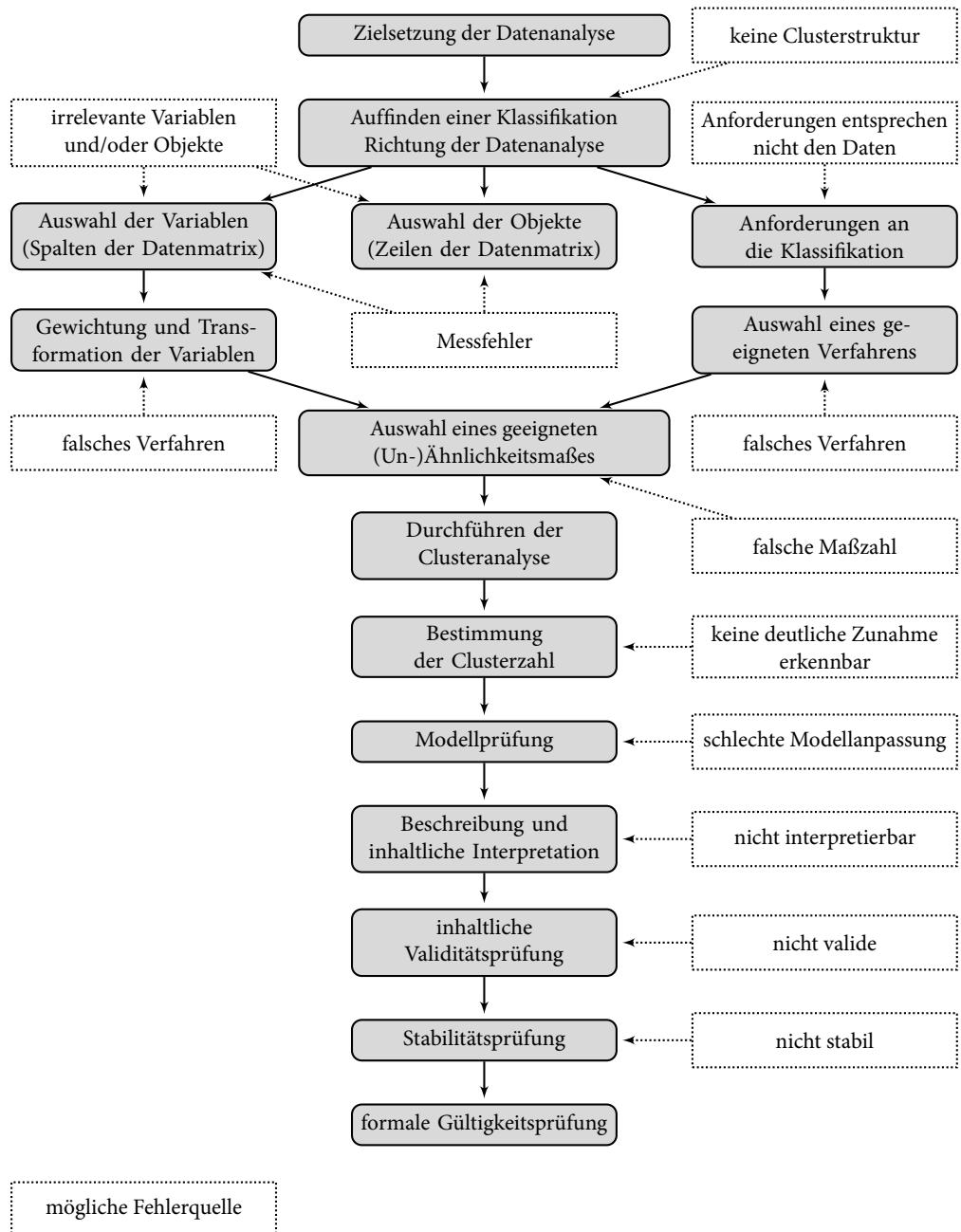


Abb. 6.2: Vorgehensweise der Lösung einer objektorientierten Klassifikationsaufgabe beim Vorliegen einer Datenmatrix

Tab. 6.6: Die 8-Clusterlösungen für das Ward-, Median- und Zentroid-Verfahren

Cluster	Ward-Verfahren	Median-Verfahren	Zentroid-Verfahren
<i>Zuordnung der Objekte</i>			
C_1	Argentinien, Chile, Kuba, Uruguay	Argentinien, Brasilien, Chile, Costa Rica, Kolumbien, Kuba, Mexiko, Panama, Paraguay, Uruguay	Argentinien, Brasilien, Chile, Costa Rica, Kolumbien, Kuba, Mexiko, Panama, Paraguay, Uruguay, Venezuela
C_2	Bolivien, Nicaragua, Peru	Bolivien, Dom. Republik, El Salvador, Honduras, Nicaragua, Peru	Bolivien
C_3	Brasilien, Costa Rica, Kolumbien, Mexiko, Panama, Paraguay	Ecuador	Dom. Republik, El Salvador, Guatemala, Honduras, Nicaragua, Peru
C_4	Dom. Republik, El Salvador, Guatemala, Honduras	Guatemala	Ecuador
C_5	Ecuador, Venezuela	Haiti	Haiti
C_6	Haiti	Jamaika	Jamaika
C_7	Jamaika	Trinidad	Peru
C_8	Trinidad	Venezuela	Trinidad
Verfahren	Ward-Verfahren	Median-Verfahren	Zentroid-Verfahren
<i>Übereinstimmungen zwischen den Verfahren (RAND-Index-Werte)</i>			
Ward	1,000		
Median	0,840	1,000	
Zentroid	0,818	0,900	1,000

einen entscheidenden Einfluss auf die Ergebnisse einer (objektorientierten) Clusteranalyse haben. Als irrelevant werden Variablen bezeichnet, die keinen Beitrag zur Trennung der Cluster leisten, in denen die untersuchten Objekte also eine homogene Verteilung, zum Beispiel eine Normal- oder Gleichverteilung, besitzen.

3. Es wurden »fehlerbehaftete« Objekte aufgenommen, zum Beispiel Ausreißer oder nichtklassifizierbare Objekte.
4. Es wurde ein Verfahren ausgewählt, das nicht den spezifizierten Anforderungen an die gesuchte Klassifikation entspricht.
5. Das ausgewählte Verfahren entspricht zwar den spezifizierten Forderungen, ist aber für die Daten ungeeignet. Bei der Spezifikation wurde zum Beispiel angenommen, dass keine Überlappungen auftreten, empirisch ist dies aber der Fall.
6. Den Daten liegt keine Clusterstruktur (im Sinne einer Klassifikation) zugrunde. Diese Schlussfolgerung wird man erst dann ziehen, wenn die anderen Fehlerursachen weitgehend ausgeschlossen werden können.

Einige der genannten Fehlerquellen wurden in Simulationsstudien untersucht (Everitt 1980, S. 99–106; Kaufman 1985; Milligan 1980, 1981; Mojena 1977; Punj und Stewart 1983

u. a.). Milligan (1980) untersuchte beispielsweise unter anderem folgende Fehlerquellen:

1. *Ausreißer* (fehlerbehaftete Objekte): Der vorgegebenen Clusterstruktur wurden 20 Prozent bzw. 40 Prozent Ausreißer hinzugefügt.
2. *Irrelevante Variablen*: Zu den durchschnittlich vier Variablen wurde eine bzw. zwei Zufallsvariablen als irrelevante Variable(n) hinzugefügt.
3. *Auswahl eines »falschen« Distanzmaßes*: Anstelle eines euklidischen Distanzmaßes wurde das Ähnlichkeitsmaß von Cattell (1949; siehe Formel 8.17 auf Seite 223) und die Q-Korrelation verwendet (siehe Formel 8.11 auf Seite 219).

Die Ergebnisse von Milligan wurden hinsichtlich dieser drei Fehlertypen für die in dieser Arbeit behandelten Verfahren reanalyisiert (siehe Tabelle 6.7): Insgesamt treten im Durchschnitt 9,69 Prozent Fehlklassifikationen auf. Dies ist eine sehr geringe Quote, wenn man bedenkt, dass in den untersuchten Modellkonstellationen im Durchschnitt 30 Prozent Ausreißer (Mittelwert aus 20 und 40 Prozent Ausreißern) vorliegen oder jede vierte Variable (1,5 von durchschnittlich 5,5 Variablen) irrelevant ist. Die beiden untersuchten Faktoren (Fehlertyp und Verfahren) erklären zusammen 52,6 Prozent der Gesamtvarianz. Der Rest ist auf Interaktionen der beiden Faktoren zurückzuführen. Die Verfahren verhalten sich somit in Abhängigkeit vom Fehlertypus unterschiedlich. Durchschnittlich betrachtet, kommt den irrelevanten Variablen die größte Bedeutung zu. Zwei irrelevante Variablen erhöhen den Anteil an falschen Zuordnungen von 9,69 auf 18,00 Prozent ($9,69 + 8,31$ Prozent), eine irrelevante nur auf 10,54 Prozent. Die Verfahren selbst unterscheiden sich nur geringfügig. Nur das K-Means-Verfahren reduziert signifikant den Anteil der Fehlklassifikationen, wenn es mit Startwerten aus dem Weighted-Average-Linkage gerechnet wird.

Zusammenfassend ist somit die Auswahl von »richtigen« Variablen bzw. die Elimination von »irrelevanten« Variablen von zentraler Bedeutung für die hier behandelten Verfahren (Green, Carmone u. a. 1990; Milligan 1980; Milligan und Cooper 1987). Die besten Ergebnisse werden hier mit dem Weighted-Average-Linkage und dem K-Means-Verfahren mit Startwerten aus dem Weighted-Average-Linkage erzielt (Milligan 1980). Die anderen Fehlertypen sind von untergeordneter Bedeutung. Nur das Ward-Verfahren und der Complete-Linkage reagieren sensibel auf Ausreißer.

Punj und Stewart (1983) kommen aufgrund einer Sichtung von 12 Simulationsstudien zu dem Ergebnis, dass K-Means-Verfahren und andere partitionierende Verfahren sowie das Ward-Verfahren und Mittelwertverfahren (Average-Linkage, Weighted-Average-Verfahren) gegenüber Fehlerquellen insgesamt weniger sensibel sind als andere Verfahren, wie zum Beispiel der Complete- oder Single-Linkage. Die eher schlechten Ergebnisse für das Ward-Verfahren von Milligan konnten hier nicht bestätigt werden.

Tab. 6.7: Verhalten von ausgewählten Verfahren bei unterschiedlichen Fehlerbedingungen (Anteil der Fehlklassifikationen, Ergebnisse aus einer Reanalyse der Daten von Milligan 1980)

Gesamtdurchschnitt	9,69*	Faktor: Verfahren	Haupteffekte
Faktor: FehlerTyp	Haupteffekte		
20 % Ausreißer	-2,32	Single-Linkage	-1,72
40 % Ausreißer	1,12	Complete-Linkage	1,77
eine irrelevante Variable	2,23	Weighted-Average	-4,10
zwei irrelevante Variablen	8,31*	Average	-2,75
Ähnlichkeitsmaß von Cattell	-4,05*	Within-Average	1,73
Q-Korrelation	-5,29*	Zentroid	3,57
erklärte Streuung in %	34,8	Median	3,65
		Ward	3,87
		K-Means mit zufälligen Startwerten	-0,36
		K-Means mit Startwerten aus dem Weighted-Average-Linkage	-5,65*
		Nächste-Nachbarn-Verfahren, Complete-Linkage für überlappende Cluster, Repräsentanten-Verfahren	nicht untersucht
erklärte Streuung in %	17,8		

*: signifikant zu einem Niveau von 95 Prozent ($p < 0,05$)

Quelle: Milligan (1980, S. 332,336), eigene Berechnung (Varianzanalyse)

Zu beachten ist, dass Simulationsstudien nur bedingt verallgemeinerbar sind, da sie von den getroffenen Modellannahmen, aber auch der Ergebnisauswertung abhängen. Hinzu kommt, dass der Großteil der Simulationsstudien in den 1970er und Anfang der 1980er Jahre durchgeführt wurde. Neuere Entwicklungen und Verfahren wurden daher nicht untersucht. So zum Beispiel äußern sich Kaufman und Rousseeuw (1990, S. 117–118) kritisch gegenüber Methoden, die Clusterzentren als Repräsentanten verwenden, da Mittelwerte allgemein gegenüber »Ausreißern« sensibel sind, ohne dies allerdings durch Simulationsstudien nachzuweisen. Trotz dieser Einschränkungen haben die zitierten Simulationsstudien entscheidend zur Identifikation möglicher Fehlerquellen und zu einer ersten groben Abschätzung ihrer Bedeutung beigetragen.

Um eine genauere Vorstellung der drei von Milligan untersuchten Fehlerquellen zu erhalten, sollen mit den Entwicklungsländerdaten folgende Rechenexperimente durchgeführt werden:

Rechenexperiment I: Wir fügen zwei fiktive »Ausreißer« mit den Werten $(-2,5, -2,5, -2,5, -2,5)$ und $(+2,5, +2,5, +2,5, +2,5)$ hinzu und untersuchen, in welcher Hinsicht sich die Clusterstruktur ändert.

Rechenexperiment II: Wir rechnen das Ward-Verfahren mit der City-Block-Metrik. Dies ist falsch, da das Ward-Verfahren die Verwendung der quadrierten euklidischen Distanzen voraussetzt.

Rechenexperiment III: Wir nehmen zwei standardnormalverteilte Zufallsvariablen als irrelevante Variablen in die Analyse auf.

Rechenexperiment I: Kein Einfluss von »Ausreißern« würde dann vorliegen, wenn diese eigenständige Cluster bilden bzw. mit einem Cluster verschmolzen werden, das bereits aus einem in den Daten vorhandenen Ausreißer besteht. In unserem Beispiel ist Haiti ein »Ausreißer«, der durch eine deutliche Unterentwicklung in allen vier Dimensionen gekennzeichnet ist. Das Muster von Haiti entspricht in etwa dem des Ausreißers 1. Wenn die Hinzunahme von Ausreißern keinen Einfluss hat, würden wir folgendes Ergebnis erwarten: Die Clusteranalyse sollte jeweils eine um 1 größere Clusterzahl als »bedeutsam« ausweisen als bei der Analyse ohne Ausreißer. Anstelle einer deutlichen Zunahme beim Übergang von elf zu zehn Clustern sollte also beim Übergang von zwölf zu elf Clustern eine deutliche Zunahme auftreten usw. Die Cluster, mit Ausnahme des von Haiti gebildeten Clusters, sollten sich dabei nicht ändern. Die 12-Clusterlösung sollte der 11-Clusterlösung entsprechen, wobei der Ausreißer 1 mit Haiti ein gemeinsames Cluster und der Ausreißer 2 ein selbständiges Cluster bildet usw. Führt man die Analyse durch (siehe Tabelle 6.8), zeigt sich tatsächlich beim Übergang von zwölf auf elf Cluster eine deutliche Zunahme des Distanzniveaus. Auch die anderen Zunahmen im Verschmelzungsniveau verschieben sich um 1 nach oben. Anstelle einer Zunahme beim Übergang von acht zu sieben Clustern liegt eine Zunahme beim Übergang von neun zu acht Clustern vor usw. Betrachten wir die 12-Clusterlösung im Detail, so entspricht diese den Erwartungen. Der zweite Ausreißer bildet ein selbständiges Cluster, der erste Ausreißer ein gemeinsames Cluster mit Haiti. Die anderen Cluster werden nicht geändert. Dies ist auch bei den anderen Lösungen der Fall. Ausreißer haben somit in dem Beispiel keinen Einfluss.

Rechenexperiment II: Die Wahl eines »falschen« Unähnlichkeitsmaßes hat ebenfalls keinen Einfluss auf die Klassifikationsergebnisse. Die Ergebnisse für die City-Block-Metrik und die quadrierte euklidische Distanz stimmen perfekt überein. Auch dieser Befund bestätigt die Simulationsstudien.

Rechenexperiment III: Die Hinzunahme von zwei irrelevanten Variablen führt – im Unterschied zu der Hinzunahme von Ausreißern – dagegen zu einer weitgehenden Zerstörung der Clusterstruktur (siehe Tabelle 6.9 auf Seite 172). Zwar tritt beim Übergang von elf auf zehn Cluster ein deutlicher Zuwachs im Distanzniveau auf, die beiden elf Clusterlösungen (mit und ohne irrelevanten Variablen) unterscheiden sich aber deutlich. In einem weiteren

Tab. 6.8: Ergebnisse des Ward-Verfahrens für die Entwicklungsländerdaten bei Hinzunahme von zwei Ausreißern

Schritt	Verschmelzung der Cluster mit den Objekten	Clusterzahl	Verschmelzungsniveau	Zunahme	Mojena-I-Teststatistik	Mojena-II-Teststatistik
1	3 – 16	23	0,096	—	—	—
2	5 – 15	22	0,440	0,344	0,933	-1,188
3	7 – 19	21	0,495	0,055	1,945	0,099
4	8 – 13	20	0,764	0,269	1,469	-0,406
5	6 – 23	19	0,852	0,088	1,615	-0,231
6	5 – 20	18	1,011	0,159	1,231	-0,603
7	7 – 17	17	1,018	0,007	2,116	0,322
8	18 – 21	16	1,387	0,369	1,797	0,000
9	8 – 10	15	1,484	0,098	2,064	0,267
10	3 – 6	14	1,766	0,281	2,025	0,231
11	1 – 12	13	1,974	0,209	1,632	-0,158
12	5 – 7	12	1,975	0,000	3,520	1,734
13	9 – 24	11	3,281	1,306	3,186	1,486
14	8 – 11	10	3,965	0,684	2,325	0,658
15	4 – 18	9	3,987	0,022	4,014	2,332
16	3 – 14	8	6,577	2,590	3,572	1,975
17	5 – 9	7	8,070	1,493	3,964	2,394
18	4 – 8	6	11,137	3,067	2,810	1,283
19	3 – 22	5	11,229	0,092	6,393	4,842
20	3 – 5	4	25,621	14,392	4,838	3,476
21	2 – 3	3	33,648	8,027	4,904	3,565
22	1 – 4	2	48,414	14,766	8,143	6,825
23	1 – 2	1	109,015	60,601	—	—

grau hinterlegt: Zunahme des Verschmelzungsniveaus

Schritt wurde daher überprüft, ob die Durchführung einer Faktorenanalyse zu einer Elimination der irrelevanten Variablen führt. Zusätzlich zu den ursprünglich dreizehn untersuchten Variablen wurden zwei irrelevanten, standardnormalverteilte Zufallsvariablen in die Faktorenanalyse einbezogen. Führt man die Faktorenanalyse durch, ergeben sich zunächst anstelle der vier bedeutsamen Faktoren fünf bedeutsame Dimensionen (siehe Tabelle 6.10 auf Seite 173). Auch die beiden Faktorstrukturen unterscheiden sich deutlich. So zum Beispiel werden die beiden Dimensionen »Bildung« und »Befriedigung existentieller Grundbedürfnisse« zu einem Faktor verschmolzen. Der Export von Rohstoffen (Abkürzung: Export) wird zu einem selbständigen Faktor. Die Ursachen für die Zerstörung der ursprünglichen Faktorstruktur sind in dem Beispiel die kleine Fallzahl sowie das Fehlen einer Einfachstruktur. Verfüngigfachen wir die Stichprobe, indem jedes Land 20-mal mit unterschiedlichen Realisierungen der beiden irrelevanten Variablen in die Analyse einbezogen wird, treten zwar noch immer fünf Faktoren mit Eigenwerten größer 1 auf, die ursprüngliche Faktorstruktur wird aber reproduziert (siehe

Tab. 6.9: Ergebnisse des Ward-Verfahrens für die Entwicklungsländerdaten bei Hinzunahme von zwei irrelevanten Variablen

11-Clusterlösung mit irrelevanten Variablen			11-Clusterlösung ohne irrelevante Variablen		
C_1 ($n = 5$)	1	Argentinien	C_1 ($n = 4$)	1	Argentinien
	5	Costa Rica		4	Chile
	13	Kolumbien		14	Kuba
	15	Mexiko		21	Uruguay
	18	Paraguay	C_2 ($n = 1$)	2	Bolivien
C_2 ($n = 1$)	2	Bolivien	C_3 ($n = 6$)	3	Brasilien
C_3 ($n = 3$)	3	Brasilien		5	Costa Rica
	6	Domenikan. Repub.		13	Kolumbien
	16	Nicaragua		15	Mexiko
C_4 ($n = 4$)	4	Chile		17	Panama
	14	Kuba		18	Paraguay
	17	Panama	C_4 ($n = 3$)	6	Domenikan. Repub.
	21	Uruguay		8	El Salvador
C_5 ($n = 1$)	7	Ecuador		11	Honduras
C_6 ($n = 2$)	8	El Salvador	C_5 ($n = 1$)	7	Ecuador
	11	Honduras	C_6 ($n = 1$)	9	Guatemala
C_7 ($n = 1$)	9	Guatemala	C_7 ($n = 1$)	10	Haiti
C_8 ($n = 1$)	10	Haiti	C_8 ($n = 1$)	12	Jamaika
C_9 ($n = 1$)	12	Jamaika	C_9 ($n = 1$)	16	Nicaragua
C_{10} ($n = 1$)	19	Peru		19	Peru
C_{11} ($n = 2$)	20	Trinidad	C_{10} ($n = 1$)	20	Trinidad
	22	Venezuela	C_{11} ($n = 1$)	22	Venezuela

Tabellen 6.11 und 6.1 auf Seite 158). Die beiden irrelevanten Variablen bilden den fünften Faktor, der mit einem Eigenwert von 1,0240 nur mehr knapp über dem Schwellenwert von 1,0 liegt und gegenüber dem Eigenwert des vierten Faktors (1,3250) einen deutlichen Abfall aufweist.

Allgemein lässt sich festhalten: Die Gefahr, dass durch irrelevante Zufallsvariablen eine vorhandene dimensionale Struktur zerstört wird, nimmt mit der Stichprobengröße ab und erhöht sich, wenn das zugrunde liegende Modell keine Einfachstruktur besitzt. Die Einfachstruktur muss allerdings sehr stark verletzt sein, damit durch die Hinzunahme von irrelevanten Variablen eine Zerstörung auftritt. Um uns von diesem Sachverhalt zu überzeugen, soll folgende zweifaktorielle Struktur untersucht werden: Die Variablen X_1 und X_2 laden mit einer Faktorladung von 0,8 auf dem ersten Faktor, die Variablen X_5 und X_6 mit einer Faktorladung von 0,8 auf dem zweiten Faktor. Die Variablen X_3 und X_4 verletzen dagegen die Annahme der Einfachstruktur, sie laden jeweils mit einer Faktorladung von 0,5 auf beiden Faktoren. Die beiden Faktoren sollen voneinander unabhängig und standardnormalverteilt sein. Die empirische Ausprägung in jeder Variablen soll neben dem oder den gemeinsamen Faktor bzw. Faktoren noch von zufälligen, ebenfalls

Tab. 6.10: Ergebnisse der Faktorenanalyse für die Entwicklungsländerdaten bei Hinzunahme von zwei irrelevanten Variablen

Variable	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
ImpNahr	-0,4887	-0,4416	-0,1303	-0,1469	-0,2623
Kalorien	0,7646	-0,3058	-0,2908	-0,2979	0,2990
LebErw	0,8652	-0,4019	-0,1659	-0,2461	0,6369
Kindsterb	-0,8008	0,2281	0,2224	0,1445	-0,5590
Alpha80	0,8986	-0,1855	0,2701	0,0683	0,6830
Einschulung	0,8285	-0,0143	0,0706	0,0983	0,3946
EinwproAr	-0,7954	0,0095	-0,3053	-0,2480	-0,5564
Wachst80	0,0478	0,7867	0,1280	0,3565	-0,1347
BevZu80	-0,3137	0,7926	-0,0786	0,5093	-0,4422
Schulden	0,2595	0,1633	0,7011	-0,0224	-0,1053
irrelev. Var. 1	-0,1196	-0,1944	0,7126	0,1775	0,3200
irrelev. Var. 2	0,0916	-0,0935	-0,7952	0,1901	0,0033
Export	-0,0300	0,4956	-0,0223	0,9036	0,2893
BruSozPr	0,6618	-0,2385	-0,0278	0,1847	0,8458
Indust80	0,4317	-0,3903	0,1839	0,3640	0,8793

grau hinterlegt: höchste Faktorladung jeder Ausprägung auf den Faktoren

Tab. 6.11: Einfluss der irrelevanten Variablen auf die Ergebnisse der Hauptkomponentenanalyse in einer größeren Stichprobe (Verfünzigfachung der Stichprobe, $n = 50 \cdot 22 = 1100$)

Variable	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
Kalorien	0,8244	-0,2166	0,2817	0,5313	-0,0222
LebErw	0,9561	-0,1994	0,5776	0,5680	-0,0136
Kindsterb	-0,9260	0,0189	-0,4607	-0,4772	0,0045
Wachst80	-0,0625	0,8025	-0,1580	0,1032	0,1807
Export	-0,2335	0,6397	0,2713	0,0845	-0,2179
ImpNahr	-0,3931	-0,5469	-0,2407	-0,4292	-0,0639
BevZu80	-0,4348	0,7080	-0,4311	-0,1745	-0,0297
BruSozPr	0,6162	0,0205	0,8436	0,4802	0,1503
Indust80	0,2700	-0,1345	0,9176	0,4418	0,0926
Schulden	-0,0519	0,1394	-0,0749	0,6389	0,0754
Alpha80	0,6350	0,1225	0,3993	0,8449	-0,1056
Einschulung	-0,4967	-0,1606	-0,5977	-0,9169	0,1073
EinwproAr	0,7200	0,0323	0,6739	0,8907	-0,0567
irrelev. Var. 1	-0,0198	-0,0325	0,0778	-0,0214	0,6624
irrelev. Var. 2	0,0003	0,0463	0,1037	-0,0699	0,6689

grau hinterlegt: höchste Faktorladung jeder Ausprägung auf den Faktoren

Tab. 6.12: Einfluss von irrelevanten Variablen auf die Ergebnisse der Hauptkomponentenanalyse (Ergebnisse eines Rechenexperiments)

	Analyse ohne irrelevante Variablen		Analyse mit irrelevanten Variablen		
	Faktor 1	Faktor 2	Faktor 1	Faktor 2	Faktor 3
X_1	0,8734	0,0391	0,8730	0,0389	0,1761
X_2	0,8518	-0,0260	0,8499	-0,0168	0,1592
X_3	0,8703	0,4500	0,8714	0,4512	0,0820
X_4	0,8580	0,5409	0,8584	0,5449	0,0619
X_5	0,2922	0,8916	0,2971	0,8831	-0,1925
X_6	0,2188	0,8678	0,2181	0,8638	0,1001
irrelevante Var. 1	—	—	0,1080	0,0924	0,7720
irrelevante Var. 2	—	—	0,0946	-0,1884	0,6895
Eigenwerte	3,2863	1,1639	3,2541	1,6431	1,0445

standardnormalverteilten Messfehlern abhängen und sich wie folgt zusammensetzen aus:

$$X_i = a_{i1} \cdot F_1 + a_{i2} \cdot F_2 + \sqrt{1 - a_{i1}^2 - a_{i2}^2} \cdot U_i ,$$

wobei U_i der zufällige Messfehler der Variablen i ist. Für X_1 und X_2 ist $a_{11} = a_{21} = 0,8$ und $a_{12} = a_{22} = 0$, für X_5 und X_6 ist $a_{51} = a_{61} = 0$ und $a_{52} = a_{62} = 0,8$ und für X_3 sowie X_4 ist $a_{31} = a_{32} = a_{41} = a_{42} = 0,5$. Fügt man zwei irrelevante Zufallsvariablen hinzu und erzeugt eine Datenmatrix mit 100 Objekten, werden drei Faktoren mit Eigenwerten größer 1 ermittelt. Der dritte Faktor setzt sich wiederum aus den beiden irrelevanten Variablen zusammen. Die ursprüngliche Faktorstruktur wird reproduziert (siehe Tabelle 6.12). Da bei der Berechnung der unrotierten Faktoren deren Varianz maximiert wird, laden die Variablen X_3 und X_4 stärker auf dem ersten Faktor.

In der Forschungspraxis kann somit in der Regel davon ausgegangen werden, dass irrelevanten Variablen eine dimensionale Struktur nicht zerstören, dass sie aber mitunter einen eigenständigen Faktor bilden. Dies ist ein äußerst positiver Befund, der aber nicht bedeutet, dass irrelevanten Zufallsvariablen aus der Clusteranalyse eliminiert sind. In dem obigen Beispiel sind alle Variablen für eine Klassifikation der Objekte »irrelevant«, da alle Variablen eine homogene Normalverteilung besitzen und somit überhaupt keine Clusterstruktur vorliegt. Die Ergebnisse einer Clusteranalyse würden somit ein reines Artefakt darstellen. Verfahren, mit denen sich überprüfen lässt, ob die Ergebnisse einer Clusteranalyse ein reines Artefakt darstellen, werden ausführlich in den nachfolgenden Abschnitten dargestellt.

7 Gewichtung und Transformation von Variablen

7.1 Vergleichbarkeit von Klassifikationsmerkmalen

Formal müssen die Variablen, die in eine deterministische Clusteranalyse einbezogen werden, »vergleichbar« sein. Vergleichbarkeit (*Kommensurabilität*) von Variablen liegt in folgenden Situationen nicht vor (*Inkommensurabilität*):¹

- Die Variablen besitzen *unterschiedliche Maßeinheiten*.
- Die Variablen besitzen *gemischtes Messniveau*.
- Die Variablen sind *hierarchisch*.

Unterschiedliche Maßeinheiten: In dem Anwendungsbeispiel des Abschnitts 6.6 wurden als Indikatoren für die wirtschaftliche Entwicklung das »Pro-Kopf-Bruttosozialprodukt« und das »jährliche Wirtschaftswachstum in den 1980er Jahren« verwendet. Diese beiden Variablen besitzen zwar das gleiche Messniveau (quantitativ), sind aber nicht vergleichbar, da das »Pro-Kopf-Bruttosozialprodukt« in einer bestimmten Währungseinheit (Dollar) gemessen wird, das »jährliche Wirtschaftswachstum in den 1980er Jahren« dagegen in Prozenten.

Gemischtes Messniveau: In eine Clusteranalyse sollen die nominalen Variablen »Geschlecht« und »berufliche Tätigkeit«, die ordinale Variable »abgeschlossene Schulbildung« und die quantitative Variable »Einkommen« einbezogen werden. In diesem Beispiel liegt Nichtvergleichbarkeit vor, da die Variablen unterschiedliches Messniveau besitzen.

Hierarchische oder bedingte Variablen: Diese liegen dann vor, wenn das Auftreten einer Variablen von dem Auftreten der Ausprägung(en) einer oder mehrerer anderer Variablen

¹ Siehe dazu zum Beispiel Schlosser (1976, S. 60–88), Sodeur (1974, S. 44–59) oder Vogel (1975, S. 50–78).

abhängt. Die Variable »derzeitiger Beruf« tritt beispielsweise nur dann auf, wenn die vorausgehende Variable »Berufstätigkeit« die Ausprägung »derzeit berufstätig« besitzt.

Daneben können inhaltliche Überlegungen zu dem Urteil der Nichtvergleichbarkeit führen: Selbst wenn alle in die Analyse einbezogenen Variablen dieselbe Maßeinheit (zum Beispiel Prozente) besitzen, wie zum Beispiel »Industrialisierungsquote in den 1980er Jahren«, »jährliches Wirtschaftswachstum in den 1980er Jahren«, kann in Frage gestellt werden, ob beide Klassifikationsmerkmale dieselbe »Maßeinheit« besitzen, ob also eine Differenz von 5 Prozent beim jährlichen Wirtschaftswachstum »dasselbe« bedeutet wie bei der Industrialisierungsquote (siehe dazu auch Fox 1982, S. 132). Weitere inhaltliche Überlegungen beziehen sich darauf, ob den Variablen gemeinsame Dimensionen zugrunde liegen, die durch eine unterschiedliche Anzahl von Indikatoren (Variablen) repräsentiert sind (Problem der Über- bzw. Unterrepräsentativität, siehe dazu Abschnitt 6.6).

7.2 Lösungsstrategien

Liegt Nichtvergleichbarkeit der Variablen vor, stehen unter anderem folgende Strategien zur Verfügung:

1. Die Variablen werden vor der Analyse transformiert bzw. gewichtet.
2. Die Variablen werden bei der Berechnung eines (Un-)Ähnlichkeitsmaßes gewichtet.
3. Die Variablen werden in der Clusteranalyse gewichtet.
4. Es werden getrennte Analysen für jeweils jene Variablengruppen gerechnet, die vergleichbar sind.²
5. Es werden durch eine vorausgehende Anaylse mittels eines Skalierungsverfahrens (zum Beispiel mittels mehrdimensionaler Skalierung, Korrespondenz- oder Faktorenanalyse) abgeleitete Variablen gebildet, die vergleichbar sind. Dies war die im Anwendungsbeispiel aus Abschnitt 6.6 gewählte Strategie.

Allgemein sollte nach Möglichkeit bei der Datenerhebung das Problem der Nichtvergleichbarkeit vermieden werden, indem bedeutungsgleiche, der jeweiligen Fragestellung angepasste Antwortskalen verwendet werden. Liegt dennoch Nichtvergleichbarkeit vor, kann eine der genannten Strategien eingesetzt werden. Strategie 1 bis 3 sind ineinander

² Beispiel: In einer Untersuchung wurden zwei Fragebatterien verwendet. Mit der ersten Fragebatterie wurden Erziehungsziele durch eine fünfstufige Antwortskala erfasst. In der zweiten Fragebatterie wurden gemeinsame familiäre Freizeitaktivitäten mit einer dichotomen Antwortskala erfasst. Um das Problem der Vergleichbarkeit zu umgehen, kann zunächst jede Fragebatterie getrennt untersucht werden.

überführbar und resultieren daher auch in identischen Ergebnissen. Wir wollen hier die erstgenannte Strategie darstellen.

7.3 Theoretische und empirische Standardisierung

Mit der Bezeichnung theoretische und empirische Standardisierung sind folgende Datentransformationen gemeint:

1. die »eigentliche« *Standardisierung* bzw. *z-Transformation*
2. die *Extremwertnormalisierung*

Die »eigentliche« *Standardisierung* bzw. *z-Transformation* kann mit Hilfe der folgenden Berechnungsvorschrift durchgeführt werden:

$$z_{gi} = \frac{x_{gi} - \bar{x}_i}{s_i} \quad \text{bzw.} \quad z_{gi} = \frac{x_{gi} - \mu_i}{\sigma_i}, \quad (7.1)$$

wobei z_{gi} der standardisierte Wert des Objekts g in der standardisierten Variablen Z_i ist. Der Wert des Objekts g in der nichtstandardisierten Variablen X_i wird mit x_{gi} und der empirische Mittelwert der Variablen X_i mit \bar{x}_i bezeichnet. Die empirische Standardabweichung ist s_i , die theoretische Standardabweichung ist σ_i und der theoretische Skalenmittelwert wird mit μ_i bezeichnet.

Die *Extremwertnormalisierung* erfolgt über:

$$z_{gi} = \frac{x_{gi} - a_i}{b_i - a_i} \quad \text{bzw.} \quad z_{gi} = \frac{x_{gi} - \alpha_i}{\beta_i - \alpha_i}, \quad (7.2)$$

wobei a_i die empirische Untergrenze der Variablen X_i ist. Die empirische Obergrenze bezeichnet b_i , α_i die theoretische Untergrenze und β_i die theoretische Obergrenze. Die Extremwertnormalisierung transformiert die Variable auf das Intervall $[0,1]$.

In einer Standardisierung können theoretische Skalenkennwerte oder empirische Verteilungskennwerte eingehen. Diese Unterscheidung ist *nur für eine objektorientierte Clusteranalyse* wichtig. Im Rahmen einer variablenorientierten Clusteranalyse werden üblicherweise Korrelationskoeffizienten zur Messung der Ähnlichkeit verwendet. Es wird somit implizit eine empirische Standardisierung durchgeführt. *Die nachfolgenden Ausführungen beziehen sich somit primär auf eine objektorientierte Clusteranalyse.*

Zur Verdeutlichung des Unterschieds von theoretischen und empirischen Kennwerten wollen wir folgendes Beispiel betrachten (siehe Abbildung 7.1 auf Seite 179). In einer Clusteranalyse soll unter anderem das Item »Eine berufstätige Mutter kann ein genauso

herzliches Verhältnis zu ihren Kindern finden wie eine Mutter, die nicht berufstätig ist« mit den Ausprägungen 1 (stimme voll zu), 2 (stimme zu), 3 (stimme eher nicht zu) und 4 (stimme überhaupt nicht zu) einbezogen werden. Die Befragten verteilen sich auf die Antwortkategorien wie folgt: 35 Prozent (Kategorie 1), 45 Prozent (Kategorie 2), 20 Prozent (Kategorie 3) und 0 Prozent (Kategorie 4). Hinsichtlich dieser Variablen können folgende Skalenkennwerte berechnet werden:

- Theoretische Unter- und Obergrenze: 1 und 4
- Empirische Unter- und Obergrenze: 1 und 3, da empirisch Ausprägung 4 nicht auftritt
- Theoretischer Skalenmittelwert: $(1 + 2 + 3 + 4)/4 = 2,5$
- Empirischer Mittelwert:

$$(1 \cdot 0,35 + 2 \cdot 0,45 + 3 \cdot 0,20) = 1,850$$
- Theoretische Skalenstandardabweichung:

$$\sqrt{((1 - 2,5)^2 + (2 - 2,5)^2 + (3 - 2,5)^2 + (4 - 2,5)^2)/4} = \sqrt{1,25} = 1,12$$
- Empirische Standardabweichung:

$$\sqrt{0,35 \cdot (1 - 1,85)^2 + 0,45 \cdot (2 - 1,85)^2 + 0,20 \cdot (3 - 1,85)^2} = 0,073$$

Die Abbildung 7.1 verdeutlicht den Unterschied zwischen theoretischen und empirischen Skalenkennwerten. Empirische Skalenwerte werden auf der Grundlage der Verteilung der untersuchten Objekte berechnet. In die Berechnung der theoretischen Skalenwerte geht dagegen die Verteilung der untersuchten Objekte (Personen) nicht ein. Sie sind daher populationsunabhängig. Die theoretischen Skalenkennwerte werden aus der Bedeutung der vorgegebenen Antwortskalen abgeleitet, wobei zur Berechnung des theoretischen Skalenmittelwerts und der theoretischen Skalenstandardabweichung eine Gleichverteilung der Objekte auf den Skalen (Variablen) angenommen wird.³ Die theoretischen Skalenkennwerte können in Abhängigkeit vom Skalentyp nach einer der in der Tabelle 7.1 auf Seite 180 wiedergegebenen Formeln berechnet werden.

Die drei Skalentypen der Tabelle 7.1 auf Seite 180 sind:

Skalentyp I: Variablen mit einer kontinuierlichen Skala in dem Wertebereich zwischen α_i und β_i , wie zum Beispiel Stimmenanteile von Parteien (Wertebereich von 0 bis 100 Prozent).

Skalentyp II: Variablen mit einer diskreten Skala mit ganzzahligen äquidistanten Ausprägungen: $X_{1i} = \alpha_i$, $X_{2i} = \alpha_i + 1, \dots$ wie zum Beispiel eine Einstellungsfrage mit den Ausprägungen 1 (stimme voll zu), 2 (stimme zu), 3 (stimme eher nicht zu), 4 (stimme überhaupt nicht zu).

Skalentyp III: Variablen mit einer diskreten Skala mit nicht äquidistanten Ausprägungen, wie zum Beispiel eine Aktivitätsfrage (zum Beispiel nach der Häufigkeit einer

³ Auch die Annahme einer Normalverteilung ist möglich (Schlosser 1976: 64). Dadurch entsteht aber ein größerer Berechnungsaufwand.

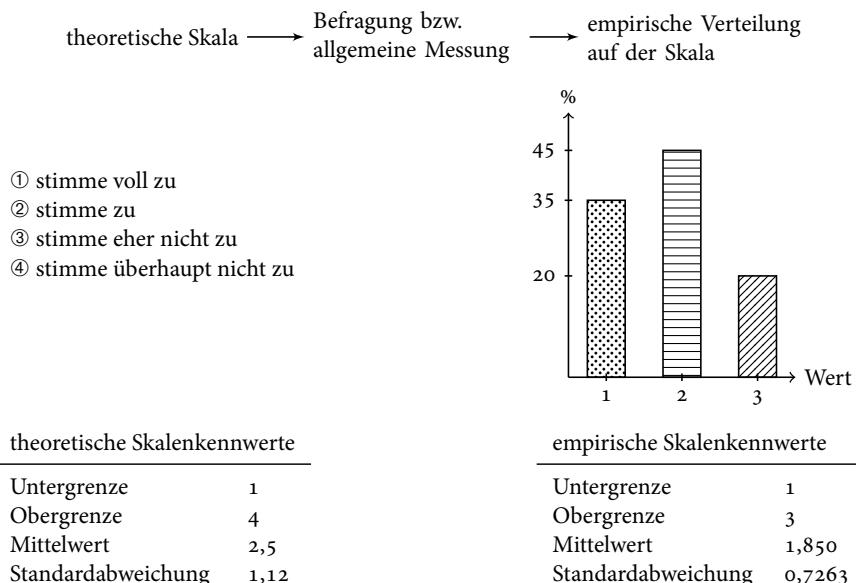


Abb. 7.1: Theoretische und empirische Skalenwerte

bestimmten Freizeitaktivität oder nach dem Kontakt mit einer bestimmten Person) mit den Ausprägungen täglich (7,0), mehrmals wöchentlich (3,5), wöchentlich (1,0) und seltener (0,20).

Da der Skalentyp II in der sozialwissenschaftlichen Praxis der Regelfall ist, enthält Tabelle 7.2 auf der nächsten Seite die theoretischen Skalenkennwerte für eine unterschiedliche Zahl von Ausprägungen. Dabei wurde von einer Kodierung ausgegangen, die mit 1 beginnt. Wird beispielsweise eine fünfstufige Skala verwendet, kann durch die Transformation $z_{gi} = (x_{gi} - 3,0) / 1,41$ eine theoretische z-Transformation durchgeführt werden, indem die entsprechenden Tabellenwerte verwendet werden. Die fünf Ausprägungen erhalten dadurch folgende standardisierte Skalenwerte: 1 : -1,42, 2 : -0,71, 3 : 0, 4 : 0,71, 5 : 1,42.

Ziel der *theoretischen Standardisierung* ist, vergleichbare Variablen zu erhalten, wobei die standardisierten Skalenwerte die ursprüngliche Bedeutung der Ausprägungen abbilden sollen. Mitunter kann dies auch durch eine *einfache Umkodierung* erreicht werden. So sind beispielsweise für die Antwortskalen in Tabelle 7.3 auf Seite 181 folgende Umkodierungen möglich, die zu einer Vergleichbarkeit führen (siehe Spalte »inhaltliche Umkodierung« in Tabelle 7.3 auf Seite 181): Die Kodierung der Variablen 1 wird beibehalten, die Ausprägung »stimme zu« der Variablen 2 erhält den Wert 2,5. Dies ist der Durchschnitt der ersten vier Ausprägungen der Variablen 1. Die Ausprägung »lehne ab« erhält den Wert 5,5. Dies ist der Durchschnitt der letzten vier Kategorien. Die Kategorie

Tab. 7.1: Berechnungsformeln für theoretische Skalenkennwerte

Symbol	Skalentypen			
	I	II	III	
Untergrenze	α_i		aus der Skala unmittelbar ablesbar	
Obergrenze	β_i		aus der Skala unmittelbar ablesbar	
Skalenmittelwert	μ_i	$\frac{\beta_i - \alpha_i}{2}$	$\frac{\beta_i - \alpha_i}{2}$	$\frac{\sum X_{ij}}{m_i}$
Skalenvarianz	σ_i^2	$\frac{(\beta_i - \alpha_i)^2}{12}$	$\frac{(m_i + 1)(m_i - 1)}{12}$	$\frac{\sum(X_{ij} - \mu_i)^2}{m_i}$
Skalenstandardabweichung	σ_i	$\sqrt{\sigma_i^2}$	$\sqrt{\sigma_i^2}$	$\sqrt{\sigma_i^2}$

Abkürzungen: m_i : Zahl der Ausprägungen der untersuchten Variablen; X_{ij} : Wert der j -ten Ausprägung der Variablen X_i

dazwischen wird doppelt verrechnet. Analog wird für die dritte Variable vorgegangen. Eine *theoretische Extremwertnormalisierung* würde hier zu dem unerwünschten Effekt führen, dass die Antwortkategorie »stimme zu« den gleichen Skalenwert erhält wie die Ausprägung »stimme stark zu«. Mit der Folge, dass in einer Clusteranalyse der Unterschied von »stimme zu« und »lehne ab« in der Variablen 2 gleich bewertet würde wie der Unterschied von »stimme stark zu« und »lehne stark ab« in der Variablen 3. Die *theoretische z-Transformation* führt diesbezüglich zu besseren Ergebnissen (siehe Tabelle 7.3). Die Kategorie »stimme stark zu« der Variablen 1 erhält mit 1,50 einen Skalenwert, der im Absolutbetrag größer ist als jener der Variablen 2 von -1,00. Die Ausprägung »stimme zu« der Variablen 2 erhält mit -1,34 ebenfalls im Absolutbetrag einen größeren

Tab. 7.2: Theoretische Skalenkennwerte für den Skalentyp II in Abhängigkeit von der Zahl der Ausprägungen

Zahl der Ausprägungen	theoretische Skalenkennwerte				
	Unter-grenze α_i	Ober-grenze β_i	Skalen-mittelwert μ_i	Skalenstandard-abweichung σ_i	
2	1	2	1,5	0,50	
3	1	3	2,0	0,82	
4	1	4	2,5	1,12	
5	1	5	3,0	1,41	
6	1	6	3,5	1,71	
7	1	7	4,0	2,00	
8	1	8	4,5	2,29	
9	1	9	5,0	2,59	
10	1	10	5,5	2,87	

Tab. 7.3: Konsequenzen einer inhaltlichen Umkodierung, theoretischen Extremwertnormalisierung und z-Transformation

	inhaltliche Umkodierung	theoretische Extremwert- normalisierung	theoretische z-Transformation
<i>Variable 1</i>			
1 »stimme stark zu«	1	0,00	-1,50
2 »stimme zu«	2	0,17	-1,00
3 »stimme eher zu«	3	0,33	-0,50
4 »dazwischen«	4	0,50	0,00
5 »lehne eher ab«	5	0,67	0,50
6 »lehne ab«	6	0,83	1,00
7 »lehne stark ab«	7	1,00	1,50
<i>Variable 2</i>			
1 »stimme zu«	2,5	0,00	-1,00
2 »lehne ab«	5,5	1,00	1,00
<i>Variable 3</i>			
1 »stimme zu«	1,5	0,00	-1,34
2 »stimme eher zu«	3,5	0,33	-0,45
3 »stimme eher nicht zu«	4,5	0,67	0,45
4 »stimme überhaupt nicht zu«	6,5	1,00	1,34

Skalenwert als die Variable 2, da es mit »stimme eher zu« noch eine weitere Kategorie gibt, mit der eine Zustimmung ausgedrückt werden kann. Der Skalenwert ist kleiner als jener von »stimme sehr zu«, da es hier drei Ausprägungen für Zustimmungen gab. Das Beispiel macht deutlich, dass *eine bestimmte theoretische Transformation* – in dem Beispiel die theoretische Extremwertnormalisierung – *nicht automatisch durchgeführt* werden soll. Vielmehr sind in jedem konkreten Anwendungsbeispiel die Konsequenzen einer Skalierung zu prüfen.

Voraussetzung für eine theoretische Transformation ist, dass die Skalenkennwerte der untersuchten Variablen definiert sind und die Variablen quantitatives Messniveau besitzen oder als »quantitativ« betrachtet werden können. Nicht alle Variablen erfüllen diese Voraussetzungen, wie man sich leicht anhand von Beispielen verdeutlichen kann:

- Die Variable »Wirtschaftswachstum« ist zwar eine kontinuierliche Skala (Skalentyp I), Unter- und Obergrenze sind aber nicht definiert.
- Das Einkommen – durch eine offene Frage erfragt – ist eine kontinuierliche Skala (Skalentyp I), die Obergrenze ist nicht definiert.
- Die Zahl der Kinder in einem Haushalt ist eine diskrete, ganzzahlige Variable (Skalentyp II). Die Obergrenze dieser Variablen ist nicht exakt definiert.

- Die Variable »berufliche Tätigkeit« erfüllt nicht diese Voraussetzung, da sie nominal-skaliert ist.

Sind die Voraussetzungen nicht erfüllt, ist eine theoretische Transformation von Variablen nicht möglich. Liegen quantitative Variablen vor, kann eine empirische Transformation, zum Beispiel in Form einer empirischen z-Transformation, durchgeführt werden. Diese hat folgende Konsequenzen:

1. *Vergleiche von Objekten sind nur mehr innerhalb einer Variablen möglich.* Eine Aussage der Art »Objekt A hat in der Variablen V_4 einen größeren Wert als Objekt B« ist zulässig, da der Vergleich innerhalb der Variablen V_4 stattfindet. Eine Aussage der Art »Objekt A hat in der Variablen V_4 einen größeren Wert als in der Variablen V_5 « ist dagegen nicht zulässig.
2. *Unterschiedliche empirische Standardabweichungen werden beseitigt.* Die empirische Varianz kann man sich als Summe der Varianz zwischen den Clustern und der Varianz innerhalb der Cluster vorstellen. Die Varianz innerhalb der Cluster kann wie bei der Varianzanalyse als Fehlervarianz interpretiert werden. Eine hohe empirische Varianz kann zwei Ursachen haben:
 - *Hohe Varianz zwischen den Clustern:* Die Variable trennt die Cluster sehr gut. In diesem Fall ist eine Standardisierung unerwünscht, da die Variable in der Analyse ein kleineres Gewicht erhalten würde.
 - *Hohe Fehlervarianz:*⁴ Die Variable trennt die Cluster nicht. Die Unterschiede sind rein zufällig. In diesem Fall führt eine Standardisierung zu einem positiven Effekt. Zufällige Unterschiede werden beseitigt bzw. reduziert.

Die Standardisierung kann günstige Effekte (geringere Gewichtung irrelevanter Variablen), aber auch negative Effekte (geringere Gewichtung von Variablen, die Cluster gut trennen) haben. Ein Vorteil ist, dass empirisch standardisierte Werte innerhalb einer Variablen leicht zu interpretieren sind. Ein hoher positiver bzw. negativer Wert (zum Beispiel von -2 bzw. +2 oder von +3 bzw. -3) bedeutet eine (relativ) hohe Abweichung vom Gesamtmittelwert. Anzumerken zur Frage, ob theoretisch oder empirisch transformiert werden soll, ist, dass der Effekt einer empirischen Standardisierung nicht überschätzt werden soll und die Ergebnisse relativ stabil sind. Dies zeigen auch die Simulationsstudien von Milligan (1980).

In der Forschungspraxis kann man sich an folgenden Regeln orientieren:

⁴ Diese zweite Fehlerursache wurde von Schlosser (1976, S. 79) in seiner Kritik an der empirischen Standardisierung übersehen. Schlosser nimmt an, dass kleine Varianzen ein hohes Fehlerausmaß bedeuten und große Varianzen ein Hinweis auf Clusterunterschiede sind.

1. Eine empirische Standardisierung muss immer dann durchgeführt werden, wenn eine theoretische Standardisierung nicht möglich ist (siehe dazu das Beispiel der Entwicklungsländerdaten).
2. Ist eine theoretische Standardisierung möglich, wird man diese durchführen, wenn größere Varianzen in einer oder mehreren Variablen ein Hinweis auf Unterschiede zwischen den Clustern (Varianz zwischen den Clustern) sind.
3. Sind größere Varianzen dagegen auf eine höhere Fehlerstreuung (Varianz in den Clustern) zurückzuführen, wird man sich für eine empirische Standardisierung entscheiden.
4. Ist eine eindeutige Entscheidung für eine theoretische oder empirische Standardisierung nicht möglich, wird man eine Analyse mit beiden Varianten rechnen.
5. Als ein weiteres Entscheidungskriterium können inferenzstatistische Überlegungen dienen. So zum Beispiel setzen statistische Signifikanztests für die euklidische Distanz oder die quadrierte euklidische Distanz empirisch standardisierte, unabhängige und normalverteilte Variablen voraus (siehe Abschnitt 8.5). Sollen derartige Tests durchgeführt werden, wird man sich für eine empirische Standardisierung entscheiden.

Von der Frage, ob die in einer objektorientierten Analyse einbezogenen, nichtvergleichbaren Variablen theoretisch oder empirisch zu standardisieren sind, ist die Interpretation einer Clusterlösung zu unterscheiden. Zur Interpretation können sowohl empirisch standardisierte Werte als auch Rohwerte und theoretisch standardisierte Werte verwendet werden, um die Vorteile der einzelnen Skalen bei der Interpretation zu nutzen. Diese sind:

1. Die empirisch standardisierten Variablen geben Anhaltspunkte, wie stark ein Cluster in einer Variablen vom Gesamtmittelwert abweicht.
2. Die nichtstandardisierten Rohwerte ermöglichen eine Interpretation in der ursprünglichen Skala und Vergleiche zwischen jenen Variablen, die in derselben Einheit gemessen sind.
3. Die theoretisch standardisierten Variablen schließlich ermöglichen auch Vergleiche zwischen jenen Variablen, die in unterschiedlichen Skalen gemessen wurden.

7.4 Hierarchische Variablen

In dem vorausgehenden Abschnitt wurde bereits darauf hingewiesen, dass Vergleichbarkeit von Variablen auch durch Umkodierungen erreicht werden kann. Durch Umkodierungen können die meisten *hierarchischen Variablen* in *nicht-hierarchische Variablen*

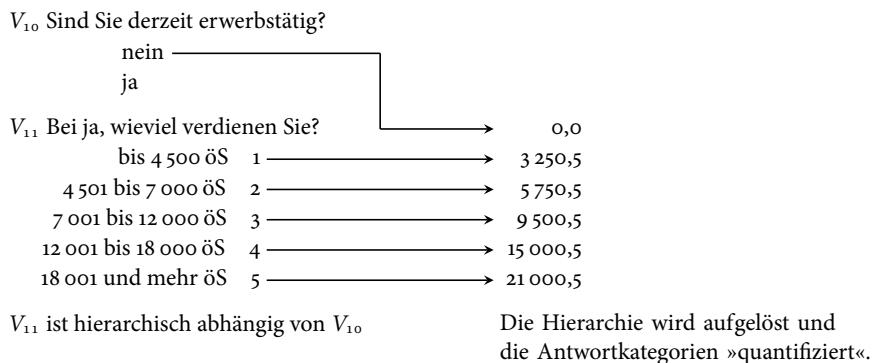


Abb. 7.2: »Quantifizierung« der Antwortkategorien durch Umkodierung von hierarchischen Variablen

aufgelöst werden. Betrachten wir dazu folgendes Beispiel: In einer Untersuchung wurden die Befragten zunächst danach gefragt, ob sie erwerbstätig sind und wie viel sie gegebenenfalls verdienen. Die Variable »Einkommen« hängt also hierarchisch von der Variablen »Erwerbstätigkeit« ab. Aus diesen beiden Variablen lässt sich eine neue Variable erzeugen, indem Personen, die derzeit nicht erwerbstätig sind, ein Einkommen von 0 zugewiesen wird. Die anderen Einkommenskategorien können durch Bildung der Intervallmittnen (zum Beispiel ist die Intervallmitte der Kategorie 2 gleich $(4,501 + 7,000)/2 = 5,750,5$) quantifiziert werden. Das in Abbildung 7.2 dargestellte Schema verdeutlicht das Vorgehen.

7.5 Gemischte Variablen

Für gemischte Variablen werden unterschiedliche Lösungsvorschläge in Bezug auf das Problem der Nichtvergleichbarkeit vorgeschlagen (Gordon 1981, 1999, S. 21–34; Opitz 1980, S. 57–64), von denen hier nur einige angeführt seien:

1. *Reduktion des Messniveaus:* Diese Strategie kann mitunter mit einem erheblichen Informationsverlust verbunden sein.
 2. *Gewichtung bzw. Transformation der Variablen.*
 3. *Gewichtung der Distanzen* (siehe dazu Abschnitt 8.7).
 4. *Verwendung der probabilistischen Clusteranalyse:* Hier tritt das Problem der Nicht-vergleichbarkeit von gemischten Variablen nicht auf, da nicht mit Distanzen oder Korrelationen gerechnet wird, sondern mit Wahrscheinlichkeiten, die unabhängig vom Messniveau auf das Intervall $[0,1]$ normiert sind (siehe dazu Abschnitt 15.4).

5. Berechnung von quantitativen Skalenwerten für nominale und ordinale Variablen mittels multipler Korrespondanzanalyse (siehe dazu Kapitel 3).
6. Einsatz spezieller Distanzmaße, wie zum Beispiel dem Distanzmaß von Gower (1971 siehe Abschnitt 12.14).

Die erste Lösungsstrategie sollte wegen des Informationsverlusts nicht gewählt werden. Wir wollen hier die zweite Strategie beschrieben. Auf die anderen Strategien wurde bereits oder wird in späteren Kapiteln eingegangen. Das Vorgehen bei der zweiten Strategie ist:

1. Die nominalen Variablen werden in Dummies aufgelöst. Die Dummies können in der weiteren Analyse wie quantitative Variablen behandelt werden.
2. Ordinale Variablen werden wie quantitative Variablen behandelt. (Alternativ können sie auch als nominal betrachtet werden.)
3. Die Variablen (Dummies, ordinale und quantitative Variablen) werden empirisch oder theoretisch standardisiert.
4. Für die Dummies der nominalen Variablen ist eine weitere Gewichtung erforderlich: Sie müssen mit 0,5 multipliziert werden, da ansonsten die Unterschiede in den nominalen Variablen ein doppeltes Gewicht erhalten würden. Von diesem Sachverhalt kann man sich leicht überzeugen. Betrachten wir dazu die Variable »Geschlecht«. Da diese Variable dichotom ist, kann sie als quantitative Variable behandelt werden. Tut man dieses, ist die maximale Differenz für Personen mit unterschiedlichem Geschlecht gleich 1. Wird das Geschlecht dagegen als nominale Variable betrachtet und in zwei Dummies aufgelöst, nimmt die maximale Differenz einen Wert von 2 an. Durch eine Gewichtung mit 0,5 der Dummies kann dieser Effekt beseitigt werden.
5. Die Unähnlichkeit zwischen den Objekten wird durch die City-Block-Metrik gemessen.
6. Die Clusteranalyse wird durchgeführt.

Anstelle der City-Block-Metrik kann auch ein anderes Distanzmaß (zum Beispiel quadrierte euklidische Distanzen) eingesetzt werden. Allerdings sind dann die Dummies anders zu gewichten. Wird beispielsweise die quadrierte euklidische Distanz verwendet, müssen die Dummies mit $\sqrt{0,5}$ multipliziert werden.

Wir wollen die Verfahrensschritte für zwei befragte Personen g und g^* darstellen (siehe Tabelle 7.4 auf der nächsten Seite). Beide Personen wurden zufällig aus den Daten von Denz (1989) ausgewählt: Die Person g ist männlich (Geschl: 1) und besucht derzeit eine Berufsschule (Schtyp: 3). Der Vater übt eine Tätigkeit mit einem mittleren Berufsprestige (BerufV: 4) aus und hat eine geringe Schulbildung (SchV: 1) abgeschlossen. Die Mutter ist erwerbstätig (ErwM: 1) und hat ebenfalls eine geringe Schulbildung (SchM: 1). Die Person

Tab. 7.4: Datenmatrix für zwei befragte Jugendliche in den untersuchten sozialstrukturellen Variablen (Abkürzungen siehe Text)

	Schtyp	Geschl	BerufV	ErwM	SchV	SchM
g	3	1	4	1	1	1
g^*	1	1	5	1	4	8

(a) Ausgangsdatenmatrix

	BHS	AHS	BS	Geschl	BerufV	ErwM	SchV	SchM
g	0	0	1	1	4	1	1	1
g^*	1	0	0	1	5	1	4	8

(b) Datenmatrix nach Dummy-Auflösung der nominalen Variablen

	BHS	AHS	BS	Geschl	BerufV	ErwM	SchV	SchM
μ_i	0,50	0,50	0,50	1,50	4,00	1,50	4,50	4,50
σ_i	0,50	0,50	0,50	0,50	2,00	0,50	2,29	2,29

(c) theoretische Skalenkennwerte

	BHS	AHS	BS	Geschl	BerufV	ErwM	SchV	SchM
g	-0,50	-0,50	0,50	-1,00	0,00	-1,00	-1,53	-1,53
g^*	0,50	-0,50	-0,50	-1,00	0,50	-1,00	-0,22	1,53

(d) standardisierte Datenmatrix (Dummies mit 0,5 multipliziert)

BHS	AHS	BS	Geschl	BerufV	ErwM	SchV	SchM	Σ
1	0	1	0	0,50	0	1,31	3,06	6,87

(e) Absolute Differenzen in den einzelnen Variablen

g^* ist ebenfalls männlich, besucht aber eine berufsbildende höhere Schule (Schtyp: 1). Der Vater hat eine mittlere Schulbildung (SchV: 4) abgeschlossen und übt derzeit einen Beruf mit einem mittleren Berufspristige (BerufV: 5) aus. Die Mutter hat eine hohe Schulbildung (SchM: 8) und ist ebenfalls erwerbstätig (ErwM: 1). Der derzeit von den Jugendlichen besuchte Schultyp ist eine nominalskalierte Variable mit drei Ausprägungen (»BHS«, »AHS« und »BS«). Diese Variable wird daher für die Analyse in drei Dummies aufgelöst. Die Person g erhält in der Dummy-Variablen »BS« den Wert 1, da sie derzeit

eine Berufsschule besucht. Die Person g^* erhält dagegen in der Dummy-Variablen »BHS« den Wert 1. Alle anderen Variablen werden als quantitativ betrachtet.⁵

Zur Erreichung der Vergleichbarkeit soll eine theoretische Standardisierung durchgeführt werden. Die theoretischen Skalenkennwerte können der Tabelle 7.2 auf Seite 180 entnommen werden und sind in der Tabelle 7.4 wiedergegeben: Die dichotomen Variablen »Geschlecht« und »Erwerbstätigkeit der Mutter« haben zwei Ausprägungen, nämlich 1 und 2. Ihr theoretischer Skalenmittelwert ist daher gleich 1,5 und ihre theoretische Standardabweichung gleich 0,5 (siehe Tabelle 7.4). Das Berufsprestige des Vaters ist eine ordinale Variable mit Werten von 1 bis 7. Für eine siebenstufige Skala ist entsprechend Tabelle 7.2 auf Seite 180 der theoretische Skalenmittelwert gleich 4,0 und die theoretische Standardabweichung gleich 2,0. Die »Variablen Schulbildung des Vaters« und »Schulbildung der Mutter« besitzen acht Ausprägungen. Die theoretische Skalenmittelwerte sind daher gleich 4,5 und die entsprechenden Standardabweichungen gleich 2,29. Die Dummy-Variablen sind ebenfalls dichotome Variablen. Ihre theoretische Standardabweichung ist daher gleich 0,5. Da sie nicht mit 1 und 2 sondern mit 0 und 1 kodiert sind, ist der theoretische Skalenmittelwert gleich 0,5. Unter Verwendung der theoretischen Skalenkennwerte werden die Ausprägungen der Personen entsprechend der Gleichung 7.1 auf Seite 177 standardisiert. Für die Person g ergibt sich in der Dummy-Variablen BHS eine Wert von $-1,0 (= (0 - 0,5) / 0,5)$. Dieser wird aus den genannten Gründen für die Analyse mit 0,5 multipliziert. In den einzelnen Variablen werden die absoluten Abweichungen für die Personen g und g^* berechnet und summiert. Dieser Vorgang ergibt die City-Block-Metrik, die einen Wert von 6,87 hat. Dieses Distanzmaß kann für alle Befragtenpaare berechnet werden. Für die dabei entstehende Distanzmatrix kann eine Clusteranalyse gerechnet werden.

Für unser Beispiel ergibt sich für alle von Denz (1982) befragten Jugendliche eine 8-Clusterlösung mit den in Tabelle 7.5 auf Seite 189 dargestellten Clusterzentren: Es liegen zwei sehr große Cluster (C_2 und C_5) vor, die durch eine mittlere Bildungs- und Berufsherkunft gekennzeichnet sind. Der Mittelwert im Berufsprestige des Vaters beträgt 3,32 bzw. 3,09 auf der siebenstufigen Prestigeskala. Die Mittelwerte in der abgeschlossenen Schulbildung der Eltern variieren zwischen 3,21 und 3,91. Bezogen auf die achtstufige Skala (1: »geringe Schulbildung«, 8: »hohe Schulbildung«) liegt somit ebenfalls eine mittlere bis geringe Schulbildung vor. Diesen beiden Clustern gehört der Großteil der in die Analyse einbezogenen Befragten (130 von 216, das heißt 60 Prozent) an. Die beiden Cluster unterscheiden sich nur darin, dass das Cluster C_2 von männlichen Jugendlichen gebildet wird, das Cluster C_5 dagegen von weiblichen. Eine BHS und BS wird zu jeweils ungefähr 40 Prozent besucht, eine AHS zu 20 Prozent. 60 Prozent der Befragten kommen

5 Für das Geschlecht und die Erwerbstätigkeit der Mutter ist eine Dummy-Auflösung nicht erforderlich, da sie nur zwei Ausprägungen besitzen. Die Variablen können unmittelbar als quantitativ betrachtet werden.

somit aus der Mittelschicht und besuchen zu jeweils 40 Prozent eine BHS oder BS und zu 20 Prozent eine AHS. Diese Verteilung auf die untersuchten Schultypen entspricht jener der Gesamtpopulation. Es liegen daher keine z-Werte größer +3 oder kleiner -3 hinsichtlich des Schultyps vor. Die Bildung der Eltern ist im Vergleich zu allen Befragten unterdurchschnittlich, daher ist bei den z-Werten auch ein Minus ausgewiesen. Neben diesen beiden Clustern gibt es noch ein weiteres relativ großes Cluster C_1 ($n = 34$) von männlichen Jugendlichen, die eine Schulbildung mit Maturaabschluss (AHS) besuchen. Die Jugendlichen kommen aus einer mittleren bis höheren Bildungs- und Berufsschicht, ihre Mütter sind erwerbstätig. Auf Seiten der weiblichen Jugendlichen wird dieser Typus in zwei Cluster aufgespalten: in ein Cluster der BHS-Schülerinnen (C_6), wo etwa zwei Drittel der Mütter erwerbstätig sind, und in ein Cluster der AHS-Schülerinnen (C_7), wo alle Mütter erwerbstätig sind. Die verbleibenden drei Cluster sind nur mehr schwach besetzt: Das Cluster C_3 ist durch BHS-Schüler und Schülerinnen aus der unteren Bildungs- und Berufsschicht gekennzeichnet. Das Cluster C_4 wird von zwei Jugendlichen gebildet, die sich trotz einer sehr hohen Bildungsherkunft für eine Berufsschule entschieden haben. Das Cluster C_8 schließlich wird von AHS-Schülerinnen und -Schülern gebildet, deren Väter aus einer sehr hohen Bildungsschicht kommen, die Mütter dagegen aus einer mittleren.

Zusammenfassend zeigen die Ergebnisse: In mittleren sozialen Schichten ist die Erwerbsbeteiligung der Mütter gering. Einer Berufsbildung – in Form einer BHS oder BS – wird ein größeres Gewicht beigemessen als dem Besuch einer AHS. In höheren sozialen Schichten liegt in der Regel eine hohe Erwerbsbeteiligung der Mütter vor. Eine Ausnahme bilden Eltern, die nicht bildungshomogam geheiratet haben. Hier besitzen die Mütter eine geringere Schulbildung und sind in der Regel nicht erwerbstätig. In diesen höheren sozialen Schichten wird bis auf wenige Ausnahmen ausschließlich eine höhere Schule besucht.

7.6 Standardisierung von Objekten

Die bisherigen Ausführungen bezogen sich ausschließlich auf die Variablen, die in die Analyse einbezogen werden sollen. Neben einer empirischen oder theoretischen Standardisierung von Variablen kann auch eine (*empirische*) Mittelwertzentrierung oder *Standardisierung von Objekten* durchgeführt werden. Dazu wird für jedes Objekt (Zeile einer Datenmatrix) der Mittelwert und die Standardabweichung in den zu analysierenden Variablen berechnet mit

$$\bar{x}_g = \frac{\sum_i x_{gi}}{m} \quad \text{und} \quad s_g = \sqrt{\frac{\sum_i (x_{gi} - \bar{x}_g)^2}{m}},$$

Tab. 7.5: Clusterzentren der 8-Clusterlösung für gemischte Merkmale

	<i>n</i>	männlich	ErwM	BerufV	SchV	SchM	BHS	AHS	BS
C_1	34	1,00	0,97	4,41	5,74	5,76	0,47	0,53	0,00
C_2	63	1,00	0,14	3,32	3,90	3,35	0,40	0,21	0,39
C_3	6	0,33	1,00	2,67	2,33	2,33	1,00	0,00	0,00
C_4	2	1,00	1,00	3,50	7,50	7,50	0,00	0,00	1,00
C_5	67	0,00	0,48	3,09	3,91	3,21	0,37	0,21	0,42
C_6	21	0,00	0,67	4,80	6,20	5,52	1,00	0,00	0,00
C_7	10	0,00	1,00	4,90	6,60	5,80	0,00	1,00	0,00
C_8	13	0,46	0,00	5,55	7,42	4,92	0,00	1,00	0,00

Abkürzungen: siehe Text

(a) Mittel- bzw. Anteilswerte

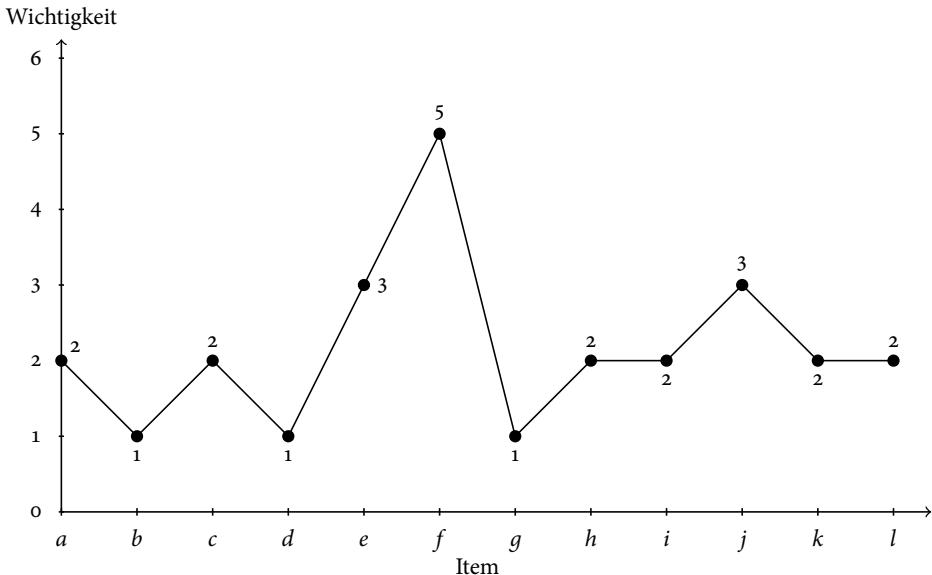
	<i>n</i>	männlich	ErwM	BerufV	SchV	SchM	BHS	AHS	BS
C_1	34	+	+	+	+	+		+	-
C_2	63	+	-		-	-			
C_3	6		+	-	-	-	+	-	-
C_4	2	+	+		+	+	-	-	+
C_5	67	-		-	-	-			
C_6	21	-	+	+	+	+	+	-	-
C_7	10	-	+	+	+	+	-	+	-
C_8	13		-	+	+		-	+	-

Abkürzungen: »+«: Werte größer 3, »-«: Werte kleiner -3

(b) z-Werte

wobei über alle Variablen m summiert wird. Der Mittelwert \bar{x}_g eines Objekts g wird in der clusteranalytischen Literatur als Profilhöhe bezeichnet, die Standardabweichung s_g als Profilstreuung oder Profilstandardabweichung. Diese Namensgebung röhrt daher, dass der Antwortvektor eines Objekts g als Profil dargestellt werden kann. Abbildung 7.3 auf der nächsten Seite gibt ein Beispiel für ein Profil aus der Wertestudie von Denz (1989). Der Befragte hat eine starke Präferenz für Ziele der demokratischen Mitbestimmung (Variablen b , d und g). Diese werden für sehr wichtig gehalten. Als wichtig gelten noch humanistische Werte (h und k), Umweltschutz (l), wirtschaftliche Werte (c und i) sowie Aufrechterhaltung der Ordnung (a). Stark abgelehnt wird eine starke Landesverteidigung (f), abgelehnt wird die Bekämpfung der Inflation (e) und die Verbrechensbekämpfung (j). Für den Befragten ergibt sich eine Profilhöhe von

$$\bar{x}_g = \frac{(2 + 1 + 2 + 1 + \dots + 2)}{12} = 2,17$$



Abkürzungen: a: Aufrechterhaltung der Ordnung, b: mehr Mitsprache bei wichtigen Regierungsentscheidungen, c: Kampf gegen steigende Preise, d: Schutz der freien Meinungsausübung, e: Erhaltung des wirtschaftlichen Wachstums, f: eine starke Landesverteidigung, g: verstärktes Mitspracherecht am Arbeitsplatz, h: menschengerechte Städte, i: stabile Wirtschaft, j: Kampf gegen das Verbrechen, k: eine humane und weniger unpersönliche Gesellschaft, l: Erhaltung der Natur.

1: sehr wichtig, 2: wichtig, 3: weniger wichtig, 4: eher unwichtig, 5: vollkommen unwichtig.

Abb. 7.3: Ein Beispiel für ein Antwortprofil eines Befragten g

und eine Profilstandardabweichung von

$$s_g = \sqrt{\frac{(2 - 2,17)^2 + \dots + (2 - 2,17)^2}{12}} = 1,07 .$$

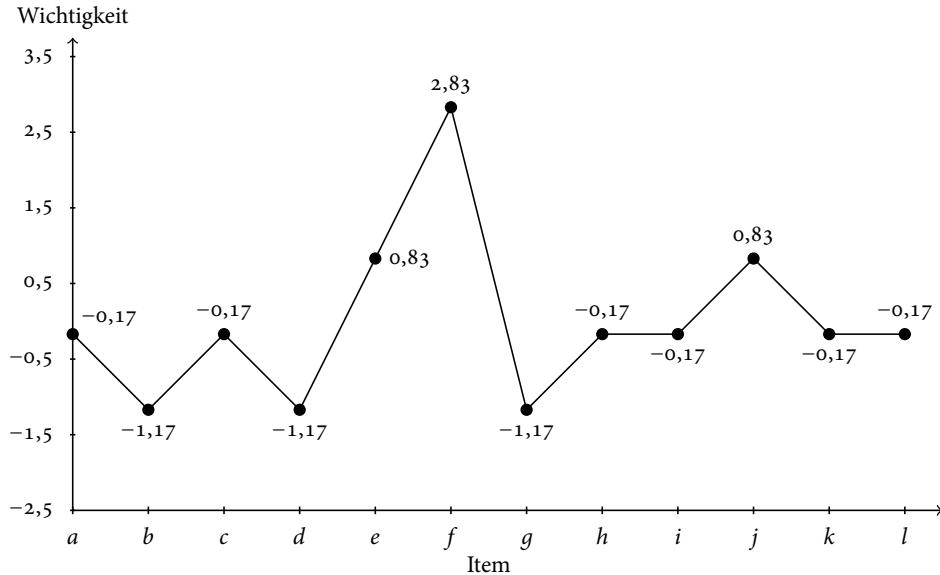
Mit diesen Kennwerten können folgende Transformationen durchgeführt werden:

- Mittelwertzentrierung
- Standardisierung

Die *Mittelwertzentrierung* ist definiert als:

$$x'_{gi} = x_{gi} - \bar{x}_g .$$

Aus dem Profil jeder Person g bzw. allgemein des Objekts g wird die Profilhöhe herausgenommen. Diese Transformation führt dazu, dass alle Personen dieselbe Profilhöhe besitzen. Inhaltlich bedeutet sie, dass nicht mehr die absolute Bewertung jedes Items

Abb. 7.4: Mittelwertzentriertes Profil der Person g der Abbildung 7.3

untersucht wird, sondern die Präferenzen jeder Person.⁶ Für das in der Abbildung 7.3 dargestellte Profil ergibt sich das in Abbildung 7.4 dargestellte *mittelwertzentrierte* Profil. Negative Zahlen drücken eine größere Präferenz aus, da »sehr wichtig« mit »1« kodiert wurde und »vollkommen unwichtig« mit »5«. »1« wird daher nach einer Mittelwertzentrierung negativ.

Die *Standardisierung* ist definiert mit:

$$x_{gi}'' = \frac{x_{gi} - \bar{x}_g}{s_g} = \frac{x'_{gi}}{s_g} .$$

Aus dem Profil jeder Person g wird die Profilhöhe und die Profilstreuung herausgenommen. Alle Profile erhalten dieselbe Profilhöhe von 0 und die Profilstandardabweichung von 1. Ist die Profilstreuung gleich 0, sind die standardisierten Profilwerte gleich null. In unserem Beispiel entspricht das standardisierte Profil weitgehend jenem des mittelwertzentrierten Profils, da die Standardabweichung nahe 1 liegt. Das Profil wird in der Vertikalen nur geringfügig zusammengestaucht. Aus dem Wert 2,83 wird der Wert 2,64 ($= 2,83/1,07$) und aus dem Wert $-1,17$ der Wert $-1,09$ ($= -1,17/1,07$).

⁶ Mathematisch wird die Mittelwertzentrierung der Objekte als ipsative Transformation bezeichnet (siehe Abschnitt 5.3.1, Horst 1965, S. 291–295).

Tab. 7.6: Rohprofile, die das gleiche standardisierte Profil besitzen

Person	Rohprofil				\bar{x}_g	s_g	standard. Profil			
	a	b	c	d			a	b	c	d
1	1	1	2	2	1,5	1,5	-1	1	1	-1
2	1	3	3	1	2,0	1,0	-1	1	1	-1
3	1	4	4	1	2,5	1,5	-1	1	1	-1
4	1	3	5	1	3,0	2,0	-1	1	1	-1
5	2	3	3	2	2,5	0,5	-1	1	1	-1
6	4	5	5	4	4,5	0,5	-1	1	1	-1

Die Standardisierung der Profile führt dazu, dass nur mehr relative Präferenzen abgebildet werden. Alle in der Tabelle 7.6 angeführten Personen besitzen das gleiche standardisierte Profil, obwohl sich die Rohprofile deutlich unterscheiden, da nur relative Präferenzen abgebildet werden. Die (absolute) Wichtigkeit der einzelnen Items spielt keine Rolle mehr. So besitzen die Person 1 und 6 das gleiche standardisierte Profil, obwohl sich ihre Antwortmuster (Rohprofile) deutlich unterscheiden. Auch die Stärke der Präferenz spielt keine Rolle mehr. So zum Beispiel hat Person 4 eine starke Präferenz für die Items a und d , da die Items b und c deutlich abgelehnt werden, die Person 1 dagegen nur eine sehr schwache Präferenz. Alle diese Unterschiede gehen bei einer Standardisierung der Objekte verloren.

Zusammenfassend hat eine Mittelwertzentrierung oder Standardisierung der Objekte somit folgende Effekte:

- Die Mittelwertzentrierung führt dazu, dass nur mehr Präferenzen interpretierbar sind. Aussagen über die Einstellung (Bewertung) einzelner Items sind nicht mehr möglich.
- Die Standardisierung führt darüber hinaus dazu, dass auch die Stärke der Präferenz eliminiert wird.

Ob in einer Analyse einer der beiden Effekte erwünscht ist, hängt von der inhaltlichen Fragestellung ab. In der Regel sind beide Effekte nicht erwünscht, so dass keine Standardisierung der Objekte durchgeführt wird. Bezuglich der Auswahl eines (Un-)Ähnlichkeitsmaßes ist dann darauf zu achten, dass dabei nicht implizit eine Standardisierung der Objekte durchgeführt wird. Dies ist beispielsweise bei der Verwendung von Korrelationskoeffizienten der Fall.

7.7 Exkurs: Die Problematik einer automatischen Orthogonalisierung

In der Literatur (siehe zum Beispiel Green und Tull 1982, S. 414; Lüdtke 1989; Kaufman 1985) werden die Variablen vor der Clusteranalyse oft *orthogonalisiert*, indem die *unrotierten Hauptkomponenten* als Variablen verwendet werden (siehe Abschnitt 5.3). Dies wird damit begründet, dass das verwendete Verfahren unabhängige Variablen bzw. einen orthogonalen Merkmalsraum⁷ voraussetzt. Eine *automatische Orthogonalisierung* ist aber nicht unproblematisch, da empirische Korrelationen ein Hinweis auf eine vorhandene Clusterstruktur sein können, die durch eine automatische Orthogonalisierung zerstört wird. Eine inhaltlich und empirisch begründete Entscheidung ist erforderlich, ob empirische Korrelationen ein Hinweis auf ein Clustermodell auf einer latenten Ebene oder auf ein Clustermodell auf einer manifesten Ebene sind. Im letzten Fall sind die Korrelationen durch die Clusterstruktur bedingt und sollten nicht durch eine automatische Orthogonalisierung beseitigt werden. Empirisch begründet werden kann eine Entscheidung durch eine Faktorenanalyse. In der weiteren Analyse werden dann aber nicht die unrotierten Faktoren (Hauptkomponenten) verwendet, sondern die inhaltlich interpretierten (rotierten) Faktoren.

Hinzu kommt, dass ein automatisches Vorgehen der oben dargestellten Art – im Unterschied zu einer dimensionalen Analyse durch eine Faktorenanalyse – nicht zur Elimination von Fehlerquellen beiträgt, wenn eine bestimmte Berechnungsmethode gewählt wird. Kaufman (1985) untersucht in Simulationsstudien, ob durch eine Orthogonalisierung Messfehler beseitigt werden können. Er vergleicht folgende Vorgehensweisen:

1. Verwendung der empirisch standardisierten Werte,
2. Verwendung der bedeutsamen ungewichteten Hauptkomponenten (Hauptkomponenten mit Eigenwerten größer 1),
3. Verwendung der bedeutsamen gewichteten Hauptkomponenten,
4. Verwendung aller ungewichteten Hauptkomponenten und
5. Verwendung aller gewichteten Hauptkomponenten.

Die Ergebnisse zeigen, dass bei der Verwendung aller ungewichteten Hauptkomponenten – eine Methode der automatischen Orthogonalisierung – wesentlich schlechtere Ergebnisse erzielt werden als mit den anderen vier Methoden. Der Anteil der Fehlklassifikationen beträgt für die untersuchten Modellkonstellationen (im Durchschnitt 5 Prozent

⁷ Die Annahme eines orthogonalen Merkmalsraumes ist nur für bestimmte statistische Testverfahren erforderlich.

Messfehler und durchschnittlich 5 Prozent fehlende Werte) 45,87 Prozent, wenn alle ungewichteten Hauptkomponenten verwendet werden. Bei den anderen Methoden variiert die Fehlerquote zwischen 2,00 Prozent, wenn alle gewichteten Hauptkomponenten (eine andere Methode der automatischen Orthogonalisierung) und 5,65 Prozent, wenn nur bedeutsame ungewichtete Hauptkomponenten verwendet werden. Bei einer Analyse mit standardisierten Werten werden im Vergleich zur Verwendung von gewichteten Hauptkomponenten (Methode 3 und 5) nur geringfügig schlechtere Ergebnisse erzielt. Der Fehlklassifikationsanteil beträgt 2,65 Prozent. Clusteranalyseprogramme, die automatisch eine Orthogonalisierung durchführen, sind daher mit Vorsicht anzuwenden. Auf jeden Fall ist hier zu prüfen, ob ungewichtete oder gewichtete Hauptkomponenten verwendet werden. Zu empfehlen sind gewichtete Hauptkomponenten. Diese sind dadurch gekennzeichnet, dass ihre Varianz gleich dem Eigenwert ist. Ungewichtete Hauptkomponenten sind dagegen auf eine Standardabweichung von 1,0 normiert.

Allgemein empfehlen wir:

1. Anstelle einer automatischen orthogonalen Transformation sollte eine Faktorenanalyse durchgeführt und mit inhaltlich daraus abgeleiteten Variablen (Gesamtpunktwerte oder Faktorwerte) weitergerechnet werden.
2. Falls eine automatische orthogonale Transformation durchgeführt werden soll, sollten gewichtete Hauptkomponenten eingesetzt werden.

8 Unähnlichkeits- und Ähnlichkeitsmaße

Bei bestimmten deterministischen Clusteranalyseverfahren (Complete-Linkage, Single-Linkage, Mittelwertverfahren, Complete-Linkage für überlappende Cluster, verallgemeinerte Nächste-Nachbarn-Verfahren, Repräsentanten-Verfahren) muss der Anwender ein für die Daten und das Analyseziel geeignetes (Un-)Ähnlichkeitsmaß auswählen. Die Situation ist somit ähnlich jener der nichtmetrischen mehrdimensionalen Skalierung. Auch hier muss eine Auswahl eines Ähnlichkeitsmaßes erfolgen, sofern nicht direkt eine (Un-)Ähnlichkeitsmatrix erhoben wurde. Ähnlichkeitsmaße sind dadurch gekennzeichnet, dass ein größerer Zahlenwert eine größere Ähnlichkeit bedeutet. Für sie soll im folgenden das Symbol \ddot{a}_{ij} bzw. \ddot{a}_{g,g^*} verwendet werden.¹ Die Indizierung (i,j) bezieht sich auf Variablenpaare, also auf eine variablenorientierte Analyse. Mit der Indizierung (g,g^*) sind dagegen Objektpaare gemeint. Ist also \ddot{a}_{g,g^*} größer $\ddot{a}_{g^{**},g^{***}}$, so sind sich die Objekte g und g^* ähnlicher als die Objekte g^{**} und g^{***} . Umgekehrt bedeutet ein kleinerer Zahlenwert bei einem Unähnlichkeitsmaß eine größere Ähnlichkeit. Für Unähnlichkeitsmaße soll das Symbol u verwendet werden. Ist u_{g,g^*} größer $u_{g^{**},g^{***}}$, so sind sich die Objekte g und g^* unähnlicher als die Objekte g^{**} und g^{***} . Die (Un-)Ähnlichkeitsmaße lassen sich in vier Gruppen einteilen:

1. *Korrelationskoeffizienten als Ähnlichkeitsmaße*: Für sie wird auch die Bezeichnung Assoziations- oder Zusammenhangsmaß verwendet.
2. *Distanzmaße (D-Maße) als Unähnlichkeitsmaße*: Diese werden nach der verallgemeinerten Minkowski-Metrik berechnet (siehe Formel 8.12 auf Seite 219).
3. *Aus Distanzmaßen oder Korrelationskoeffizienten abgeleitete (Un-)Ähnlichkeitsmaße*: Durch monotone Transformationen können aus Distanzmaßen oder Korrelationsmaßen weitere Ähnlichkeits- oder Unähnlichkeitsmaße abgeleitet werden.
4. *Andere (Un-)Ähnlichkeitsmaße*: Dabei handelt es sich um eine Restgruppe von Koeffizienten, die für spezifische Fragestellungen und Messniveaus entwickelt wurden.

¹ Zur besseren Lesbarkeit verwenden wir ein Komma zwischen den Indizes, sofern dies erforderlich ist.

8.1 Auswahl eines (Un-)Ähnlichkeitsmaßes

Die Frage, welches bzw. welche (Un-)Ähnlichkeitsmaße in einem Anwendungsfall zu wählen sind, hängt vor allem von der Richtung der Datenanalyse (objekt- oder variablenorientierte Auswertung) ab. Allgemein lässt sich festhalten:

1. Für eine variablen- und objektorientierte Analyse können im Prinzip sowohl Korrelationskoeffizienten als auch Distanzmaße und daraus abgeleitete (Un-)Ähnlichkeitsmaße verwendet werden.
2. Nicht im Rahmen einer objektorientierten Clusteranalyse eingesetzt werden können Korrelationskoeffizienten für nominalskalierte Variablen mit mehr als zwei Ausprägungen (*Kontingenzkoeffizient C*, *Cramérs V* usw.). Sie stellen insofern eine Ausnahme dar.
3. Für eine variablenorientierte Clusteranalyse eignen sich insbesondere Korrelationskoeffizienten. Distanzmaße sind dagegen nur bedingt geeignet.
4. In einer objektorientierten Clusteranalyse dagegen sind Korrelationskoeffizienten nur bedingt geeignet, da bei deren Berechnung implizit eine Standardisierung der Objekte durchgeführt wird. Die damit verbundenen Effekte (Elimination der Profilhöhe und -streuung) sind in der Regel unerwünscht (siehe Abschnitt 7.6). In einer objektorientierten Datenanalyse werden daher üblicherweise Distanzmaße oder daraus abgeleitete (Un-)Ähnlichkeitsmaße verwendet.
5. Durch monotone Transformationen von Korrelations- oder Distanzmaßen ändern sich die Ergebnisse bei jenen Clusteranalyseverfahren, die nicht invariant gegenüber monotonen Transformationen sind: Werden dieselben Daten beispielsweise mit einem Distanzmaß, zum Beispiel der City-Block-Metrik, und mit einem daraus abgeleiteten Ähnlichkeitsmaß mit einem Verfahren, das nicht invariant gegenüber monotonen Transformationen ist, untersucht, ergeben sich unterschiedliche Klassifikationen. Die Verfahren sind allerdings relativ robust gegenüber derartigen Transformationen. Nur wenn bei der Berechnung implizit unterschiedliche Gewichtungen und Transformationen der Variablen durchgeführt werden, ergeben sich unterschiedliche Ergebnisse. So zum Beispiel wird bei der Verwendung der Q-Korrelation (Produkt-Moment-Korrelation zwischen Objekten) eine implizite Standardisierung der Objekte durchgeführt. Die Ergebnisse einer Clusteranalyse, bei der Q-Korrelationen als Ähnlichkeitsmaße verwendet werden, unterscheiden sich daher stärker von einer Analyse mit quadrierten euklidischen Distanzen als die Ergebnisse, die bei Verwendung der City-Block-Metrik anstelle der quadrierten euklidischen Distanzen erzielt werden.

Nachfolgend sollen in Abhängigkeit vom Messniveau einige Ähnlichkeits- und Unähnlichkeitsmaße beschrieben werden. Da dabei dichotomen Variablen eine Sonderstellung einnehmen, werden sie getrennt behandelt. Der Schwerpunkt liegt auf der Darstellung der Berechnung von (Un-)Ähnlichkeiten zwischen Objekten (Zeilen einer Datenmatrix). Nur gelegentlich wird auf die Berechnung der (Un-)Ähnlichkeit von Variablen eingegangen.

8.2 Dichotome Variablen

Für *dichotome Variablen* wurden eine Vielzahl von (Un-)Ähnlichkeitsmaßen entwickelt. Zunächst können bei dichotomen Variablen die *City-Block-Metrik* und die *quadrierte euklidische Distanz* als Distanzmaße und der *Produkt-Moment-Korrelationskoeffizient* (Φ -Korrelation) als Ähnlichkeitsmaß berechnet werden. Dabei werden die dichotomen Variablen wie quantitative Variablen behandelt.

Der Φ -Koeffizient als Ähnlichkeitsmaß: Diesen erhält man, wenn man für dichotome Variablen die Berechnungsformel der Produkt-Moment-Korrelation für quantitative Variablen anwendet. Da sich der Φ -Koeffizient als Korrelationskoeffizient nur bedingt für eine objektorientierte Analyse eignet, soll hier nur die Formel für ein Variablenpaar (i, j) wiedergegeben werden:

$$\Phi_{ij} = \text{PMK}_{ij} = r_{ij} = \frac{\frac{1}{n} \cdot \sum_g (x_{gi} - \bar{x}_i) \cdot (x_{gj} - \bar{x}_j)}{\sqrt{\frac{1}{n} \sum_g (x_{gi} - \bar{x}_i)^2 \cdot \frac{1}{n} \sum_g (x_{gj} - \bar{x}_j)^2}},$$

wobei PMK für Produkt-Moment-Korrelation und r für R -Korrelation steht. \bar{x}_i ist der Mittelwert der Variablen i , \bar{x}_j jener der Variablen j . Sind die Variablen mit 0/1 kodiert, sind die Mittelwerte gleich den Anteilswerten der 1-Antworten. Im Nenner stehen die Standardabweichungen der beiden Variablen.

Die City-Block-Metrik als Unähnlichkeitsmaß: Die City-Block-Metrik ergibt sich aus der verallgemeinerten Minkowski-Metrik (siehe Formel 8.12 auf Seite 219), wenn $r = q = 1$ gesetzt wird:

$$\text{CITY}_{g,g^*} = \sum_i |x_{gi} - x_{g^*i}| \quad \text{bzw.} \quad \text{CITY}_{ij} = \sum_g |x_{gi} - x_{gj}|. \quad (8.1)$$

Bei dichotomen Variablen kann der Absolutbetrag $|x_{gi} - x_{g^*i}|$ in einer Variablen i für ein Objektpaar (g, g^*) bzw. der Absolutbetrag $|x_{gi} - x_{gj}|$ in einem Objekt für ein Variablenpaar (i, j) nur die Werte 0 oder 1 annehmen. »0« bedeutet eine Übereinstimmung, »1« eine Nichtübereinstimmung. In einer objektorientierten Analyse ist also die City-Block-

Tab. 8.1: Interpretation der City-Block-Metrik bei dichotomen Variablen

Objekt	X_1	X_2	X_3	X_4	
g	1	0	0	1	
g^*	0	1	0	1	
$ x_{gi} - x_{g^*i} $	1	1	0	0	CITY = 2

Anmerkung: Die Objekte g und g^* stimmen in X_3 und X_4 überein, in X_1 und X_2 dagegen nicht.

(a) objektorientierte Betrachtung

Objekte	X_i	X_j	$ x_{gi} - x_{gj} $	
1	1	1	0	
2	1	0	1	
3	0	0	0	
4	0	0	0	
5	1	1	0	
				CITY = 1

Anmerkung: Die Variablen i und j besitzen in den Objekten 1, 3, 4 und 5 eine Übereinstimmung und in 2 keine.

(b) variablenorientierte Betrachtung

Metrik zwischen zwei Objekten g und g^* gleich der Zahl der Variablen minus der Zahl der Übereinstimmungen in allen Variablen, also die Zahl der Nichtübereinstimmungen, bei einer variablenorientierten Analyse die Zahl der Nichtübereinstimmungen in allen Objekten. In der Tabelle 8.1 besitzen die Objekte g und g^* in den Variablen X_1 und X_2 unterschiedliche Ausprägungen. In den Variablen X_3 und X_4 liegen dieselben Merkmalsausprägungen vor. Die Zahl der Nichtübereinstimmungen ist gleich 2. Berechnet man die City-Block-Metrik nach der Formel 8.1 auf der vorherigen Seite ergibt sich ein Wert von 2. Bei der variablenorientierten Betrachtung ergibt sich aufgrund der einzigen Nichtübereinstimmung in Objekt 2 hingegen ein Wert von 1.

Quadrierte euklidische Distanz: Da bei dichotomen Variablen die absoluten Abweichungen $|x_{gi} - x_{g^*i}|$ bzw. $|x_{gi} - x_{gj}|$ nur die Werte 1 und 0 annehmen können, ist die City-Block-Metrik gleich der quadrierten euklidischen Distanz:

$$\text{QEUKLID}_{g,g^*} = \text{CITY}_{g,g^*} \quad \text{bzw.} \quad \text{QEUKLID}_{ij} = \text{CITY}_{ij},$$

wobei QEUKLID für die quadrierte euklidische Distanz steht.

Tab. 8.2: Variablen- und objektorientierte Vierfeldertafel als Ausgangspunkt der Berechnung von (Un-)Ähnlichkeitskoeffizienten für dichotome Variablen

		Objekt g^*					Variable j		
Objekt g				Σ	Variable i				Σ
	0	1	Σ			0	1	Σ	
0	1	1	2	Σ	Variable i	0	2	0	2
	1	1	2			1	1	2	3
Σ		2	2	4	Σ		3	2	5

(a) objektorientiert (b) variablenorientiert

Zur Berechnung der behandelten Maßzahlen lassen sich *Vierfeldertafeln* konstruieren. Abhangig von der Richtung der Analyse kann fur jedes Objektpaar (g, g^*) bzw. jedes Variablenpaar (i, j) eine Vierfeldertafel gebildet werden. Fur die Tabelle 8.1 ergeben sich die in der Tabelle 8.2 dargestellten Vierfeldertafeln. Allgemein enthalt eine Vierfeldertafel – die hier verwendete Notation ist in Tabelle 8.3 dargestellt – folgende Informationen:

1. *Übereinstimmungen hinsichtlich des Besitzes*: Dies ist die Zelle $(1,1)$ mit einer Häufigkeit von d .
 2. *Übereinstimmungen hinsichtlich des Nichtbesitzes*: Dies ist die Zelle $(0,0)$ mit einer Häufigkeit von a .
 3. *Nichtübereinstimmungen*: Dies sind die Zellen $(1,0)$ und $(0,1)$ mit einer Häufigkeit von $c + b$.

Durch eine unterschiedliche Gewichtung dieser drei Informationen lassen sich eine Reihe von (Un-)Ähnlichkeitskoeffizienten erzeugen. Die dabei verwendete Ausgangsformel ist

$$\ddot{a}_{ij} \quad \text{bzw.} \quad \ddot{a}_{g,g^*} = \frac{\alpha \cdot a + \beta \cdot d}{\delta \cdot a + \beta \cdot d + \gamma \cdot (b + c)} \quad (8.2)$$

mit α , β , γ und δ als Gewichtungsfaktoren. In Abhangigkeit von der Wahl der Gewichte ergeben sich die in der Tabelle 8.4 auf der nachsten Seite dargestellten Ähnlichkeitsmae.

Tab. 8.3: Notation Vierfeldertafel

		Variable j oder Objekt g^*		Σ
		0	1	
Variable i oder Objekt g	0 (Nichtbesitz)	a	b	$a + b = S_1$
	1 (Besitz)	c	d	$c + d = S_2$
Σ (gesamt)		$a + c = S_3$	$b + d = S_4$	$a + b + c + d = S$

Tab. 8.4: Ähnlichkeitsmaße für dichotome Variablen

Ähnlichkeitsmaß	Berechnungsformel	Beispiel ^{a)}	Eigenschaften
JACCARD I	$\frac{d}{d + b + c}$	$\frac{1}{1 + 1 + 1} = \frac{1}{3} = 0,33$	Gemeinsamer Nichtbesitz (o,o) geht nicht in die Berechnung ein.
DICE	$\frac{2d}{2d + b + c}$	$\frac{2 \cdot 1}{2 \cdot 1 + 1 + 1} = \frac{2}{4} = 0,50$	Gemeinsamer Nichtbesitz (o,o) geht nicht in die Berechnung ein, gemeinsamer Besitz wird doppelt gewichtet.
SOKAL & SNEATH I	$\frac{d}{d + 2(b + c)}$	$\frac{1}{1 + 2 \cdot (1 + 1)} = \frac{1}{5} = 0,20$	Gemeinsamer Nichtbesitz (o,o) geht nicht in die Berechnung ein, Nichtübereinstimmungen (o,1) und (1,o) werden doppelt gewichtet.
RUSSEL & RAO	$\frac{d}{d + a + b + c}$	$\frac{1}{1 + 1 + 1 + 1} = \frac{1}{4} = 0,25$	Gemeinsamer Nichtbesitz wird nicht als Ähnlichkeit betrachtet, geht aber in den Nenner ein.
Simple-Matching	$\frac{d + a}{d + a + b + c}$	$\frac{1 + 1}{1 + 1 + 1 + 1} = \frac{2}{4} = 0,50$	Besitz und Nichtbesitz werden gleich gewichtet.
SOKAL & SNEATH II	$\frac{2(d + a)}{2(d + a) + b + c}$	$\frac{2 \cdot (1 + 1)}{2 \cdot (1 + 1) + 1 + 1} = \frac{4}{6} = 0,67$	Übereinstimmungen (o,o) und (1,1) werden doppelt gewichtet.
ROGERS & TANIMOTO	$\frac{d + a}{d + a + 2(b + c)}$	$\frac{1 + 1}{3 + 2 + 2(0 + 1)} = \frac{5}{7} = 0,71$	Nichtübereinstimmungen (o,1) und (1,o) werden doppelt gewichtet.

a) Zahlenwerte aus der objektorientierten Vierfeldertafel der Tabelle 8.2 auf der vorherigen Seite

Die Ähnlichkeitsmaße wurden nach dem Autor bzw. den Autoren bezeichnet, die das jeweilige Maß entwickelt haben (Steinhausen und Langer 1977, S. 55; Vogel 1975, S. 93–108). Die römische Nummerierung bedeutet, dass der Autor bzw. die Autoren mehrere Maße entwickelt haben. Alle Ähnlichkeitsmaße variieren zwischen 0 und 1. Ein Wert von 0 bedeutet, dass keine Ähnlichkeit (keine Übereinstimmung) vorliegt. Ein Wert von 1 bedeutet dagegen perfekte Ähnlichkeit (perfekte Übereinstimmung).

Alle Ähnlichkeitsmaße lassen sich durch die Transformation

$$u_{ij} = 1 - \ddot{a}_{ij} \quad \text{bzw.} \quad u_{g,g^*} = 1 - \ddot{a}_{g,g^*} \quad (8.3)$$

in Unähnlichkeitsmaße transformieren. Für den JACCARD-I-Koeffizient beispielsweise ergibt sich folgendes Unähnlichkeitsmaß:

$$u_{g,g^*} \quad \text{bzw.} \quad u_{ij} = 1 - \frac{d}{d + b + c} = \frac{d + b + c - d}{d + b + c} = \frac{b + c}{d + b + c}.$$

Die Beziehung zu der allgemeinen Berechnungsformel 8.2 auf Seite 199 und den in der Übersicht angeführten Berechnungsformeln lässt sich leicht herstellen. Sie soll daher nur exemplarisch für den JACCARD-I-Koeffizient verdeutlicht werden. Setzt man in Gleichung 8.2 auf Seite 199 $\alpha = \delta = 0$, $\beta = 1$ und $\gamma = 1$, ergibt sich der JACCARD-I-Koeffizient:

$$\text{JACCARD-I} = \frac{0 \cdot a + 1 \cdot d}{0 \cdot a + 1 \cdot d + 1 \cdot (b + c)} = \frac{d}{d + b + c}. \quad (8.4)$$

In Bezug auf die zu Beginn des Kapitels 8 vorgenommene Einteilung der (Un-)Ähnlichkeitsmaße gehören die in der Übersichtstabelle 8.4 enthaltenen Maßzahlen – mit Ausnahme des Simple-Matching-Koeffizienten – der vierten Restgruppe der anderen (Un-)Ähnlichkeitsmaße an. Der Simple-Matching-Koeffizient (SMK) lässt sich dagegen als lineare Transformation aus der City-Block-Metrik ableiten mit:

$$\text{SMK}_{ij} = 1 - \frac{\text{CITY}_{ij}}{n} \quad \text{bzw.} \quad \text{SMK}_{g,g^*} = 1 - \frac{\text{CITY}_{g,g^*}}{m},$$

wobei n die Zahl der Objekte und m die Zahl der Variablen ist. Der Simple-Matching-Koeffizient ist gleich dem Anteil der Übereinstimmungen des Besitzes und Nichtbesitzes.

Der *Übereinstimmungskoeffizient* κ . Er ist wie folgt definiert (Fleiss 1981, S. 217–225):

$$\kappa = \frac{U^{\text{beob}} - U^{\text{erw}}}{1 - U^{\text{erw}}}, \quad (8.5)$$

wobei U^{beob} der beobachtete Anteil der Übereinstimmungen ist. U^{beob} ist also gleich dem Simple-Matching-Koeffizienten. U^{erw} ist der bei statistischer Unabhängigkeit erwartete Anteil an Übereinstimmungen. Dieser lässt sich aufgrund von drei unterschiedlichen Nullmodellen der statistischen Unabhängigkeit berechnen, die für den Fall einer objektorientierten Vierfeldertafel dargestellt werden sollen:

1. *Nullmodell ohne Restriktionen*: In diesem Fall besitzen alle vier Zellen der Vierfeldertafel dieselbe Wahrscheinlichkeit. Der zufällig erwartete Anteil der Übereinstimmungen ist gleich 0,5. Für den Übereinstimmungskoeffizienten κ ergibt sich somit ein Wert von:

$$\kappa_{g,g^*}^{(o)} = \frac{\text{SMK}_{g,g^*} - 0,5}{1 - 0,5}.$$

2. *Nullmodell mit vorgegebener Profilhöhe:* Es wird angenommen, dass der Anteil der 1-Antworten für jedes Objekt g gegeben ist. Die Randsummen der objektorientierten Vierfeldertafel sind somit fixiert. Der erwartete Anteil an Übereinstimmungen bei gegebenen Randsummen ist gleich

$$\frac{S_1 S_3}{SS} + \frac{S_2 S_4}{SS}$$

mit $SS = S \cdot S$ als Produkt der Gesamtfallzahl. $1 - U^{\text{erw}}$ ist daher:

$$\begin{aligned} 1 - \left(\frac{S_1 S_3}{SS} + \frac{S_2 S_4}{SS} \right) &= \frac{(S_1 + S_2)(S_3 + S_4) - S_1 S_3 - S_2 S_4}{SS} \\ &= \frac{S_1 S_4 + S_2 S_3}{SS}. \end{aligned}$$

Der empirisch beobachtete Übereinstimmungsanteil U^{beob} ist gleich dem Simple-Matching-Koeffizienten $(a + d)/S$. $U^{\text{beob}} - U^{\text{erw}}$ ist daher gleich:

$$\begin{aligned} \frac{S(a + d) - S_1 S_3 - S_2 S_4}{SS} &= \frac{(a + b + c + d)(a + d) - (a + b)(a + c) - (c + d)(b + d)}{SS} \\ &= \frac{(a + b)(a + d - a - c) + (c + d)(a + d - b - d)}{SS} \\ &= \frac{2(ad - bc)}{SS}. \end{aligned}$$

Für den Übereinstimmungskoeffizienten κ ergibt sich daher die Darstellung:

$$\kappa_{g,g^*}^{(1)} = \frac{2(ad - bc)}{S_1 S_4 + S_2 S_3}. \quad (8.6)$$

Zu beachten ist, dass bei Verwendung dieses Nullmodells wie bei Φ eine Elimination der Profilhöhe stattfindet. Dies schränkt die Anwendung des nach Gleichung 8.6 berechneten Koeffizienten im Rahmen einer objektorientierten Datenanalyse ein.

3. *Nullmodell mit vorgegebenen Anteilswerten der Variablen:* Im Unterschied zum vorausgehenden Nullmodell wird hier angenommen, dass die Anteilswerte der Variablen eine konstante Größe sind. Bezeichnen wir diese allgemein mit $p_{i(0)}$ und $p_{i(1)}$, wobei $p_{i(0)}$ der Anteil der Ausprägung 0 in der Variablen i und $p_{i(1)}$ ($= 1 - p_{i(0)}$) der Anteil der Ausprägung 1 ist, ist der erwartete Anteil an Übereinstimmungen gleich:

$$U^{\text{erw}} = \frac{1}{m} \sum_i (p_{i(0)} p_{i(0)} + p_{i(1)} p_{i(1)}).$$

Für den Übereinstimmungskoeffizienten ergibt sich daher die Berechnungsformel

$$\kappa_{g,g^*}^{(2)} = \frac{\text{SMK}_{g,g^*} - \frac{1}{m} \sum_i (p_{i(0)} p_{i(0)} + p_{i(1)} p_{i(1)})}{1 - \frac{1}{m} \sum_i (p_{i(0)} p_{i(0)} + p_{i(1)} p_{i(1)})}.$$

Im Unterschied zu $\kappa_{g,g^*}^{(1)}$ ist die erwartete Übereinstimmung eine für alle Objekte konstante Größe. Für $p_{i(0)} = 0,5$ gilt: $\kappa_{g,g^*}^{(2)} = \kappa_{g,g^*}^{(0)}$. Der nachteilige Effekt von Gleichung 8.6 tritt daher nicht auf.

In der Regel wird man sich bei einer objektorientierten Clusteranalyse für das erste Nullmodell entscheiden, sofern nicht angenommen werden kann, dass die Profilhöhen oder die Anteilswerte der Variablen konstante Größen sind. Allgemein können Werte größer 0,75 in den meisten Anwendungssituationen als ausgezeichnete überzufällige Übereinstimmung interpretiert werden. Werte zwischen 0,40 und 0,75 können als befriedigende oder gute Übereinstimmung und Werte kleiner 0,40 als unbefriedigende Übereinstimmung interpretiert werden (Fleiss 1981, S. 218).

Welches (Un-)Ähnlichkeitsmaß in einem konkreten Anwendungsfall zu verwenden ist, muss aufgrund der Datenkonstellation und der Richtung der Datenanalyse inhaltlich begründet werden. In einer variablenorientierten Analyse wird man in der Regel Φ oder einen anderen Korrelationskoeffizienten für dichotome Variablen verwenden. In einer objektorientierten besteht die primäre Entscheidung darin, welchen Informationswert Übereinstimmungen hinsichtlich des Nichtbesitzes haben. Kommt ihnen kein Informationswert zu, kann der JACCARD-I-Koeffizient verwendet werden. Besitzen sie dagegen denselben Informationswert, wird die City-Block-Metrik oder ein daraus abgeleitetes Ähnlichkeitsmaß wie der Simple-Matching-Koeffizient verwendet. Wir wollen nachfolgend anhand von zwei Beispielen darstellen, wie eine derartige Entscheidung begründet werden kann.

Beispiel 1 (Begründung für eine Gleichgewichtung des Besitzes und Nichtbesitzes von Merkmalen): In Teil 1 wurden die Freizeitaktivitäten von Kindern untersucht. Die einzelnen Freizeitaktivitäten wurden in Form einer Liste den Kindern zur Beantwortung vorgelegt. Die befragten Kinder sollten jene Aktivitäten nennen, die sie in den letzten 14 Tagen ausgeübt haben. In diesem Beispiel stellt das Ausüben einer Aktivität den »Besitz« dar, das Nichtausüben den »Nichtbesitz«. Eine unterschiedliche Gewichtung des Besitzes oder Nichtbesitzes müßte damit begründet werden, dass das Nichtausüben einer Tätigkeit einen geringeren Informationsgehalt hat als das Ausüben einer Tätigkeit. Eine derartige Begründung ist aber nur schwer möglich. Man wird sich daher für die City-Block-Metrik oder einem daraus abgeleiteten Ähnlichkeitsmaß entscheiden.

Beispiel 2 (Begründung für eine Vernachlässigung von Übereinstimmungen hinsichtlich des Nichtbesitzes): In der Untersuchung über die Lebenssituationen von Kindern wurden zusätzlich qualitative Leitfadengespräche mit den Kindern geführt und nach der bei Mayring (1988) angegebenen Methode der inhaltlichen Strukturierung verkodet. Schematisch ergibt sich dabei die in Tabelle 8.5 auf der nächsten Seite dargestellte Datenmatrix, wenn die dichotomen Variablen V_1 bis V_5 ausgewertet werden sollen. Die Verkodungen

Tab. 8.5: Datenkonstellation aus der Verkodung eines qualitativen Datenmaterials

Interview		Verkodung			
1	$V_1 = 1$	$V_2 = 1$	$V_3 = 1$	$V_4 = 1$	$V_5 = 0$
2	$V_2 = 1$	$V_4 = 1$	$V_5 = 0$		
3	$V_1 = 1$	$V_5 = 1$			
usw.					

bedeuten: Im Interview 1 treten die Variablen V_1 , V_2 , V_3 und V_4 auf, die Variable V_5 tritt dagegen nicht auf. Im Interview 2 tritt nur die Variable V_2 und V_4 auf, im Interview 3 nur die Variable V_1 und V_5 . Die Variablen können beispielsweise sein: V_1 : »Aufreten von Konflikten bezüglich des Fernsehens mit dem Vater«, V_2 : »Aufreten von Konflikten bezüglich des Fernsehens mit der Mutter«, V_3 : »Aufreten von Konflikten bezüglich der Schule mit dem Vater«, V_4 : »Aufreten von Konflikten bezüglich der Schule mit der Mutter«, V_5 : »Aufreten von Konflikten in der Schule«.

Bei dieser Datenkonstellation ist es sinnvoll, den Nichtbesitz (0-Kodierung) anders zu gewichten, da das Auftreten einer Kodierung (1-Kodierung) einen anderen Informationsgehalt besitzt². Eine 0-Kodierung kann nämlich auch dann auftreten, wenn diese Dimension im Interview nicht ausgesprochen wurde bzw. nicht auftreten konnte. So zum Beispiel können im Interview 2 die Variablen V_1 und V_3 deshalb nicht auftreten, da das Kind alleine von der Mutter erzogen wird oder sich das Interview auf Konflikte mit der Mutter konzentrierte. Auch die Dimension »Konflikte in der Schule« (V_5) kann nicht angesprochen worden sein. In diesem Beispiel ist es daher sinnvoll, Übereinstimmung hinsichtlich des Nichtbesitzes kein bzw. zumindest ein geringeres Gewicht beizumessen. Man kann sich daher für den JACCARD-I-Koeffizienten entscheiden, der Übereinstimmungen hinsichtlich des Nichtauftretens ignoriert.

Wahrscheinlichkeitsverteilungen und Signifikanztests für (Un-)Ähnlichkeitsmaße: Signifikanztests für Ähnlichkeitsmaße bzw. Unähnlichkeitsmaße sind allgemein für drei Fragestellungen bedeutsam:

1. Es kann geprüft werden, ob die Ähnlichkeit für zwei zufällig ausgewählte Objektpaare g und g^* signifikant größer der Ähnlichkeit für ein bestimmtes Nullmodell ist.
2. Für eine Clusteranalyse können Signifikanzschwellen zur Bestimmung der Clusterzahl bestimmt werden. Die Clusterzahl wird im Verschmelzungsschema dort festgelegt, wo der nachfolgende Schritt zu einer Verschmelzung von Clustern führt, deren Ähnlichkeit nicht mehr überzufällig ist. Mit »überzufällig« ist dabei gemeint,

² Auch bei einer Inhaltsanalyse, beispielsweise von Medienberichten, ist häufig eine ungleiche Gewichtung der 0-Ausprägungen im Vergleich zu den 1-Ausprägungen sinnvoll.

dass die Ähnlichkeit signifikant größer der bei einem bestimmten Nullmodell erwarteten Ähnlichkeit ist.

3. Die Erwartungswerte der Wahrscheinlichkeitsverteilungen können zur Normierung verwendet werden. Diese Technik wurde bei der Entwicklung der Übereinstimmungskoeffizienten $\kappa_{g,g^*}^{(0)}$, $\kappa_{g,g^*}^{(1)}$ und $\kappa_{g,g^*}^{(2)}$ angewendet.

Signifikanz des Φ -Koeffizienten: Für Φ kann bei Vorliegen bestimmter Kriterien (siehe zum Beispiel Siegel 1976, S. 107) die Signifikanz mit Hilfe des – aus der Tabellenanalyse bekannten – χ^2 -Tests berechnet werden, da gilt $\chi^2 = S\Phi^2$, wobei S die Fallzahl (n bei einer variablenorientierten Analyse und m bei einer objektorientierten Analyse) ist. Bei einer kleinen Fallzahl ($S < 20$) ist der χ^2 -Test nicht mehr angebracht. Eine derartige Situation ist häufig in einer objektorientierten Datenanalyse gegeben, da hier die Fallzahl gleich der Zahl der untersuchten Variablen ist. In diesem Fall kann mit dem exakten Fisher-Test bzw. mit dessen Modifikation durch Tocher gearbeitet werden (siehe zum Beispiel Siegel 1976, S. 94–111).

Für den *Simple-Matching-Koeffizienten* kann unter Verwendung des Nullmodells ohne Restriktionen die Wahrscheinlichkeitsverteilung über die Binomialverteilung bestimmt werden mit (Jain und Dubes 1988, S. 21):

$$P(a + d = \text{zufällig}) = \sum_{j=a+d}^S \binom{S}{j} \cdot 0,5^S,$$

wobei S die Fallzahl und $a + d$ die Zahl der Übereinstimmungen ist. In dem Beispiel der objektorientierten Vierfeldertafel der Tabelle 8.2 auf Seite 199 ist die Zahl der Übereinstimmungen gleich 2 und $S = 4$. Die Wahrscheinlichkeit, dass rein zufällig zwei oder mehr Übereinstimmungen auftreten, ist gleich:

$$P(2 = \text{zufällig}) = \sum_{j=2}^4 \binom{4}{j} \cdot 0,5^4 = 0,5^4(6 + 4 + 1) = 0,6875. \quad (8.7)$$

Rein zufällig treten mit einer Wahrscheinlichkeit von ungefähr 69 Prozent zwei oder mehr Übereinstimmungen auf. Dies ist die Fehlerwahrscheinlichkeit für das Verwerfen der Nullhypothese, dass die Übereinstimmungen rein zufällig auftreten. Da die Signifikanzschwelle über einem üblicherweise verwendeten Wert von 10, 5 oder 1 Prozent liegt, wird man in dem Beispiel die empirisch beobachtete Zahl von Übereinstimmungen als nicht überzufällig betrachten.

Der *JACCARD-I-Koeffizient* besitzt folgende Wahrscheinlichkeitsverteilung, wenn von dem Nullmodell ohne Restriktionen ausgegangen wird (Jain und Dubes 1988, S. 22):

$$P(\text{JACCARD I} > \ddot{\alpha}) = \sum_{x \geq \ddot{\alpha}} \sum_{i,j} \binom{S}{i} \binom{S-i}{j} \cdot 0,5^{S+i+j},$$

wobei in der zweiten Summe über jene Werte summiert wird, wo $x = j/(S - i)$ und $0 \leq i \leq S$ und $0 \leq j \leq S - i$ ist. In der ersten Summe wird über alle möglichen Werte summiert, die auftreten können, wenn der JACCARD-I-Koeffizient größer oder gleich dem empirisch beobachteten Wert \ddot{a} ($= d/(d+b+c)$) ist.

Für die anderen in der Tabelle 8.4 auf Seite 200 wiedergegebenen Koeffizienten ist die Wahrscheinlichkeitsverteilung nur teilweise bekannt (Jain und Dubes 1988, S. 23). Die Signifikanz des *Übereinstimmungskoeffizienten* κ_{g,g^*}^1 kann über einen z-Wert geprüft werden (Fleiss 1981, S. 21).

Die behandelten Wahrscheinlichkeitsverteilungen für die einzelnen Koeffizienten gehen von *unterschiedlichen Nullmodellen* aus. Diese sind, wenn wir auf die Ausführungen beim Übereinstimmungskoeffizienten κ zurückgreifen:

1. Für Φ_{g,g^*} und κ_{g,g^*}^1 wird das Nullmodell mit gegebener Profilhöhe für g und g^* angenommen.
2. Für die City-Block-Metrik und den Simple-Matching-Koeffizienten wird dagegen ein Nullmodell ohne Restriktionen verwendet.

Bei den nachfolgenden Koeffizienten werden wir auch auf das dritte bei der Entwicklung des Übereinstimmungskoeffizienten κ eingeführte Nullmodell (Nullmodell mit gegebenen Anteilswerten) zurückgreifen. Wir werden auf dieses Modell als *Nullmodell einer homogenen Population* zurückgreifen. Hier wird angenommen, dass eine homogene Verteilung ohne Clusterstruktur mit bestimmten, bekannten oder zu schätzenden Verteilungskennwerten in den einzelnen Variablen vorliegt. Bei quantitativen Variablen wird dabei als homogene Verteilung eine Gleichverteilung oder Normalverteilung angenommen werden.

Unabhängig von der Wahl der homogenen Verteilung setzen die meisten Testverfahren, die von dem Modell einer homogenen Population ausgehen, voraus, dass die Variablen unabhängig voneinander erfasst wurden. Dies ist zum Beispiel nicht der Fall, wenn Rangreihen vorliegen, da hier der Rangplatz, der einem Item zugewiesen wird, nicht mehr bei einem anderen Item auftreten kann. Kann dagegen jedes Item unabhängig von einem anderen Item beantwortet werden, ist diese Annahme erfüllt. Unabhängigkeit bedeutet also nicht, dass die untersuchten Variablen unkorreliert sind, die Variablen müssen nur unabhängig voneinander erhoben werden.

Tab. 8.6: In der Tabellenanalyse verwendete nominale Assoziationsmaße (Schmierer 1975)

Assoziationsmaß	Berechnungsformel	Eigenschaften
Kontingenzkoeffizient C	$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$	C liegt zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang). Der Wert von 1 wird aber nur bei großen Tabellen, zum Beispiel bei einer 30×30 Tabelle annähernd erreicht.
korrigierter Kontingenzkoeffizient C_{korr}	$C_{\text{korr}} = \frac{C}{\sqrt{w(w-1)}}$	C_{korr} liegt ebenfalls zwischen 0 und 1. Der nachteilige Effekt von C wird beseitigt.
Cramérs V (bzw. korrigiertes Tschuprowsche T)	$V = \sqrt{\frac{\chi^2}{n(w-1)}}$	V liegt ebenfalls zwischen 0 und 1 und ist als Verallgemeinerung von Φ definiert.

Abkürzungen: n : Zahl der Fälle, die in die Tabelle eingehen; w : Minimum der Zahl der Zeilen und Spalten

8.3 Nominale Variablen

Nominale Variablen mit mehr als zwei Ausprägungen nehmen insofern eine Sonderstellung ein, als die aus der Tabellenanalyse bekannten und auf Pearson (1904) zurückgehenden Assoziations- bzw. Korrelationsmaße (siehe Übersichtstabelle 8.6) zwar für eine *variablenorientierte Datenanalyse*, nicht aber für eine *objektorientierte Clusteranalyse* geeignet sind. Dieser Sachverhalt soll am Beispiel der (fiktiven) Datenmatrix der Tabelle 8.7 dargestellt werden. Die erste Person beispielsweise ist männlich, der Vater ist Angestellter und die Person studiert Betriebswirtschaftslehre (BWL).

Die Berechnung eines (*korrigierten*) *Kontingenzkoeffizienten* C_{korr} oder eines anderen nominalen Assoziationsmaßes für zwei Personen, wie sie für eine objektorientierte Datenanalyse erforderlich wäre, ist nicht sinnvoll, da sich unabhängig von den konkreten Ausprägungen nach Elimination der leeren Zeilen und Spalten Assoziationsmaße von 0 ergeben würden, da der χ^2 -Wert gleich 0 ist. Dieser Sachverhalt kann leicht veranschau-

Tab. 8.7: (Fiktive) Datenmatrix für nominalskalierte Variablen

Person	Geschlecht	Beruf des Vaters	Studienrichtung
1	m (1)	Angestellter (2)	BWL (1)
2	m (1)	Arbeiter (3)	BWL (1)
3	w (2)	Beamter (1)	Soziologie (2)
4	w (2)	Angestellter (2)	BWL (1)
5	w (2)	Beamter (1)	vWl (3)

Tab. 8.8: Objektorientierte Tabelle für die Personen 1 und 2 der Tabelle 8.7 auf der vorherigen Seite

		Person 2								
Person 1		m	w	Bea.	Ang.	Arb.	BWL	Soz.	vwl	Σ
m		1	0	0	0	1	1	0	0	3
w		0	0	0	0	0	0	0	0	0
Bea.		0	0	0	0	0	0	0	0	0
Ang.		1	0	0	0	1	1	0	0	3
Arb.		0	0	0	0	0	0	0	0	0
BWL		1	0	0	0	1	1	0	0	3
Soz.		0	0	0	0	0	0	0	0	0
vwl		0	0	0	0	0	0	0	0	0
Σ		3	0	0	0	3	3	0	0	9

(a) vollständige Rohtabelle

		Person 2			
Person 1		m	Arb.	BWL	Σ
m		1	1	1	3
Ang.		1	1	1	3
BWL		1	1	1	3
Σ		3	3	3	9

(b) reduzierte Tabelle ohne leere Zeilen und Spalten

licht werden, indem für die beiden ersten Personen die entsprechende objektorientierte Tabelle konstruiert wird (siehe Tabelle 8.8): Nach der Elimination der nicht besetzten Zeilen und Spalten ergibt sich eine (3×3) -Tabelle, in der alle Zellen mit einer Häufigkeit von 1 besetzt sind. Der entsprechende χ^2 -Wert ist gleich 0. Unabhängig von der konkreten Ausprägung ergäbe sich für alle Personenpaare eine Ähnlichkeit von 0, wenn zum Beispiel der Kontingenzkoeffizient C als Ähnlichkeitsmaß verwendet wird.

Für eine *objektorientierte Clusteranalyse* können dagegen folgende Maßzahlen berechnet werden:

1. Simple-Matching-Koeffizient (SMK^{nom}) für nominale Variablen,
2. κ -Koeffizient (κ^{nom}) für nominale Variablen und
3. City-Block-Metrik (CITY) bzw. quadrierte euklidische Distanz (QEUKLID).

Zur Berechnung dieser Maßzahlen werden die nominalen Variablen in ihre Dummies aufgelöst (siehe Abschnitt 7.5). Die drei genannten Koeffizienten können dann analog wie

bei den dichotomen Variablen berechnet werden. Für den Simple-Matching-Koeffizienten ist allerdings eine etwas andere Normierung der City-Block-Metrik erforderlich.

Der *Simple-Matching-Koeffizient* SMK^{nom} berechnet sich mit:

$$\text{SMK}_{g,g^*}^{\text{nom}} = 1 - \frac{\text{CITY}_{g,g^*}}{2m},$$

wobei CITY_{g,g^*} die *City-Block-Metrik* zwischen den Objekten g und g^* ist:

$$\text{CITY}_{g,g^*} = \sum_i \sum_j |x_{gi(j)} - x_{g^*i(j)}|.$$

Im Unterschied zu dichotomen Variablen ist der Wert der City-Block-Metrik in einer Variablen i im Falle einer Nichtübereinstimmung 2 und nicht 1 (siehe dazu Abschnitt 7.5). Wenn m die Zahl der Variablen ist, beträgt die maximale Distanz somit $2m$. Bei der Berechnung des Simple-Matching-Koeffizienten aus der City-Block-Metrik muss daher durch $2m$ dividiert werden. Da die Dummies dichotome Variablen sind, ist die City-Block-Metrik wiederum gleich der quadrierten euklidischen Distanz.

Der *Übereinstimmungskoeffizient* κ^{nom} für nominale Variablen (Fleiss 1981, S. 218–220) wird nach der allgemeinen Formel 8.5 auf Seite 201 berechnet mit:

$$\kappa^{\text{nom}} = \frac{U^{\text{beob}} - U^{\text{erw}}}{1 - U^{\text{erw}}},$$

wobei der Anteil der beobachteten Übereinstimmungen gleich dem Simple-Matching-Koeffizienten für nominalskalierte Variablen ist ($U^{\text{beob}} = \text{SMK}^{\text{nom}}$). Zur Berechnung der erwarteten Häufigkeiten kann das Nullmodell einer homogenen Population mit oder ohne Restriktionen (mit oder ohne vorgegebenen Anteilsraten) verwendet werden. Die erwarteten Übereinstimmungen berechnen sich mit

$$U^{\text{erw}} = \frac{1}{m} \sum_i \sum_j p_{i(j)} p_{i(j)},$$

wobei beim Nullmodell mit vorgegebenen Anteilsraten $p_{i(k)}$ der Anteil der Ausprägung k der Variablen i in der Gesamtpopulation ist. Für das Nullmodell ohne Restriktionen gilt $p_{i(k)} = 1/m_i$ (m_i : Zahl der Ausprägungen der Variablen i).

Die Berechnung der Maßzahlen soll für die fiktive Datenmatrix der Tabelle 8.7 auf Seite 207 veranschaulicht werden. Nach Durchführung der Dummy-Auflösung ergibt sich die in der Tabelle 8.9 auf der nächsten Seite dargestellte Datenmatrix. Die City-Block-Metrik für die Personen (Objekte) 1 und 2 ist gleich:

$$\text{CITY}_{1,2} = |1 - 1| + |0 - 0| + |0 - 0| + |1 - 0| + |0 - 1| + |1 - 1| + |0 - 0| + |0 - 0| = 2,$$

Tab. 8.9: In Dummies aufgelöste Datenmatrix der Tabelle 8.7 auf Seite 207

Person	Geschl.		Beruf d. V.			Studienrichtung		
	m	w	Bea.	Ang.	Arb.	BWL	Soz.	vwl
<i>g</i>	$X_{1(1)}$	$X_{1(2)}$	$X_{2(1)}$	$X_{2(2)}$	$X_{2(3)}$	$X_{3(1)}$	$X_{3(2)}$	$X_{3(3)}$
1	1	0	0	1	0	1	0	0
2	1	0	0	0	1	1	0	0
3	0	1	1	0	0	0	1	0
4	0	1	0	1	0	1	0	0
5	0	1	1	0	0	0	0	1
$p_{i(j)}$	0,4	0,6	0,4	0,4	0,2	0,6	0,2	0,2

da sich die beiden ersten Personen nur hinsichtlich des Berufs des Vaters unterscheiden. Insgesamt gingen drei Variablen ($m = 3$) in die Berechnung ein. Der Simple-Matching-Koeffizient ist daher gleich:

$$\text{SMK}_{1,2}^{\text{nom}} = 1 - \frac{\text{CITY}_{1,2}}{2m} = 1 - \frac{2}{2 \cdot 3} = 0,67 .$$

Die beiden Personen stimmen in zwei Drittel der untersuchten Variablen überein. Als Nullmodell zur Berechnung des Übereinstimmungskoeffizienten soll das Nullmodell mit vorgegebenen Anteilswerten verwendet werden. Der Anteil der zufällig erwarteten Übereinstimmungen ist gleich:

$$U^{\text{erw}} = \frac{1}{3}(0,4 \cdot 0,4 + 0,6 \cdot 0,6 + \dots + 0,2 \cdot 0,2) = 0,56 .$$

Damit ergibt sich für κ :

$$\kappa_{1,2}^{\text{nom}} = \frac{0,67 - 0,56}{1 - 0,56} = 0,25 .$$

Dieser Wert ist kleiner dem von Fleiss genannten Schwellenwert von 0,4. Man wird von keiner zufriedenstellenden (überzufälligen) Übereinstimmung sprechen.

Wahrscheinlichkeitsverteilungen und Signifikanzprüfung: Für den Simple-Matching-Koeffizienten und die City-Block-Metrik lässt sich über die verallgemeinerte Binomialverteilung ein Signifikanztest konstruieren. Auch die Testlogik von κ lässt sich leicht verallgemeinern (Fleiss 1981, S. 219–225).

8.4 Ordinale Variablen

Bei *ordinalen Variablen* können für eine variablen- oder objektorientierte Clusteranalyse zunächst die aus der Tabellenanalyse bekannten ordinalen Zusammenhangsmaße bzw. Korrelationsmaße verwendet werden (siehe Übersichtstabelle 8.10 auf der nächsten Seite). Zu ihrer Berechnung werden vier Summenwerte benötigt: Die Summe S_{ij}^+ bzw. S_{g,g^*}^+ der gleichgerichteten Beziehungen für zwei Variablen i und j bzw. zwei Objekte g und g^* , die Summe S_{ij}^- bzw. S_{g,g^*}^- der entgegengerichteten Beziehungen, die Summe S_{ii}^o bzw. $S_{g,g}^o$ der Bindungen in der Variablen i bzw. in dem Objekt g und die Summe S_{jj}^o bzw. S_{g^*,g^*}^o in der Variablen j bzw. in dem Objekt g^* . Für zwei Variablen i und j sind diese Summenwerte wie folgt definiert:

$$S_{ij}^+ = \sum_g \sum_{g^* > g} c_{ij}^+(g, g^*) ,$$

wobei gilt $c_{ij}^+(g, g^*) = 1$, wenn in den Variablen i und j für das Personenpaar (g, g^*) eine gleichgerichtete Beziehung vorliegt. Die Person g hat in beiden Variablen i und j größere (oder kleinere) Zahlenwerte. Formal ausgedrückt: $(x_{gi} > x_{g^*i}$ und $x_{gj} > x_{g^*j})$ oder $(x_{gi} < x_{g^*i}$ und $x_{gj} < x_{g^*j})$. Die Summe der entgegengerichteten Beziehungen ist analog definiert mit:

$$S_{ij}^- = \sum_g \sum_{g^* > g} c_{ij}^-(g, g^*) ,$$

wobei gilt $c_{ij}^-(g, g^*) = 1$, wenn in den Variablen i und j für das Personenpaar (g, g^*) eine entgegengerichtete Beziehung vorliegt: $(x_{gi} > x_{g^*i}$ und $x_{gj} < x_{g^*j})$ oder $(x_{gi} < x_{g^*i}$ und $x_{gj} > x_{g^*j})$.

Die Bindungssummen berechnen sich mit:

$$S_{ii}^o = \sum_g \sum_{g^* > g} c_{ii}^o(g, g^*) \quad \text{und} \quad S_{jj}^o = \sum_g \sum_{g^* > g} c_{jj}^o(g, g^*) ,$$

wobei $c_{ii}^o(g, g^*)$ bzw. $c_{jj}^o(g, g^*)$ gleich 1 ist, wenn in der Variablen i bzw. j für das Personenpaar (g, g^*) eine Bindung vorliegt, wenn also gilt: $x_{gi} = x_{g^*i}$ bzw. $x_{gj} = x_{g^*j}$.

Für zwei Personen (bzw. allgemein für zwei Objekte) lassen sich die Summenwerte analog definieren. Anstelle von Personenpaaren gehen die Variablenpaare (i, j) mit $j > i$ in die Berechnung ein. Für zwei Personen g und g^* ist beispielsweise die Summe der gleichgerichteten Beziehungen wie folgt definiert:

$$S_{g,g^*}^+ = \sum_i \sum_{j > i} c_{g,g^*}^+(i, j) ,$$

Tab. 8.10: Assoziationsmaße (Korrelationskoeffizienten) für ordinale Variablen

Assoziationsmaß	variablenorientiert	objektorientiert
Kendalls τ	$\tau_{i,j} = \frac{S_{i,j}(+) - S_{i,j}(-)}{m(m-1)/2}$	$\tau_{g,g^*} = \frac{S_{g,g^*}^+ - S_{g,g^*}^-}{n(n-1)/2}$
γ	$\gamma_{i,j} = \frac{S_{i,j}(+) - S_{i,j}(-)}{S_{i,j}(+) + S_{i,j}^-}$	$\gamma_{g,g^*} = \frac{S_{g,g^*}^+ - S_{g,g^*}^-}{S_{g,g^*}^+ + S_{g,g^*}^-}$
korrigiertes Kendalls τ (τ_b)	$\tau_{i,j} = \frac{S_{i,j}(+) - S_{i,j}(-)}{\sqrt{S_{i,i}S_{j,j}}}$ mit $S_{i,i} = S_{i,j}^+ + S_{i,j}^- + S_{i,i}^o$ und $S_{j,j} = S_{i,j}^+ + S_{i,j}^- + S_{j,j}^o$	$\tau_{g,g^*} = \frac{S_{g,g^*}^+ - S_{g,g^*}^-}{\sqrt{S_{g,g}S_{g^*,g^*}}}$ mit $S_{g,g^*} = S_{g,g^*}^+ + S_{g,g^*}^- + S_{g,g^*}^o$ und $S_{g^*,g^*} = S_{g,g^*}^+ + S_{g,g^*}^- + S_{g^*,g^*}^o$

Abkürzungen: m : Zahl der Variablen; n : Zahl der Objekte (Fälle)

wobei $c_{g,g^*}^+(i,j)$ gleich 1 ist, wenn bei den Personen g und g^* für das Variablenpaar (i,j) eine gleichgerichtete Beziehung vorliegt, wenn also gilt: $(x_{gi} > x_{gj}$ und $x_{g^*i} > x_{g^*j})$ oder $(x_{gi} < x_{gj}$ und $x_{g^*i} < x_{g^*j})$. Aus den Summenwerten lassen sich die aus der Tabellenanalyse bekannten Assoziationsmaße bzw. Korrelationskoeffizienten berechnen, die in der Tabelle 8.10 zusammengefasst sind (Goodman und Kruskal 1954, 1959, 1963, 1972; Schmierer 1975, S. 96–101; Siegel 1976, S. 203–212).

Da die Berechnung von Korrelationskoeffizienten für Objekte wegen der Dominanz der variablenorientierten Datenanalyse ungewohnt ist, soll sie hier exemplarisch für eine fiktive Datenmatrix verdeutlicht werden (siehe Tabelle 8.11). Zur Berechnung der ordinalen Assoziationsmaße für die Objekte 1 und 2 erzeugen wir uns eine neue »Datenmatrix« mit den Variablenpaaren und den Funktionswerten $c_{g=1,g^*=2}^{(\dots)}(i,j)$ (siehe Tabelle 8.12). Aus den Summenwerten können entsprechend den Berechnungsformeln der Tabelle 8.10 die ordinalen Korrelationskoeffizienten berechnet werden. Da sehr viele Bindungen vorliegen, unterscheiden sich die Zusammenhangsmaße in dem Beispiel deutlich: Ken-

Tab. 8.11: Datenmatrix zur Berechnung von ordinalen Korrelationskoeffizienten für zwei Objekte

Person (Objekt)	Items (Variablen)			
	a	b	c	d
1	2	1	2	1
2	1	1	2	1

Anmerkung: Zur Bedeutung der Items a bis d siehe Abbildung 7.3 auf Seite 190.

Tab. 8.12: Datenmatrix der Variablenpaare (i, j) für das Objektpaar $(1, 2)$

Variablenpaar (i, j)	gleichgerichtete Beziehungen $c_{1,2}^+(i, j)$	entgegenger. Beziehungen $c_{1,2}^-(i, j)$	Bindungen in Person 1 $c_{1,1}^0(i, j)$	Bindungen in Person 2 $c_{2,2}^0(i, j)$
(a, b)	0	0	0	1
(a, c)	0	0	1	0
(a, d)	0	0	0	1
(b, c)	1	0	0	0
(b, d)	0	0	1	1
(c, d)	1	0	0	0
Σ	2	0	2	3

dalls τ ist gleich 0,33, da zwei gleichgerichtete Beziehungen vorliegen und insgesamt $(4 - 1)/2 = 6$ Variablenpaare untersucht werden ($\tau = (2 - 0)/(4 \cdot (4 - 1)/2) = 0,33$). Da keine entgegengerichteten Beziehungen vorliegen und die Bindungen in γ nicht eingehen, ist $\gamma = 1$ ($\gamma = (2 - 0)/(2 + 0) = 1$). Für das korrigierte Kendalls τ_b ergibt sich ein Wert von

$$\tau_b = \frac{2 - 0}{\sqrt{(2 + 0 + 2)(2 + 0 + 3)}} = 0,41 .$$

Es treten somit deutliche Abweichungen zwischen γ und τ_b auf. Die Entscheidung für ein bestimmtes Maß hängt davon ab, welche Bedeutung Bindungen zugeschrieben wird. Bei γ wird angenommen, dass Bindungen keinen Informationswert haben. Sie werden daher eliminiert. Kendalls τ nimmt dagegen an, dass Bindungen zwar keinen Informationswert bezüglich des Zusammenhangs (gleichgerichtete Beziehungen minus entgegengerichtete Beziehungen) haben, bei der Normierung aber zu berücksichtigen sind, da die Berechnung auf allen Variablenpaaren basiert. Im Nenner steht daher die Zahl der Variablenpaare. Bei Kendalls τ_b schließlich wird der Zusammenhang so normiert, dass das Ergebnis strukturgleich dem Produkt-Moment-Korrelationskoeffizienten ist (Denz 1977).

Bei der Anwendung der ordinalen Korrelationskoeffizienten im Rahmen einer objektorientierten Datenanalyse ist wiederum zu beachten, dass wie bei allen Korrelationskoeffizienten implizit eine Standardisierung der Objekte stattfindet. Ist dieser Effekt nicht erwünscht, wird man Distanzmaße und daraus abgeleitete (Un-)Ähnlichkeitsmaße verwenden. Wir wollen hier folgende Maßzahlen darstellen:

1. City-Block-Metrik,
2. Canberra-Metrik,
3. JACCARD-II-Koeffizient,

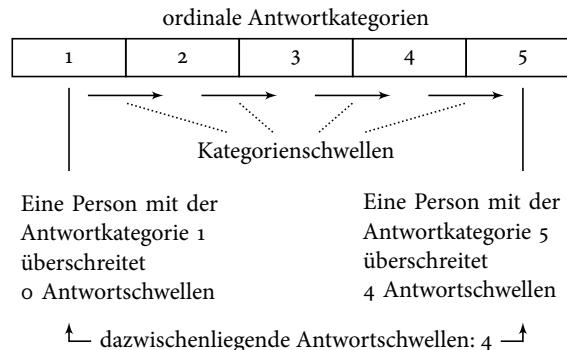


Abb. 8.1: Modellvorstellung der City-Block-Metrik für ordinale Variablen

4. verallgemeinerter Simple-Matching-Koeffizient SMK^{ord} für ordinale Variablen und
5. Übereinstimmungskoeffizient κ^{ord} für ordinale Variablen.

Die *City-Block-Metrik* berechnet sich nach der bereits bekannten Formel:

$$\text{CITY}_{g,g^*} = \sum_i |x_{gi} - x_{g^*,i}| .$$

Sind die ordinalen Variablen ganzzahlig mit 1, 2, 3 usw. kodiert, ist die City-Block-Metrik in einer Variablen gleich der Zahl der Kategorienschwellen, die zwischen den Antworten der Person g und g^* liegen. Betrachten wir dazu ein Beispiel: Es liegt eine fünfstufige Antwortskala vor. Die Person g besitzt in der Variablen i den Wert 1 (zum Beispiel »sehr wichtig«) und die Person g^* den Wert 5 (zum Beispiel »vollkommen unwichtig«). Zwischen diesen beiden Antworten liegen vier Kategorienschwellen: Die Schwelle zwischen der ersten und zweiten Antwortkategorie, die Schwelle zwischen der zweiten und dritten, jene zwischen der dritten und vierten sowie jene zwischen der vierten und fünften Kategorie. Ist die Variable mit 1, 2, 3, 4 und 5 kodiert, erhalten wir die Anzahl der zwischen g und g^* liegenden Schwellen mit Hilfe der City-Block-Metrik. Sie ist $|1 - 5| = 4$. Abbildung 8.1 verdeutlicht diesen Gedankengang. Die City-Block-Metrik besitzt somit eine ordinale Interpretation. Bei der Anwendung ist darauf zu achten, dass die Antwortkategorien fortlaufend nummeriert sind. Kodierungen der Art 1, 3, 4 und 9 sind nicht erlaubt. Die Kodierung muss aber nicht unbedingt mit 1 beginnen, da die City-Block-Metrik wie alle Distanzmaße invariant gegen Translationen (Verschiebungen um eine Konstante) ist.

Bei der *Canberra-Metrik* wird angenommen, dass absolut betrachtet gleiche Unterschiede geringer zu gewichten sind, wenn höhere Zahlenwerte vorliegen. Die Canberra-Metrik wird berechnet mit:

$$\text{CANBERRA}_{g,g^*} = \sum_i \frac{|x_{gi} - x_{g^*,i}|}{x_{gi} + x_{g^*,i}} .$$

Der Effekt, der bei der Canberra-Metrik durchgeführten Gewichtung bewirkt beispielsweise, dass Unterschiede zwischen den Kategorien 4 und 5 (zum Beispiel zwischen »lehne ab« und »lehne stark ab«) geringer gewichtet werden als Unterschiede zwischen den Kategorien 1 und 2 (zum Beispiel zwischen »stimme stark zu« und »stimme zu«). Die Verwendung der Canberra-Metrik ist somit nur zulässig, wenn inhaltlich begründet werden kann, dass größere Kodierungen weniger über Unterschiede aussagen als kleinere Kodierungen. Bei symmetrischen Einstellungsskalen (Likertskalen), die von einer starken Zustimmung (zum Beispiel 1) bis zu einer starken Ablehnung (zum Beispiel 5) gehen, ist eine derartige Begründung nicht möglich. Dagegen könnte bei Rangreihenverfahren, bei denen die Befragten ersucht werden, eine Menge von Items in eine Rangreihe zu bringen, eine derartige Gewichtung wie folgt begründet werden: Höhere Rangplätze³ sagen weniger über die Präferenz einer Person aus. Die Unterschiede sollen daher geringer gewichtet werden.

Eine implizite Gewichtung in umgekehrter Richtung findet beim *JACCARD-II-Koeffizienten* statt. Er wird wie folgt berechnet:

$$\text{JACCARD II}_{g,g^*} = \frac{\sum_i x_{gi} + \sum_i x_{g^*i} - 2 \sum_i \min(x_{gi}, x_{g^*i})}{\sum_i x_{gi} + \sum_i x_{g^*i} - \sum_i \min(x_{gi}, x_{g^*i})} \quad (8.8)$$

und ist als Verallgemeinerung des JACCARD-I-Koeffizienten definiert (siehe Formel 8.4 auf Seite 201). Er ist im Unterschied zu der City-Block-Metrik und der Canberra-Metrik ein Ähnlichkeitsmaß. Ein Wert von 1 bedeutet, dass beide Personen g und g^* dieselben Antwortprofile besitzen. Der JACCARD-II-Koeffizient nimmt an, dass Unterschiede um so größer zu gewichten sind, je größer die Summe der Profilhöhe der beiden Personen g ($\sum x_{gi}$) und g^* ($\sum x_{g^*i}$) ist. Personen mit einer größeren Profilhöhe werden daher in der Tendenz als weniger ähnlich bzw. als unähnlicher betrachtet als Personen mit geringerer Profilhöhe. Dieser Effekt wird in der Tabelle 8.13 auf der nächsten Seite dargestellt. Obwohl die City-Block-Metrik für beide Vergleichsprofile gleich 1 ist, ergibt sich für das Vergleichsprofil II mit einem Wert von 0,056 (= $1/18$) eine geringere Ähnlichkeit zu g , da beide Objekte mit einer Profilhöhe von 18 und 17 eine größere Höhe besitzen als im Vergleichsprofil I (Profilhöhen von 1 und 5). Die Canberra-Metrik führt zu einer anderen Schlussfolgerung. Sie nimmt für das Vergleichsprofil II einen Wert von 0,11 (= $1/9$) an. Dieser ist kleiner als jener von 0,33 (= $1/3$) für das Vergleichsprofil I. Vergleichsprofil II wird daher zu g als ähnlicher eingestuft als Vergleichsprofil I, da die Canberra-Metrik ein Unähnlichkeitsmaß ist und ein kleinerer Wert eine größere Ähnlichkeit ausdrückt.

³ Ein höherer Rangplatz bedeutet, dass das Item weniger wichtig ist. Das wichtigste Item hat den Rangplatz 1, das zweitwichtigste den Rangplatz 2 usw.

Tab. 8.13: Beispiel zur Berechnung der City-Block-Metrik, der Canberra-Metrik und des JACCARD-II-Koeffizienten und Veranschaulichung der impliziten Gewichtung durch diesen

	Vergleichsprofil I				\sum	Vergleichsprofil II				\sum	Maßzahl	
	x_1	x_2	x_3	x_4		x_1	x_2	x_3	x_4			
g	2	1	2	1	6	5	4	5	4	18		
g^*	1	1	2	1	5	4	4	5	4	17		
min	1	1	2	1	5	4	4	5	4	17		
$2 \cdot \text{min}$	2	2	4	1	10	8	8	10	8	34		
	$\frac{6 + 5 - 10}{6 + 5 - 5} =$				$\frac{1}{6}$	$\frac{18 + 17 - 34}{18 + 17 - 17} =$				$\frac{1}{18}$	JACCARD II	
	$ x_{g,i} - x_{g^*,i} $	1	0	0	0	1	1	0	0	0	1	CITY
	$\frac{ x_{g,i} - x_{g^*,i} }{x_{g,i} + x_{g^*,i}}$	$\frac{1}{3}$	$\frac{0}{2}$	$\frac{0}{4}$	$\frac{0}{2}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{0}{8}$	$\frac{0}{10}$	$\frac{0}{8}$	$\frac{1}{9}$	CANBERRA

Schließlich lassen sich aus der City-Block-Metrik wiederum ein *Simple-Matching-Koeffizient* SMK^{ord} und daraus ein *Übereinstimmungskoeffizient* κ^{ord} berechnen. Der Simple-Matching-Koeffizient SMK^{ord} für ordinale Variablen ist wie folgt definiert:

$$\text{SMK}_{g,g^*}^{\text{ord}} = 1 - \frac{1}{m} \sum_i \frac{|x_{gi} - x_{g^*i}|}{m_i - 1},$$

wobei m_i die Zahl der Ausprägungen der ordinalen Variablen i ist. Besitzen alle ordinalen Variablen dieselbe Zahl k von Ausprägungen, vereinfacht sich die Berechnung zu:

$$\text{SMK}_{g,g^*}^{\text{ord}} = 1 - \frac{1}{mk} \sum_i |x_{gi} - x_{g^*i}| = 1 - \frac{1}{mk} \cdot \text{CITY}_{g,g^*}. \quad (8.9)$$

Aus dem Simple-Matching-Koeffizienten kann unter Verwendung der allgemeinen Definition des Übereinstimmungskoeffizienten κ in Formel 8.5 auf Seite 201 der Übereinstimmungskoeffizient κ^{ord} für ordinale Variablen berechnet werden mit (Fleiss 1981, S. 223–225):

$$\kappa^{\text{ord}} = \frac{U^{\text{beob}} - U^{\text{erw}}}{1 - U^{\text{erw}}},$$

mit $U^{\text{beob}} = \text{SMK}^{\text{ord}}$ und

$$U^{\text{erw}} = \frac{1}{m} \sum_i \left(\underbrace{\sum_{k^*} \sum_{k^{**}} \left(1 - \frac{|k^* - k^{**}|}{m_i - 1} \right) \cdot p_{i(k^*)} p_{i(k^{**})}}_{= w_{i(k^*, k^{**})}} \right). \quad (8.10)$$

Tab. 8.14: Matrix der Gewichtungsfaktoren w_i

		k^{**}				
k^*	1	2	3	4	5	
1	1,00	0,75	0,50	0,25	0,00	
2	0,75	1,00	0,75	0,50	0,25	
3	0,50	0,75	1,00	0,75	0,50	
4	0,25	0,50	0,75	1,00	0,75	
5	0,00	0,25	0,50	0,75	1,00	

Summiert wird in jeder Variablen i über alle Ausprägungskombinationen (k^*, k^{**}) , wobei $p_{i(k^*)}$ der Anteilswert der Ausprägung k^* in der Variablen i und $p_{i(k^{**})}$ der Anteilswert der Ausprägung k^{**} in der Variablen i ist, wenn von dem Nullmodell gegebener Anteils-werte ausgegangen wird. Wird das Nullmodell ohne Restriktionen verwendet, sind die Anteilswerte $p_{i(k)} = 1/m_i$.

Die Berechnung der hier behandelten Ähnlichkeits- und Unähnlichkeitsmaße ist in der Tabelle 8.13 dargestellt. Aus der City-Block-Metrik lässt sich der Simple-Matching-Koeffizient entsprechend Formel 8.9 berechnen mit $\text{SMK}^{\text{ord}} = 1 - (1/4 \cdot 5) \cdot \text{CITY} = 1 - (1/20) \cdot 1 = 19/20$, da alle Variablen fünf Ausprägungen besitzen. Zur Berechnung des Übereinstimmungskoeffizienten κ^{ord} benötigen wir die Randverteilung der Variablen, wenn von dem Nullmodell einer homogenen Population mit gegebenen Anteilswerten ausgegangen wird. Zur Vereinfachung wird angenommen, dass alle vier Variablen dieselben Randverteilungen mit $p_{i(1)} = 0,4$, $p_{i(2)} = 0,3$ und $p_{i(3)} = p_{i(4)} = p_{i(5)} = 0,1$ besitzen. Tabelle 8.14 zeigt beispielhaft, wie die Matrix der Gewichtungsfaktoren w_i der Formel 8.10 in einer Variablen i berechnet werden. Für die Anzahl der Ausprägungen gilt $m_i = 5$. Nach Formel 8.10 beträgt der Gewichtungsfaktor für die Ausprägungskombination (2,3):

$$w_{i(2,3)} = 1 - \frac{|2-3|}{5-1} = 0,75.$$

Die Gewichte nehmen mit der Distanz $|k^* - k^{**}|$ zwischen den Ausprägungen ab. Um die erwartete Übereinstimmung für eine Variable zu erhalten, werden die Elemente der Gewichtungsmatrix mit den bei statistischer Unabhängigkeit erwarteten Anteilswerten $p_{i(k^*)}p_{i(k^{**})}$ multipliziert. Es ergeben sich die in Tabelle 8.15 auf der nächsten Seite dargestellten Werte: Bei gegebener Randverteilung ist die rein zufällig erwartete Übereinstimmung in einer Variablen gleich 0,6500. Da alle vier Variablen dieselben Randverteilungen besitzen sollen, ist U^{erw} ebenfalls gleich 0,65. Damit ergibt sich ein Übereinstimmungskoeffizient κ^{ord} von $(19/20 - 0,65)/(1 - 0,65) = 0,86$. Dieser liegt über dem von Fleiss angeführtem Schwellenwert von 0,75 für eine gute Übereinstimmung. Die beiden Merkmalsprofile stimmen also überzufällig überein.

Tab. 8.15: Matrix zur Berechnung der erwarteten Übereinstimmung U^{erw} mit Formel 8.10 auf Seite 216

	k^{**}	1	2	3	4	5	
k^*	$p_{i(k^*)}$	0,4	0,3	0,1	0,1	0,1	Σ
1	0,4	0,1600	0,0900	0,0200	0,0100	0,0000	0,2800
2	0,3	0,0900	0,0900	0,0225	0,0150	0,0075	0,2250
3	0,1	0,0200	0,0225	0,0100	0,0075	0,0050	0,0650
4	0,1	0,0100	0,0150	0,0075	0,0100	0,0075	0,0500
5	0,1	0,0000	0,0075	0,0050	0,0075	0,0100	0,0100
	Σ	0,2800	0,2250	0,0650	0,0500	0,0300	0,6500

Anmerkung: In den Zellen sind die Summanden aus Formel 8.10 auf Seite 216 dargestellt, die als Produkt der Gewichtungsfaktoren aus Tabelle 8.14 auf der vorherigen Seite, $p_{i(k^*)}$ und $p_{i(k^{**})}$ berechnet werden, zum Beispiel oben links: $1 \cdot 0,4 \cdot 0,4 = 0,16$.

Wahrscheinlichkeitsverteilungen und Signifikanztests: Signifikanztests für die ordinalen Korrelationskoeffizienten können beispielsweise in Denz (1977); Kendall (1962); Lienert (1973); Lohse, Ludwig u. a. (1982, S. 169–175) oder Siegel (1976, S. 209) nachgelesen werden. Tabellenwerte für kleine Stichprobengrößen ($n < 40$) sind unter anderem bei Kendall (1962, S. 173) und Lienert (1975, S. 237–246) angeführt. Die Signifikanzprüfung von κ^{ord} ist in Fleiss (1981, S. 224–225) dargestellt. Die Signifikanzprüfung der City-Block-Metrik wird in Abschnitt 8.5 behandelt.

Zusammenfassend können wir festhalten, dass zur Messung der Ähnlichkeit bei ordinalen Variablen die aus der Tabellenanalyse bekannten ordinalen Korrelationskoeffizienten (Kendalls τ und τ_b , Korrelationskoeffizient γ etc.) verwendet werden können. Für eine objektorientierte Clusteranalyse sind diese – wie Korrelationskoeffizienten allgemein – nur bedingt geeignet, da implizit eine Standardisierung der Objekte vorgenommen wird. In der Regel wird daher die City-Block-Metrik, ein anderes Distanzmaß oder ein daraus abgeleitetes (Un-)Ähnlichkeitsmaß verwendet. Eine ordinale Interpretation der Distanzmaße ist zulässig, sofern die ordinalen Variablen ganzzahlig ohne Sprünge kodiert sind, also beispielsweise mit 1, 2, 3 usw. Sollen die Unterschiede in Abhängigkeit von den Antworten oder der Profilhöhe gewichtet werden, kann die Canberra-Metrik oder der JACCARD-II-Koeffizient verwendet werden. Eine andere Art der Gewichtung ist durch Verwendung der quadrierten euklidischen Distanzen denkbar (siehe dazu Abschnitt 8.5).

8.5 Quantitative Variablen

Bei *quantitativen Variablen* kann zur Messung der Ähnlichkeit die *Produkt-Moment-Korrelation* verwendet werden. Wird sie für ein Variablenpaar (i, j) berechnet, wird von einer *R-Korrelation* gesprochen. Die Produkt-Moment-Korrelation zwischen zwei Objekten g und g^* wird als *Q-Korrelation* bezeichnet:

$$q_{g,g^*} = \frac{\sum_i (x_{gi} - \bar{x}_g)(x_{g^*i} - \bar{x}_{g^*})}{\sqrt{\sum_i (x_{gi} - \bar{x}_g)^2 (x_{g^*i} - \bar{x}_{g^*})^2}}. \quad (8.11)$$

In der Regel wird in einer objektorientierten Clusteranalyse jedoch wegen der nachteiligen Effekte von Korrelationsmaßen mit einem Distanzmaß gearbeitet, insbesondere werden verwendet:

1. City-Block-Metrik,
2. euklidische Distanz,
3. quadrierte euklidische Distanz und
4. Chebychev-Distanz.

Diese Distanzmaße lassen sich aus der *verallgemeinerten Minkowski-Metrik*

$$d(q,r)_{g,g^*} = \left[\sum_i |x_{gi} - x_{g^*i}|^r \right]^{1/q} \quad (8.12)$$

berechnen, indem für die Metrikparameter q und r entsprechende Werte eingesetzt werden.⁴ Diese sind:

- | | | | |
|--------------|-----|--------------|---|
| $r = 1$ | und | $q = 1$ | für die City-Block-Metrik, |
| $r = 2$ | und | $q = 2$ | für die euklidische Distanz, |
| $r = 2$ | und | $q = 1$ | für die quadrierte euklidische Distanz, |
| $r = \infty$ | und | $q = \infty$ | für die Chebychev-Distanz. |

⁴ Die verallgemeinerte Minkowski-Metrik unterscheidet sich von der bei der nichtmetrischen mehrdimensionalen Skalierung eingeführten (gewöhnlichen) Minkowski-Metrik dadurch, dass bei der verallgemeinerten Minkowski-Metrik zwei Metrikparameter q und r verwendet werden. Bei der (gewöhnlichen) Minkowski-Metrik ist dagegen $q = r$. Dies gewährleistet, dass die sogenannten Metrikeigenschaften (siehe zum Beispiel Sixtl 1982, S. 282) erfüllt sind.

Tab. 8.16: Bedeutung der Metrikparameter der verallgemeinerten Minkowski-Metrik

r	$r = q$		$q = 1$	
	g und g^*	g und g^{**}	g und g^*	g und g^{**}
1	3,00	3,00	3,00	3,00
2	1,73	3,00	3,00	9,00
3	1,44	3,00	3,00	27,00
4	1,32	3,00	3,00	108,00
:	:	:	:	:
10	1,12	3,00	3,00	59 049,00
:	:	:	:	:
∞	1,00	3,00	n. def.	n. def.

Abkürzungen: n. def.: nicht definiert bzw. unendlich

Da für $r = 2$ gilt: $|x_{gi} - x_{g^*i}|^2 = (x_{gi} - x_{g^*i})^2$, ist die Bildung des Absolutbetrages für die quadrierte euklidische Distanz und die euklidische Distanz nicht erforderlich. Wir erhalten somit folgende Berechnungsformeln:

$$\text{CITY}_{g,g^*} = d(r=1, q=1)_{g,g^*} = \sum_i |x_{gi} - x_{g^*i}| , \quad (8.13)$$

$$\text{EUKLID}_{g,g^*} = d(r=2, q=2) = \sqrt{\sum_i (x_{gi} - x_{g^*i})^2} , \quad (8.14)$$

$$\text{QEUKLID}_{g,g^*} = d(r=2, q=1) = \sum_i (x_{gi} - x_{g^*i})^2 , \quad (8.15)$$

$$\text{CHEBYCHEV}_{g,g^*} = d(r=\infty, q=\infty)_{g,g^*} = \max_i |x_{gi} - x_{g^*i}| . \quad (8.16)$$

Der Metrikparameter r führt dazu, dass die Unterschiede in den einzelnen Variablen verschieden gewichtet werden. Ein größerer Metrikparameter r bewirkt, dass größere Unterschiede in weniger Variablen stärker gewichtet werden als kleine Unterschiede in vielen Variablen. Wir wollen diesen Sachverhalt anhand eines Beispiels verdeutlichen. Es soll die Unähnlichkeit des Merkmalsprofils (2, 1, 1, 2) der Person g zu den Merkmalsprofilen (1, 2, 2, 2) und (2, 1, 1, 5) der Personen g^* und g^{**} berechnet werden. Die Personen g und g^* unterscheiden sich also durch viele kleine Abweichungen, während bei den Personen g und g^{**} nur eine große Abweichung in der vierten Variablen vorliegt. Es ergeben sich die in der Tabelle 8.16 dargestellten Werte.

Die Unähnlichkeit zwischen g und g^* nimmt mit dem Metrikparameter r ab, da den vielen kleinen Abweichungen in der Berechnung ein immer kleineres Gewicht beigemessen wird. Die Unähnlichkeit zwischen g und g^{**} bleibt konstant, da nur in einer Variablen ein großer Unterschied vorliegt. Die Objekte g und g^{**} werden mit zunehmendem Metrikparameter im Vergleich zu den Objekten g und g^* unähnlicher. Der Metrikparameter q hat

die Funktion einer Rücknormierung auf die ursprüngliche Skaleneinheit. Zur Verdeutlichung des Effekts der Rücknormierung wurde in der Tabelle auch mit $q = 1$ gerechnet. Man sieht, dass der Effekt der Gewichtung mit dem Metrikparameter r bei konstantem q immer stärker ausfällt. Während für $r = q = 10$ die Distanz zwischen g und g^* ungefähr 2,7-mal ($3,00/1,12$) so groß ist wie zwischen g und g^* , ist sie für $r = 10$ und $q = 1$ bereits 19 683-mal ($59\,049,00/3,00$) so groß. Mit Ausnahme der quadrierten euklidischen Distanz wird in der Regel r gleich q gewählt.

Wahrscheinlichkeitsverteilungen und statistische Signifikanztests: Für die Produkt-Moment-Korrelation lässt sich die Signifikanz mit Hilfe der Fisherschen z-Transformation prüfen (Fisz 1980, S. 424).⁵ Die Fishersche z-Statistik ist definiert als:

$$z_{ij} = \frac{\frac{1}{2} \log \left(\frac{1+r_{ij}}{1-r_{ij}} \right)}{\sqrt{\frac{1}{n-3}}} \quad \text{bzw.} \quad z_{g,g^*} = \frac{\frac{1}{2} \log \left(\frac{1+q_{g,g^*}}{1-q_{g,g^*}} \right)}{\sqrt{\frac{1}{m-3}}},$$

wobei r_{ij} die Produkt-Moment-Korrelation zwischen dem Variablenpaar (i,j) und q_{g,g^*} die Produkt-Moment-Korrelation zwischen dem Objektpaar (g,g^*) ist. In der Regel wird ein Schwellenwert von 2,0 bzw. 1,96 verwendet, der eine Signifikanz von ungefähr 95 Prozent besitzt.

Für die verallgemeinerte Minkowski-Metrik lässt sich allgemein keine Wahrscheinlichkeitsverteilung angeben. Dies ist aber für die euklidische Distanz und die quadrierte euklidische Distanz für das Nullmodell einer homogenen Population möglich, wobei als homogene Verteilung eine Normalverteilung in allen in die Clusterbildung einbezogenen Variablen angenommen wird. Die Variablen müssen gleiche Mittelwerte $\mu_i = \mu$ und Standardabweichungen $\sigma_i = 1$ besitzen⁶ und voneinander paarweise unabhängig sein. Unter diesen Annahmen lässt sich zeigen, dass die Größe $Q_{EUKLID}/2$ für zwei zufällig ausgewählte Personen (bzw. allgemein Objekte) g und g^* eine χ^2 -Verteilung mit m Freiheitsgraden besitzt, wobei m die Zahl der Variablen ist (Bock 1974, S. 36). Die Momente einer χ^2 -Verteilung mit m Freiheitsgraden sind (siehe zum Beispiel Fisz 1980, S. 399): $E(\chi^2) = m$ und $Var(\chi^2) = 2m$. Daraus lässt sich der Erwartungswert für die quadrierten euklidischen Distanzen berechnen mit: $E(Q_{EUKLID}) = 2m$ und $Var(Q_{EUKLID}) = 8m$.

Aufgrund dieses Befundes lässt sich nun prüfen, ob die quadrierte euklidische Distanz zwischen zwei Objekten g und g^* signifikant von 0 verschieden ist. Die berechnete quadrierte euklidische Distanz wird durch 2 dividiert und mit dem entsprechenden

⁵ Bei kleinen Stichproben ist im Zähler die Größe $\frac{r_{ij}}{2(n-1)}$ bzw. $\frac{q_{g,g^*}}{2(m-1)}$ zu addieren (Fisz 1980, S. 424). Fisz (1980, S. 421–423) gibt auch die exakte Wahrscheinlichkeitsverteilung der Produkt-Moment-Korrelation für den Fall von zwei normalverteilten Variablen an.

⁶ Sind diese Annahmen nicht erfüllt, können sie durch eine Standardisierung mit $\frac{x_i - \mu_i}{\sigma_i}$ erreicht werden.

Tab. 8.17: Erwartungswerte einiger Distanzmaße für angenommene homogene Normalverteilung

	Erwartungswert	Varianz
City-Block-Metrik ^a	$1,14m$	$0,73m^2$
euklidische Distanz	$\sqrt{2m - 1}$	1
quadrierte euklidische Distanz	$2m$	$8m$

m: Zahl der Variablen

a) Simulationswerte von Schlosser (1976, S. 126–128, 282–284)

Tabellenwert der χ^2 -Verteilung mit m Freiheitsgraden und einem vorgegebenen Signifikanzniveau verglichen. Beträgt die quadrierte euklidische Distanz zwischen zwei Objekten in vier Variablen beispielsweise 2,4, so ist diese zu einem Signifikanzniveau von 95 Prozent bzw. einem Fehlerniveau von 5 Prozent nicht signifikant von 0 verschieden, da der entsprechende kritische Schwellenwert bei vier Freiheitsgraden gleich 7,779 ist. Da die euklidische Distanz die Wurzel aus der quadrierten euklidischen Distanz ist, ist die euklidische Distanz für ($m > 30$) approximativ normalverteilt mit Erwartungswert $\sqrt{2m - 1}$ und einer Varianz von 1 (Bock 1974, S. 36; Fisz 1980, S. 401). Bei kleinerem m kann der Erwartungswert und die Varianz exakt berechnet werden (Bock 1974, S. 36; Schlosser 1976, S. 129).

Zur Bestimmung der Momente der City-Block-Metrik hat Schlosser (1976, S. 124–128) umfangreiche Simulationsstudien durchgeführt. Zu Vergleichszwecken wurde dabei auch die euklidische Distanz mituntersucht. Als H_0 -Verteilungen wurden unter anderem eine Standardnormal-, eine Gleich- und eine Dreiecksverteilung verwendet. Die Ergebnisse lassen sich wie folgt zusammenfassen (siehe Tabelle 8.17)⁷: Die Schätzung des Erwartungswertes der euklidischen Distanz mit $\sqrt{2m - 1}$ ist weitgehend unabhängig von der angenommenen homogenen Verteilung. Für den Erwartungswert der City-Block-Metrik gilt dies nur eingeschränkt. Für die Annahme einer Gleichverteilung, einer Normalverteilung oder einer Dreiecksverteilung lässt sich der Erwartungswert der City-Block-Metrik allerdings relativ gut mit $1,14m$ bestimmen.

⁷ Die Simulationswerte von Schlosser beziehen sich nur auf eine Variable. Werden m Variablen untersucht, ist daher der bei Schlosser angeführte Erwartungswert für die City-Block-Metrik von 1,14 mit m zu multiplizieren. Die in Schlosser angegebenen Standardabweichungen beziehen sich auf die Standardabweichungen der Stichprobenmittelwerte. Da wir den Fall eines Objektpaares betrachten, wurden die Standardabweichungen auf ein zufälliges Objektpaar zurückgerechnet und der häufigste Wert verwendet. Beispiel: Bei einer Stichprobengröße von $N = 20$ Objektpaaren wird in Tabelle 5.2.4 in Schlosser (1976, S. 126) eine Standardabweichung des Stichprobenmittels von 0,19 für die City-Block-Metrik angegeben. Für ein Zahlenpaar ergibt sich folglich eine Standardabweichung von $\sqrt{20 \cdot 0,19} = 0,85$ und eine Varianz von ungefähr 0,73. Bei m unabhängigen Variablen ist die Varianz somit gleich $0,73m^2$.

Die Erwartungswerte können zur Normierung der Distanzmaße verwendet werden. Beim Koeffizient r_p der Profilähnlichkeit von Cattell (1949) beispielsweise wird die quadrierte euklidische Distanz wie folgt normiert:

$$r_p = \frac{E(QEUKLID) - QEUKLID}{E(QEUKLID) + QEUKLID} = \frac{2m - QEUKLID}{2m + QEUKLID}. \quad (8.17)$$

Cattells Koeffizient der Profilähnlichkeit r_p ist ein Ähnlichkeitsmaß mit Werten zwischen -1 und $+1$. Ein Wert größer 0 bedeutet, dass die Ähnlichkeit des untersuchten Objektpaares über dem bei einer homogenen Population erwarteten Wert liegt. Vorausgesetzt werden bei der Verwendung von r_p standardisierte Variablen.

Mit den Erwartungswerten kann ferner geprüft werden ob ein zufällig ausgewähltes Objektpaar g und g^* einer homogenen Population angehört.⁸ Für die City-Block-Metrik und die euklidische Distanz lässt sich dazu ein z -Wert berechnen mit

$$z = \frac{d - E(d)}{\sigma(d)}.$$

Zur Prüfung der Signifikanz kann eine Normalverteilung angenommen werden, da die Simulationsstudien von Schlosser zeigen, dass für $m > 10$ die City-Block-Metrik und die euklidische Distanz annähernd normalverteilt sind. Bei der Berechnung des Signifikanzniveaus ist zu beachten, dass eine einseitige Fragestellung vorliegt. Es soll geprüft werden, ob d größer als der Erwartungswert ist. Wird also ein kritischer Wert von 2 verwendet, ist das entsprechende Fehlerniveau – im Unterschied zu einer zweiseitigen Fragestellung – gleich $2,5$ Prozent. Eine konservativere Teststatistik besteht in der Verwendung der *Chebychevschen Ungleichung* (Klastorin 1983; Puri und Sen 1971, S. 10–11), die keine Verteilungsannahmen trifft. Entsprechend der Chebychevschen Ungleichung kann das einseitige Fehlerniveau mit $1/z^2$ berechnet werden. Ist beispielsweise z gleich $2,17$, so ergibt sich entsprechend der Chebychevschen Ungleichung ein Fehlerniveau von $21,23$ Prozent, während bei Verwendung der Normalverteilungsapproximation ein Fehlerniveau von $1,5$ Prozent vorliegen würde. Dies zeigt, dass die Chebychevsche Ungleichheit eine sehr konservative Teststrategie ist.

⁸ Diese erfüllen immer die Voraussetzung, dass die Mittelwerte gleich sind und die Standardabweichungen einen Wert von 1 haben.

8.6 A-priori-Prüfung auf Vorhandensein einer Clusterstruktur

Bei den bisherigen Ausführungen über die Wahrscheinlichkeitsverteilungen von (Un-)Ähnlichkeitsmaßen wurde nur der Fall untersucht, dass zwei zufällig ausgewählte Objekte g und g^* vorliegen. Die diesbezüglichen Ergebnisse lassen sich nicht automatisch auf die Frage übertragen, ob alle in einer Analyse einbezogenen Objekte einer homogenen Population angehören. Werden nämlich die (Un-)Ähnlichkeitsmaße für alle Objektpaare untersucht, ist die Annahme des Vorliegens von unabhängigen Stichprobenrealisierungen verletzt, da in die Berechnung der (Un-)Ähnlichkeit zwischen g und g^* und zwischen g und g^{**} dasselbe Objekt g eingeht (Jain und Dubes 1988, S. 213). Bei einer größeren Objektmenge ($n > 50$) kann dieses Problem dadurch umgangen werden, dass nicht alle möglichen $\frac{n(n-1)}{2}$ (Un-)Ähnlichkeiten untersucht werden, sondern nur $\frac{n}{2}$ (Un-)Ähnlichkeiten, wobei jedes Objekt nur einmal in die Berechnung eingeht. Dadurch kann geprüft werden, wie stark die Verteilung der empirischen (Un-)Ähnlichkeiten von der bei einer homogenen Population erwarteten Verteilung abweicht. Aufgrund der Ausführungen in den vorausgehenden Kapiteln können unter anderem folgende Wahrscheinlichkeitsverteilungen beim Vorliegen einer homogenen Population angegeben werden:

1. *χ^2 -Verteilung für die quadrierten euklidischen Distanzen* bei quantitativen, empirisch standardisierten Variablen,
2. *Normalverteilung für die City-Block-Metrik und euklidische Distanz* bei quantitativen, empirisch standardisierten Variablen und
3. *Binomialverteilung für die City-Block-Metrik und den Simple-Matching-Koeffizient* bei dichotomen Variablen. Die Binomialverteilung kann für $n > 30$ durch eine Normalverteilung approximiert werden.

Als Testverfahren wird man bei quantitativen Variablen bzw. bei Variablen, die wie quantitative Variablen behandelt werden können, den *Kolmogorov-Smirnov-Test* (Massey Jr. 1951; Siegel 1976, S. 46–50) verwenden, da dieser eine größere Stärke besitzt als der χ^2 -Anpassungstest (Siegel 1976, S. 42–46), wenn jeder Messwert nur einmal auftritt. Bei dichotomen Variablen ist die Annahme, dass jeder Messwert nur einmal auftritt, nicht erfüllt, so dass man hier den auf Pearson (1900) zurückgehenden χ^2 -Anpassungstest verwenden wird.

Wir wollen das Vorgehen anhand der Wertedaten von Denz veranschaulichen. Es soll untersucht werden, ob bezüglich der materialistischen und postmaterialistischen Wertorientierungen (mittlere Gesamtpunktwerte für die beiden Dimensionen) eine homogene Verteilung vorliegt. Als Unähnlichkeitsmaß soll die euklidische Distanz verwendet

werden. Liegt keine Clusterstruktur vor, müssten die euklidischen Distanzen approximativ eine Normalverteilung mit dem Mittelwert $\sqrt{2m - 1} = \sqrt{2 \cdot 2 - 1} = \sqrt{3} = 1,73$ und eine Standardabweichung von 1 besitzen, da $m = 2$ Variablen analysiert wurden. Die empirische Verteilungsfunktion der euklidischen Distanzen wurde aufgrund von $n/2 = 2^{18}/2 = 109$ Distanzen berechnet. Dabei wurde wie folgt ausgewählt:

1. Die Gesamtpunktswerte wurden empirisch standardisiert, um die genannten Anwendungsvoraussetzungen (gleiche Mittelwerte der Variablen und Varianzen von 1) zu erfüllen.
2. Die Datenmatrix wurde in eine zufällige Anordnung gebracht, um Reihenfolgeeffekte durch eine systematische Dateneingabe (zuerst alle AHS-Schüler, dann alle BHS-Schüler und schließlich alle BS-Schüler) auszuschließen.
3. Anschließend wurden die euklidischen Distanzen berechnet: Die erste Distanz wurde aus den Personen 1 und 2, die zweite aus den Personen 3 und 4, die dritte aus den Personen 5 und 6 berechnet.

Das Vorgehen wurde auch mit standardnormalverteilten Zufallsvariablen gerechnet, um eine Vorstellung über die Brauchbarkeit des Testverfahrens zu gewinnen. Die Ergebnisse sind in der Tabelle 8.18 auf der nächsten Seite dargestellt. Der Tabelle ist zunächst zu entnehmen, dass bei standardnormalverteilten Zufallsdaten der Mittelwert weitgehend dem Erwartungswert von 1,73 entspricht. Auch die Standardabweichung von 0,96 entspricht der erwarteten Standardabweichung von 1,0. Für die Zufallsdaten kann die Nullhypothese des Vorliegens einer homogenen Population nicht verworfen werden. Die Fehlerquoten betragen 62,1 bzw. 26,5 Prozent. Für die empirischen Daten kann dagegen die Nullhypothese verworfen werden. Die Ergebnisse sind ein Hinweis, dass keine homogene Population vorliegt und in den Daten Cluster vorhanden sind.

Eine Reihe von weiteren Tests ist verfügbar, die eine A-priori-Prüfung erlauben, ob den Daten überhaupt eine Clusterstruktur zugrunde liegt (siehe dazu zusammenfassend Bock 1989; Jain und Dubes 1988, S. 201–221). Die Grundlogik dieser Tests besteht darin, dass geprüft wird, ob in dem von den Variablen aufgespannten Merkmalsraum Häufungen von Objekten vorliegen oder nicht. Da sich diese Verfahren nur bedingt für höherdimensionale Merkmalsräume (mehr als zwei Dimensionen) eignen, soll hier nur auf die entsprechende Literatur (Bock 1989; Jain und Dubes 1988) verwiesen werden. In Abschnitt 9.2 werden wir zeigen, wie die Ergebnisse des Single-Linkage zur Prüfung auf das Vorhandensein einer Clusterstruktur verwendet werden kann.

Allgemein ist zu dem hier dargestellten Testverfahren anzumerken, dass das Verfahren das Vorliegen von empirisch standardisierten Variablen voraussetzt. In Abschnitt 7.3 wurde ausführlich die Frage diskutiert, ob mit empirisch standardisierten Werten gerechnet werden soll oder nicht. Statistische Signifikanzüberlegungen sollten dabei eine

Tab. 8.18: Ergebnisse der A-priori-Prüfung auf Vorliegen einer Clusterstruktur in der materialistischen und postmaterialistischen Wertorientierung

Kenngrößen	empirische Daten	standard-normalverteilte Zufallsdaten	Erwartungswerte ^{a)} der euklidischen Distanz
<i>n</i>	109	109	109
Mittelwert	1,41	1,79	1,73
Standardabweichung	1,10	0,96	1,00
<i>Kolmogorov-Smirnov-Test</i>			
maximale Differenz	0,1544	0,0732	—
Fehlerwahrscheinlichkeit in %	1,1	62,1	—
<i>χ²-Anpassungstest</i>			
χ²	41,96	20,22	—
Freiheitsgrade ^{b)}	15	17	—
Fehlerwahrscheinlichkeit in %	0,1	26,5	—

a) Erwartungswerte berechnet nach Tabelle 8.17 auf Seite 222.

b) Zusammenfassungen verursachen unterschiedliche Freiheitsgrade.

untergeordnete Rolle spielen. Für die Anwendungspraxis bedeutet dies: Erscheint aus den in Abschnitt 7.3 dargestellten Gründen eine empirische Standardisierung nicht sinnvoll, sollte das hier dargestellte Testverfahren nicht eingesetzt werden.

8.7 Gewichtung von Variablen und Distanzen, Standardisierung von Objekten

In einer Clusteranalyse sind folgende Gewichtungen zu unterscheiden:

1. Die *Gewichtung von Variablen durch eine Standardisierung* bzw. allgemein durch Multiplikation mit einem Gewichtungsfaktor. Durch diese kann Nichtvergleichbarkeit beseitigt werden.
2. Die (*gleichmäßige*) *Gewichtung von Distanzen* bzw. allgemein von (Un-)Ähnlichkeitsmaßen der Distanzen in einer Variablen mit:

$$u_{g,g^*} = \frac{\sum_i w_i u(i)_{g,g^*}}{\sum_i w_i} \quad \text{bzw.} \quad \ddot{a}_{g,g^*} = \frac{\sum_i w_i \ddot{a}(i)_{g,g^*}}{\sum_i w_i},$$

wobei $u(i)_{g,g^*}$ bzw. $\ddot{a}(i)_{g,g^*}$ die (Un-)Ähnlichkeit zwischen g und g^* in der Variablen i ist und w_i das Gewicht der Variablen i . Mit $w_i = 1$ beispielsweise können mittlere (Un-)Ähnlichkeitskoeffizienten berechnet werden. Wird die City-Block-

Metrik verwendet, ist die Gewichtung der Variablen mit $w_i = 1/s_i$ oder $w_i = 1/\sigma_i$ gleich einer empirischen oder theoretischen Standardisierung.

3. Die *Gewichtung von Ausprägungen*. Sie kann bei dichotomen Variablen unter anderem durch den JACCARD-I-Koeffizienten, bei ordinalen Variablen durch den JACCARD-II-Koeffizienten und die Canberra-Metrik erreicht werden.
4. Die *Gewichtung durch die Metrikparameter r und q*.

Von diesen Gewichtungsmöglichkeiten sind wiederum Transformationen hinsichtlich der untersuchten Objekte zu unterscheiden. Bezuglich der in die Analyse einbezogenen Objekte sind folgende Entscheidungen möglich (siehe Abschnitt 7.6):

1. *Verwendung der Rohprofile der Objekte*: Die gesamte Information geht in die Berechnung ein.
2. *Verwendung der mittelwertzentrierten Profile der Objekte*: Unterschiede in der Profilhöhe werden eliminiert.
3. *Verwendung der standardisierten Profile der Objekte*: Unterschiede in der Profilhöhe und der Profilstreuung werden eliminiert.

Mittelwertzentrierte und standardisierte Profile können durch die in Abschnitt 7.6 dargestellten Transformationen berechnet werden. Soll mit standardisierten Profilen gerechnet werden, kann folgendermaßen vorgegangen werden: Die Profile werden vor der Analyse mittelwertzentriert oder standardisiert. In der anschließenden Clusteranalyse kann dann mit der verallgemeinerten Minkowski-Metrik gerechnet werden. Daneben können Ähnlichkeitsmaße verwendet werden, bei deren Berechnung implizit eine Mittelwertzentrierung oder Standardisierung der Profile vorgenommen wird. Neben dem Q-Korrelationskoeffizienten können zum Beispiel die von Zegers und Berge (1985) entwickelten Ähnlichkeitsmaße eingesetzt werden:

- *Koeffizient der Gleichheit*: Zwei Profile werden als gleich betrachtet, wenn ihre Rohprofile identisch sind, wenn also gilt: $x_{gi} = x_{g^*i}$.
- *Koeffizient der Additivität*: Zwei Profile werden als gleich betrachtet, wenn ihre mittelwertzentrierten Profile identisch sind, wenn also gilt: $x_{gi} = x_{g^*i} + a$.
- *Koeffizient der Proportionalität*: Zwei Profile werden als gleich betrachtet, wenn sie proportional zueinander sind, wenn also gilt: $x_{gi} = bx_{g^*i}$.
- *Koeffizient der Linearität*: Zwei Profile werden als gleich betrachtet, wenn ihre standardisierten Profile identisch sind, wenn also gilt: $x_{gi} = bx_{g^*i} + a$. Dies ist auch die Annahme der Q-Korrelation.

8.8 Fehlende Werte

Bisher wurde bei der Darstellung der (Un-)Ähnlichkeitsmaße das Problem *fehlender Werte* bzw. *nicht valider Angaben* ausgeklammert. Beim Vorliegen fehlender Werte kann folgendermaßen vorgegangen werden:

- *Paarweises Ausscheiden*: In die Berechnung der (Un-)Ähnlichkeit jedes Objektpaares g und g^* werden nur jene *Variablen* einbezogen, in denen g und g^* valide Werte haben.
- *Fallweises Ausscheiden*: Besitzt ein Objekt in einer oder mehreren Variablen einen fehlenden Wert, wird das *ganze Objekt* aus der Analyse eliminiert. Diese Methode kann mitunter zu einer beträchtlichen Reduktion der Fallzahl führen.
- *Schätzwerte für die fehlenden Werte*, wie zum Beispiel eine Mittelwertsubstitution: Es werden drei Variablen x_1 , x_2 und x_3 in eine Clusteranalyse einbezogen. Das Objekt g besitzt in x_3 einen fehlenden Wert. Dieser wird als Mittelwert aus x_1 und x_2 geschätzt ($x_{g3} = (x_{g1} + x_{g2})/2$). Diese Methode setzt voraus, dass die Variablen x_1 , x_2 und x_3 eine gemeinsame Dimension messen und gleich schwierig sind, das heißt, dass ihre Mittelwerte annähernd gleich sind. Wird mit empirisch standardisierten Variablen gerechnet, ist letztere Bedingung immer erfüllt, da die standardisierten Variablen einen Mittelwert von 0 haben.
- *Verwendung von Gesamtpunktwerten oder Faktorwerten* (siehe Abschnitt 5.3.1).
- *Verwendung von imputierten Werten* (Little und Rubin 2002), wobei die Clusteranalyse selbst zur Imputation eingesetzt werden kann. Es wird zunächst eine Analyse mittels fallweisen Ausscheidens gerechnet. Zur Vermeidung des Datenverlustes werden die Objekte mit fehlenden Werten dann den Clusterzentren zugeordnet.

Das paarweise Ausscheiden (Holm 1975) soll exemplarisch für eine objektorientierte Analyse dargestellt werden: Betrachten wir dazu die in der Tabelle 8.19 wiedergegebene Datenmatrix: Die Person g besitzt in der Variablen x_4 einen fehlenden Wert, die Person g^* in x_3 . Die Variablen x_3 und x_4 werden daher aus der Berechnung herausgenommen. Für die City-Block-Metrik ergibt sich in den Variablen x_1 , x_2 und x_5 ein Wert von 6. Um Vergleichbarkeit mit den Distanzen zwischen anderen Objektpaaren zu erhalten, ist eine Normierung erforderlich, andernfalls würden Objektpaare mit mehr validen Werten größere Distanzwerte erhalten. Dazu bestehen zwei Möglichkeiten: Es wird entweder mit mittleren (Un-)Ähnlichkeitsmaßen gerechnet oder die mittleren (Un-)Ähnlichkeitsmaße werden auf die ursprüngliche Variablenanzahl reskaliert. In unserem Beispiel würde sich eine mittlere City-Block-Metrik von $2^{(6/3)}$ ergeben, da drei Variablen in die Berechnung eingehen. Bei einer Reskalierung auf die ursprüngliche Variablenanzahl von fünf Variablen würde sich ein Wert von 10 ($5 \cdot 2$) ergeben.

Um Divisionen mit 0 zu vermeiden und stabile Schätzwerte für die (Un-)Ähnlichkeiten zu erhalten, ist in der Analyse die Definition eines Schwellenwertes für die Mindestzahl

Tab. 8.19: Veranschaulichung der Methode des paarweisen Ausscheidens für die City-Block-Metrik

Person (Objekt)	x_1	x_2	x_3	x_4	x_5	Σ
g	2	1	2	KW	6	
g^*	1	5	KW	1	5	
$ x_{gi} - x_{g^*i} $	1	4	—	—	1	6

Anmerkung: mittlere CITY = $6/3 = 2$; auf ursprüngliche Variablenzahl (5) reskaliert: CITY = $5 \cdot 2 = 10$

Abkürzung: KW: fehlender Wert

valider Werte eines Objektes erforderlich. Dieser Wert sollte aufgrund unserer Erfahrungen mindestens 33 Prozent sein, das heißt, jedes Objekt g sollte mehr als 33 Prozent valide Werte haben. Ist diese Bedingung nicht erfüllt, wird das entsprechende Objekt eliminiert.

Kaufman (1985) hat in Simulationsstudien die Effekte der Mittelwertsubstitution, der Methode des paarweisen Ausscheidens und des fallweisen Ausscheidens auf die Ergebnisse des Ward-Verfahrens untersucht. Die Ergebnisse zeigen, dass die Methode der Behandlung fehlender Werte nur einen geringen Einfluss auf den Anteil von Fehlklassifikationen hat. Der mittlere Anteil an Fehlklassifikationen unter den untersuchten Modellkonstellationen (1 bis 10 Prozent Messfehler und 0 bis 10 Prozent fehlende Werte) beträgt 10,79 Prozent für das fallweise Ausscheiden, 11,03 Prozent für die Methode der Mittelwertsubstitution und 13,26 Prozent für die Methode des paarweisen Ausscheidens. Wenn die vollkommen ungeeignete Methode, alle ungewichteten Hauptkomponenten zu verwenden, aus der Berechnung herausgenommen wird, beträgt der Fehler 2,97 Prozent für das fallweise Ausscheiden, 3,15 Prozent für die Mittelwertsubstitution und 3,32 Prozent für das paarweise Ausscheiden. Die Unterschiede zwischen den drei Methoden sind sehr gering, so dass – bei allen Problemen der Verallgemeinerung von Simulationsstudien – davon ausgegangen werden kann, dass in der Forschungspraxis die Methode des paarweisen Ausscheidens nicht viel schlechter ist als andere Methoden der Behandlung fehlender Werte. Die Methode des paarweisen Ausscheidens hat den Vorteil, dass in der Regel – im Unterschied zur Methode des fallweisen Ausscheidens – alle Objekte in die Analyse eingehen und dass – im Unterschied zur Methode der Mittelwertsubstitution – keine zusätzlichen Annahmen erforderlich sind.

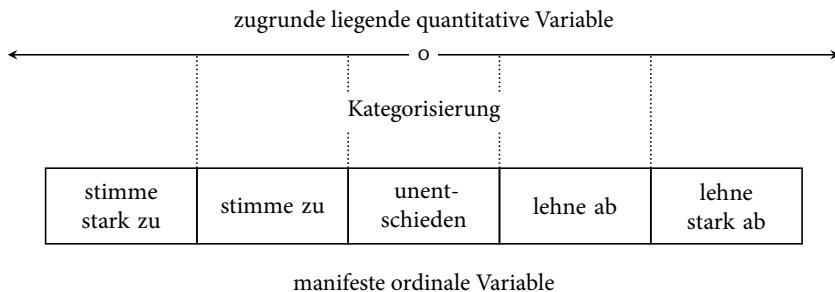


Abb. 8.2: Messtheoretische Modellvorstellung für ordinale Variablen

8.9 Exkurs: Quantifizierung und Konsequenzen der Kategorisierung

Im Rahmen der variablenorientierten Datenanalyse wurden zahlreiche Studien durchgeführt, unter welchen Bedingungen ordinale Variablen wie quantitative behandelt werden können (siehe zusammenfassend Bacher 1986 sowie Abschnitt 5.3.1). Die Studien bestätigen weitgehend die Praxis, dass ordinale und dichotome Variablen wie quantitative Variablen behandelt werden können. Diese Befunde lassen sich auch auf eine variablenorientierte Clusteranalyse übertragen. Wir wollen uns hier daher mit der Frage beschäftigen, ob in einer objektorientierten Analyse ordinale Variablen wie quantitative behandelt werden können.

Aufgrund der bisherigen Ausführungen kann diese Frage wie folgt beantwortet werden: Die formale Behandlung von ordinalen Variablen als quantitative Variablen ist zulässig, da die absoluten Abweichungen $|x_{gi} - x_{g^*i}|$, wie sie bei der verallgemeinerten Minkowski-Metrik verwendet werden, eine ordinale Interpretation besitzen, wenn die ordinalen Variablen ganzzahlig ohne Lücken kodiert sind (siehe Abschnitt 8.4). Ordinale Variablen können somit – formal betrachtet – wie quantitative Variablen behandelt werden. Von dieser formalen Möglichkeit ist die messtheoretische Frage zu unterscheiden, welche Konsequenzen die Kategorisierung einer zugrunde liegenden quantitativen Variablen in ordinale Antwortkategorien hat. Dieser Fragestellung liegt folgende Modellvorstellung zugrunde: Die untersuchten Variablen sind eigentlich quantitativ, sie können aber nur ordinal gemessen werden. Abbildung 8.2 verdeutlicht diese Annahme. Wenn wir von diesem Modell ausgehen, lautet die zu untersuchende Fragestellung: Welche Konsequenzen hat eine Kategorisierung auf die Ergebnisse einer (objektorientierten) Clusteranalyse? Wird eine auf der latenten Ebene zugrunde liegende Clusterstruktur zerstört? Wird durch die Kategorisierung eine künstliche Clusterstruktur erzeugt?

Der Fall einer homogenen, gleichverteilten Population wurde von Schlosser (1976, S. 130–134) durch Simulationsstudien ausführlich untersucht. Die Ergebnisse zeigen, dass die Kategorisierung auf die euklidische Distanz keinen Einfluss hat. Bei der City-Block-Metrik tritt dagegen eine Unterschätzung der Erwartungswerte auf. Für die Clusteranalyse bedeutet dies: Die Erwartungswerte des H_0 -Modells einer homogenen gleichverteilten Population können für die euklidischen Distanzen auch bei ordinalen Variablen angewendet werden. Für die City-Block-Metrik ist dies bei wenigen Kategorien (kleiner vier) nicht der Fall.

Der Einfluss einer Kategorisierung auf eine vorhandene Clusterstruktur kann mit den Ergebnissen von Schlosser nicht beantwortet werden. Wir haben dazu in Bacher (1996) Modellrechnungen durchgeführt, die wie folgt zusammengefasst werden können:

1. Der Effekt der Kategorisierung hängt von den Schwellenwerten ab. Sind diese gleich den Grenzen der Cluster bzw. liegen diese nahe an den Clustergrenzen, kann die Kategorisierung mitunter – bei wenigen Kategorien – sogar einen positiven Effekt haben. Die Cluster werden besser erkennbar.
2. Dieser Effekt kann bei wenigen Kategorien (bis drei bzw. vier) bei der City-Block-Metrik und der euklidischen Distanzen auftreten.
3. Bei vier und mehr Kategorien (bei der City-Block-Metrik mehr als vier Kategorien) treten nur mehr geringfügige Abweichungen im Vergleich zu den nichtkategorisierten Ergebnissen auf.

Zu beachten ist, dass durch eine Kategorisierung eine Clusterstruktur künstlich erzeugt werden kann. Wird beispielsweise eine homogene Standardnormalverteilung an den Schwellen $-0,3$ und $0,3$ trichotomisiert, entsteht eine zweigipelige Verteilung: Die erste Kategorie (Werte bis $-0,3$) hat einen Anteilswert von 38,2 Prozent, die mittlere Kategorie von 23,6 Prozent und die dritte Kategorie von 38,2 Prozent. Diese zweigipelige Verteilung wird man mitunter als einen Hinweis auf eine Clusterstruktur interpretieren. Die Gefahr, durch eine Kategorisierung eine künstliche Clusterstruktur zu erzeugen, nimmt mit der Zahl der Kategorien und der Variablenzahl ab.

9 Nächste-Nachbarn- und Mittelwertverfahren

Bei den *Nächste-Nachbarn-Verfahren* werden die Cluster so gebildet, dass jedes Klassifikationsobjekt eine bestimmte Anzahl von nächsten Nachbarn in dem Cluster, dem es angehört, hat oder dass jedes Klassifikationsobjekt in dem Cluster zumindest einen B-ten nächsten Nachbarn (zum Beispiel einen dritt nächsten Nachbarn) hat (siehe Abschnitt 6.2). Als Basismodell zur Beschreibung der Nächste-Nachbarn-Verfahren wird der *Complete-Linkage* ausgewählt (Abschnitt 9.1).

9.1 Der Complete-Linkage als Basismodell

9.1.1 Der hierarchisch-agglomerative Algorithmus

Der *Complete-Linkage* geht – wie bereits in Abschnitt 6.2 beschrieben – von einer sehr strengen Forderung hinsichtlich der Homogenität der Cluster aus: Alle Objekte eines Clusters sollen zueinander nächste Nachbarn sein. Die Bildung der Cluster erfolgt nach dem hierarchisch-agglomerativen Algorithmus, der aus folgenden Schritten besteht:

Schritt 1: Jedes Klassifikationsobjekt bildet zu Beginn ein selbständiges Cluster. Setze daher die Clusterzahl K gleich der Klassifikationsobjektzahl n .

Schritt 2: Suche das Clusterpaar $(\{p\}, \{q\})$ mit der größten Ähnlichkeit bzw. der geringsten Unähnlichkeit, verschmelze das Clusterpaar zu einem neuen Cluster $\{p, q\}$ und reduziere die Clusterzahl K um 1 ($K = K - 1$).

Schritt 3: Prüfe, ob K gleich 1 ist. Ist dies der Fall, beende den Algorithmus, da alle Klassifikationsobjekte einem einzigen Cluster angehören. Bei »nein« fahre mit Schritt 4 fort.

Schritt 4: Berechne die Ähnlichkeit bzw. Unähnlichkeit des neu gebildeten Clusters $\{p, q\}$ zu den verbleibenden Clustern i .

Schritt 5: Gehe zu Schritt 2.

Tab. 9.1: Unähnlichkeitsmatrix zwischen sieben politischen Parteien

	KP	SP	AP	LIB	ZP	CVP	KON
Kommunistische Partei (KP)	0	8,7	25,3	33,7	37,9	49,3	50,2
Sozialistische Partei (SP)	8,7	0	14,8	19,0	33,2	50,5	40,0
Arbeiterpartei (AP)	25,3	14,8	0	10,0	17,8	21,3	24,3
Liberale (LIB)	33,7	19,0	10,0	0	10,5	18,9	12,9
Zentrumspartei (ZP)	37,9	33,2	17,8	10,5	0	7,6	8,1
Christliche Volkspartei (CVP)	49,3	50,5	21,3	18,9	7,6	0	7,3
Konservative (KON)	50,2	40,0	24,3	12,9	8,1	7,3	0

Um die an die Cluster gestellte Homogenitätsforderung zu erfüllen, werden im Schritt 4 des Algorithmus die (Un-)Ähnlichkeiten des neugebildeten Clusters $\{p,q\}$ zu den verbleibenden Clustern i wie folgt berechnet:

$$u_{(p+q),i}^{\text{neu}} = \max(u_{pi}, u_{qi}) \quad \text{bzw.} \quad \ddot{a}_{(p+q),i}^{\text{neu}} = \min(\ddot{a}_{pi}, \ddot{a}_{qi}) \quad (9.1)$$

Diese Berechnung garantiert die geforderte strenge Homogenität in den Clustern des Complete-Linkage. Wir wollen das Vorgehen für die in der Tabelle 9.1 wiedergegebene Unähnlichkeitsmatrix darstellen. Sie ist einer Untersuchung von Lund (1974) über das Parteiensystem von Norwegen entnommen und in Hamerle und Pape (1984, S. 678) zitiert.

Da sieben Objekte vorliegen, wird der hierarchisch-agglomerative Algorithmus sechsmal durchlaufen. Die Schritte sind in der Abbildung 9.1 dargestellt. Vor der ersten Verschmelzung wird jedes Objekt als selbständiges Cluster betrachtet. Im ersten Verschmelzungsschritt wird entsprechend dem Schritt 2 des Algorithmus das Clusterpaar mit der geringsten Unähnlichkeit ausgewählt und verschmolzen. Die beiden Cluster mit der geringsten Unähnlichkeit (7,3) sind das Cluster {CVP} und {KON}, sie werden daher verschmolzen. Vor einer erneuten Verschmelzung werden die Unähnlichkeiten des neu gebildeten Clusters {CVP, KON} zu den verbleibenden Cluster entsprechend der Vorschrift 9.1 neu berechnet. Das von dem Objekt {KP} gebildete Cluster besitzt zu dem Cluster {CVP} eine Unähnlichkeit von 49,3 und zu dem Cluster {KON} eine Unähnlichkeit von 50,2. In die neue Distanzmatrix wird der größere Wert (50,2) eingetragen. Bei allen anderen Clustern wird analog verfahren. Es ergibt sich die in der Abbildung 9.1 dargestellte »neue« Unähnlichkeitsmatrix, wobei für das Cluster {CVP, KON} nur das erste Objekt {CVP} eingetragen ist. Für diese neue Unähnlichkeitsmatrix wird der Algorithmus erneut durchlaufen: Die Cluster {ZP} und {CVP, KON} werden verschmolzen und die Unähnlichkeiten neu berechnet. Der Verschmelzungsvorgang wird solange wiederholt, bis nur mehr ein Cluster vorliegt, und in einem Verschmelzungsschema zusammengefasst (siehe Tabelle 9.2 auf Seite 236).

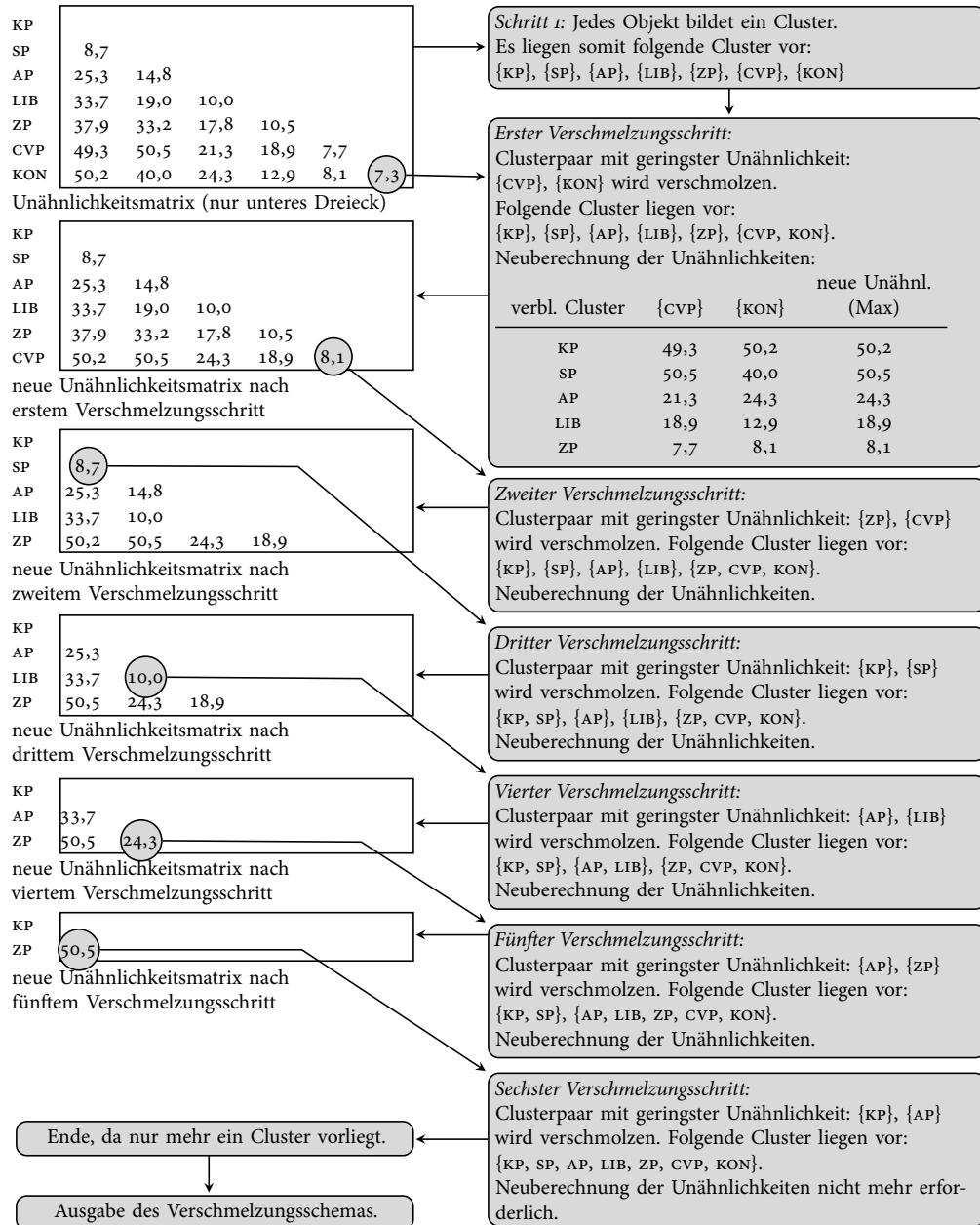


Abb. 9.1: Verschmelzungsschritte und -ergebnisse für den Complete-Linkage

Tab. 9.2: Verschmelzungsschema für das wahlsoziologische Beispiel der Tabelle 9.1 auf Seite 234 für den Complete-Linkage

Verschmelzung der Cluster mit den Objekten	Zahl der Cluster	Verschmelzungsniveau	Zuwachs
{CVP} – {KON}	6	7,3	0,0
{ZP} – {CVP, KON}	5	8,1	0,8
{KP} – {SP}	4	8,7	0,6
{AP} – {LIB}	3	10,0	1,3
{AP, LIB} – {ZP, CVP, KON}	2	24,3	14,3
{KP, SP} – {AP, LIB, ZP, CVP, KON}	1	50,5	26,2

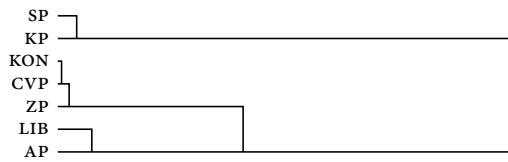
Das für eine bestimmte Clusterzahl angegebene *Verschmelzungsniveau* lässt sich wie folgt interpretieren: Alle paarweisen Unähnlichkeiten der Objekte in jedem der Cluster sind kleiner/gleich dem angegebenen Verschmelzungsniveau. Oder: Alle Objekte eines Clusters sind zum angegebenen Verschmelzungsniveau (Unähnlichkeitsniveau) zueinander nächste Nachbarn. Für die 3-Clusterlösung gilt also: Alle paarweisen Unähnlichkeiten der Objekte in jedem der drei Cluster sind kleiner/gleich 10,0. Die maximale Unähnlichkeit in Cluster 1 ist 8,7, in Cluster 2 beträgt sie 10,0 und in Cluster 3 ist sie gleich 8,1 (siehe Tabelle 9.3).

Bei den hierarchisch-agglomerativen Verfahren werden bei n Objekten $n - 1$ Clusterlösungen gebildet. Zwischen diesen besteht eine hierarchische Beziehung: Eine Clusterbildung kann in einem späteren Verschmelzungsschritt nicht mehr rückgängig gemacht werden. Die entstehende Hierarchie ist monoton: Das Verschmelzungsniveau bleibt gleich oder nimmt zu (Ausnahmen: Within-Average-Linkage, Zentroid- und Median-Verfahren). Diese Hierarchie lässt sich in Form eines sogenannten Dendrogramms darstellen (siehe Abbildung 9.2).

Tab. 9.3: Paarweise Unähnlichkeiten in den Clustern für die 3-Clusterlösung des Complete-Linkage

	Cluster 1		Cluster 2		Cluster 3		
	KP	SP	AP	LIB	ZP	CVP	KON
Kommunistische Partei (KP)	0	8,7	25,3	33,7	37,9	49,3	50,2
Sozialistische Partei (SP)	8,7	0	14,8	19,0	33,2	50,5	40,0
Arbeiterpartei (AP)	25,3	14,8	0	10,0	17,8	21,3	24,3
Liberale (LIB)	33,7	19,0	10,0	0	10,5	18,9	12,9
Zentrumspartei (ZP)	37,9	33,2	17,8	10,5	0	7,6	8,1
Christliche Volkspartei (CVP)	49,3	50,5	21,3	18,9	7,6	0	7,3
Konservative (KON)	50,2	40,0	24,3	12,9	8,1	7,3	0

grau hinterlegt: Unähnlichkeit innerhalb der Cluster



Das Dendrogramm ist wie folgt zu lesen:

1. Die Objekte {KON} und {CVP} besitzen die größte Ähnlichkeit (geringste Unähnlichkeit). Sie werden daher zunächst verschmolzen.
2. Zu diesem gebildeten Cluster {KON, CVP} wird das Objekt {ZP} hinzugefügt.
3. Als nächstes werden die Objekte {SP} und {KP} zu einem Cluster zusammengefasst.
4. Als nächstes wird aus den Objekten {LIB} und {AP} ein Cluster gebildet.
5. Die Cluster {KON, CVP, ZP} und {LIB, AP} werden verschmolzen.
6. Alle Objekte werden zu einem Cluster verschmolzen.

Abb. 9.2: Dendrogramm für den Complete-Linkage

9.1.2 Hierarchische Darstellung von Ähnlichkeitsbeziehungen

Die hierarchisch-agglomerativen Verfahren können zum Auffinden einer hierarchischen Darstellung von Ähnlichkeitsbeziehungen zwischen Variablen (variablenorientierte Analyse) oder Objekten (objektorientierte Analyse) eingesetzt werden, sofern das berechnete Verschmelzungsschema monoton ist. Die hierarchische Ähnlichkeitsbeziehung lässt sich in Form eines Dendrogramms darstellen. Zur Überprüfung, wie gut die *hierarchische Darstellung* den Daten angemessen ist, wird aus dem Verschmelzungsschema oder dem Dendrogramm eine theoretische (Un-)Ähnlichkeitsmatrix nach folgender Regel gebildet:¹ *Die theoretische Unähnlichkeit u_{g,g^*} bzw. theoretische Ähnlichkeit \ddot{a}_{g,g^*} zwischen dem Objekt g und g^* ist gleich dem Verschmelzungsniveau, zu dem die beiden Objekte zu einem Cluster verschmolzen werden.*

Für das Verschmelzungsschema der Tabelle 9.2 ergeben sich folgende Werte: Im ersten Schritt werden die Objekte {KON} und {CVP} zu einem Cluster verschmolzen. Ihre theoretische Unähnlichkeit ist daher gleich dem Verschmelzungsniveau von 7,3 ($u_{\text{CVP}, \text{KON}} = 7,3$). Im zweiten Schritt werden die Cluster {ZP} und {CVP, KON} zu einem Niveau von 8,1 verschmolzen, also die Objektpaare (<{ZP}, {CVP}) und (<{ZP}, {KON}>). Die theoretischen Unähnlichkeiten für diese Objektpaare sind gleich 8,1. Im dritten Schritt werden die Cluster {KP} und {SP}, also das Objektpaar (<{KP}, {SP}), zu einem Niveau von 8,7 verschmolzen. Die theoretische Unähnlichkeit zwischen {KP} und {SP} ist daher gleich $u_{\text{KP}, \text{SP}} = 8,7$. Geht

¹ Aus Gründen der Einfachheit wird nur die Regel für eine objektorientierte Analyse wiedergegeben. Für eine variablenorientierte Analyse ist anstelle der Indizierung (g, g^*) die Indizierung (i, j) zu verwenden.

Tab. 9.4: Theoretische Unähnlichkeitsmatrix zwischen den sieben politischen Parteien für den Complete Linkage

	KP	SP	AP	LIB	ZP	CVP	KON
Kommunistische Partei (KP)	0	8,7	50,5	50,5	50,5	50,5	50,5
Sozialistische Partei (SP)	8,7	0	50,5	50,5	50,5	50,5	50,5
Arbeiterpartei (AP)	50,5	50,5	0	10,0	24,3	24,3	24,3
Liberale (LIB)	50,5	50,5	10,0	0	10,5	24,3	24,3
Zentrumspartei (ZP)	50,5	50,5	24,3	24,3	0	8,1	8,1
Christliche Volkspartei (CVP)	50,5	50,5	24,3	24,3	8,1	0	7,3
Konservative (KON)	50,5	50,5	24,3	24,3	8,1	7,3	0

grau hinterlegt: Unähnlichkeiten innerhalb der Cluster

man alle Verschmelzungsschritte durch, ergibt sich die in der Tabelle 9.4 wiedergegebene theoretische Unähnlichkeitsmatrix, die auch als *kophenetische Matrix* bezeichnet wird.

Zur Beantwortung der Frage, wie gut die hierarchische Darstellung den Daten angemessen ist, werden die empirische und theoretische Unähnlichkeitsmatrix verglichen. Im Prinzip können dazu die im Teil I (Abschnitte 3.3.4, 4.2 und 4.5) entwickelten Maßzahlen (GFID-Index, STRESS-Koeffizient) verwendet werden. Wegen der Invarianzeigenschaft des Complete-Linkage gegenüber monotonen Transformationen ist der Einsatz eines ordinalen Zusammenhangsmaßes sinnvoll. Also zum Beispiel des γ -Korrelationskoeffizienten. Er wird nach der in Abschnitt 8.4 dargestellten Logik berechnet: Summiert wird dabei über alle Paare, die aus dem unteren Dreieck der beiden Unähnlichkeitsmatrizen gebildet werden können. Die Tabelle 9.5 veranschaulicht die Berechnungslogik. In dem Rechenschema sind die Objekte fortlaufend mit 1 beginnend nummeriert. Die empirischen Unähnlichkeiten sind der Tabelle 9.1 auf Seite 234 entnommen, jene der theoretischen Unähnlichkeit der Tabelle 9.4. In der mit » S^+ « überschriebenen Spalte stehen die gleichgerichteten Beziehungen, in » S^- « die entgegengerichteten Beziehungen. Insgesamt treten 143 gleichgerichtete Beziehungen und 6 entgegengerichtete Beziehungen auf. Für den Koeffizienten γ ergibt sich ein Wert von:

$$\gamma = \frac{143 - 6}{143 + 6} = 0,919 .$$

Es liegt eine sehr gute Modellanpassung vor.

Schwellenwerte, ab denen γ als »hoch« bezeichnet werden kann, fehlen. Hier kann man sich an den in Tabelle 3.20 auf Seite 74 angeführten Schwellenwerten für Faktorladungen orientieren und einen γ -Koeffizienten von 0,65 beispielsweise als ausreichende Modellanpassung interpretieren.

Im Hinblick auf einen *Signifikanztest für γ* schlagen Jain und Dubes (1988, S. 167–170) zur Entwicklung einer Teststatistik die Verwendung der sogenannten »random graph

Tab. 9.5: Veranschaulichung der Berechnung von γ

Objektpaare	empirische Unähnlichkeit	theoretische Unähnlichkeit	Beziehungen zw. den Objektpaaren		
			S^+	S^-	S^o
2 – 1	8,7	8,7			
3 – 1	25,3	50,5	1	0	
3 – 2	14,8	50,5	1	0	
:	:	:	:	:	:
3 – 1	25,3	50,5			
3 – 2	14,8	50,5	0	0	1
4 – 1	33,7	50,5	0	0	1
:	:	:	:	:	:
7 – 3	24,3	24,3			
7 – 4	12,9	24,3	0	0	1
7 – 5	8,1	8,1	1	0	
7 – 6	7,3	7,3	1	0	
7 – 4	12,9	24,3			
7 – 5	8,1	8,1	1	0	
7 – 6	7,3	7,3	1	0	
7 – 5	8,1	8,1			
7 – 6	7,3	7,3	1	0	
Σ			143,0	6,0	61

Abkürzungen: 1: KP; 2: SP; 3: AP; 4: LIB; 5: ZP; 6: CVP; 7: KON; siehe auch Tabelle 9.4.

hypothesis« vor. Diese geht von dem Nullmodell aus, dass die Werte in den Zellen der empirischen Unähnlichkeitsmatrix rein zufällig sind. Die Berechnung einer Verteilung von γ unter diesem Nullmodell verläuft nach folgenden Schritten:

Schritt 1: Erzeuge eine zufällige Permutation der Werte in der empirischen Unähnlichkeitsmatrix.

Schritt 2: Führe für diese zufällig permutierte Unähnlichkeitsmatrix den Complete-Linkage durch.

Schritt 3: Berechne γ für die zufällig permutierte Unähnlichkeitsmatrix.

Schritt 4: Wiederhole die Schritte 1 bis 3 eine bestimmte Anzahl Mal.

Durch dieses Vorgehen erhält man eine Wahrscheinlichkeitsverteilung für γ . Simulationsstudien (Hubert 1974) zeigen, dass die Teststatistik

$$z(\gamma) = n\gamma - 1,8 \ln(n)$$

annähernd standardnormalverteilt ist, wobei n die Zahl der untersuchten Objekte ist. Für unser Beispiel ergibt sich eine Teststatistik von $z(\gamma) = 7 \cdot 0,919 - 1,8 \ln(7) = 2,93$. Da der Wert größer 2 ist, wird man das berechnete γ zu einem Niveau von 95 Prozent ($p < 0,05$ zweiseitig, $p < 0,025$ einseitig) als signifikant betrachten. In unserem Beispiel ist die Übereinstimmung somit überzufällig. Wir können sagen, dass ein signifikanter Zusammenhang zwischen der hierarchischen Darstellung und der empirischen Ähnlichkeitsstruktur besteht.

Bei den in Abschnitt 9.5 behandelten Mittelwertverfahren wird man anstelle von γ die *kophenetische Korrelation* (KOPH) verwenden, da die Mittelwertverfahren metrische (intervallskalierte) Information voraussetzen. Beim Complete-Linkage und anderen, gegenüber monotonen Transformationen invarianten Verfahren, hat dagegen die kophenetische Korrelation die ungünstige Eigenschaft, dass sie sich nach monotonen Transformationen ändert, während die Hierarchie unverändert bleibt. Wir wollen ihre Berechnung dennoch bereits hier an dem wahlsoziologischen Beispiel aus Gründen der Vollständigkeit darstellen. Die kophenetische Korrelation ist definiert als Produkt-Moment-Korrelation zwischen der empirischen und theoretischen Unähnlichkeitsmatrix:

$$\text{KOPH}(\mathbf{U}, \tilde{\mathbf{U}}) = \frac{s(\mathbf{U}, \tilde{\mathbf{U}})}{s(\mathbf{U})s(\tilde{\mathbf{U}})}$$

mit der Kovarianz zwischen den Elementen der empirischen Unähnlichkeitsmatrix und der theoretischen Unähnlichkeitsmatrix $s(\mathbf{U}, \tilde{\mathbf{U}})$, der Standardabweichung der Elemente der empirischen Unähnlichkeitsmatrix $s(\mathbf{U})$ und der Standardabweichung der Elemente der theoretischen Unähnlichkeitsmatrix $s(\tilde{\mathbf{U}})$. In die Berechnung geht nur das untere Dreieck der empirischen und theoretischen Unähnlichkeitsmatrix ein. Die Berechnung ist in der Tabelle 9.6 dargestellt. Es ergeben sich folgende Werte:

$$\begin{aligned} s(\mathbf{U}, \tilde{\mathbf{U}}) &= \frac{1}{21} \cdot 20796,61 - \left(\frac{1}{21} \cdot 501,30 \right) \cdot \left(\frac{1}{21} \cdot 693,000 \right) = 202,56, \\ s(\mathbf{U}) &= \left[\frac{1}{21} \cdot 16342,49 - \left(\frac{1}{21} \cdot 501,30 \right)^2 \right]^{1/2} = \sqrt{208,37}, \\ s(\tilde{\mathbf{U}}) &= \left[\frac{1}{21} \cdot 29405,64 - \left(\frac{1}{21} \cdot 693,00 \right)^2 \right]^{1/2} = \sqrt{311,27}, \\ \text{KOPH}(\mathbf{U}, \tilde{\mathbf{U}}) &= \frac{202,56}{\sqrt{208,37} \cdot \sqrt{311,27}} = 0,795. \end{aligned}$$

Die kophenetische Korrelation beträgt somit 0,795. Zur Interpretation können wiederum die Schwellenwerte aus Abschnitt 3.3.4 für die Faktorladungen verwendet werden. Bei der Berechnung wurde die rechentechnische Vereinfachung der Varianz und der Kovarianz mit $s(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}$ verwendet (für $x = y$ ergibt sich die Varianz, für $x \neq y$ die Kovarianz).

Tab. 9.6: Berechnung von KOPH

Objektpaar (i,j)	empirische Unähnlichkeit $u_{i,j}$	berechnete Unähnlichkeit $\tilde{u}_{i,j}$	Hilfsgrößen zur Berechnung der kophenetischen Korrelation		
			$u_{i,j}^2$	$\tilde{u}_{i,j}^2$	$u_{i,j}\tilde{u}_{i,j}$
1 – 2	8,7	8,7	75,69	75,69	75,69
1 – 3	25,3	50,5	640,09	2 550,25	1 277,65
2 – 3	14,8	50,5	219,04	2 550,25	747,40
1 – 4	33,7	50,5	1 135,69	2 550,25	1 701,85
2 – 4	19,0	50,5	361,00	2 550,25	959,50
3 – 4	10,0	10,0	100,00	100,00	100,00
1 – 5	37,9	50,5	1 436,41	2 550,25	1 913,95
:	:	:	:	:	:
5 – 6	7,6	8,1	57,76	65,61	61,56
1 – 7	50,2	50,5	2 520,04	2 550,25	2 535,10
2 – 7	40,0	50,5	1 600,00	2 550,25	2 020,00
3 – 7	24,3	24,3	590,49	590,49	590,49
4 – 7	12,9	24,3	166,41	590,49	313,47
5 – 7	8,1	8,1	65,61	65,61	65,61
6 – 7	7,3	7,3	53,29	53,29	53,29
Σ	501,3	693,0	16 342,49	29 405,64	20 796,61

Anmerkung: zu den Objektbezeichnungen siehe Tabelle 9.5 auf Seite 239.

9.1.3 Maßzahlen zur Bestimmung der Clusterzahl

Wie bei den unvollständigen Clusteranalyseverfahren kann bei den hierarchisch-agglomerativen Verfahren ein Scree-Test zur Bestimmung der Clusterzahl durchgeführt werden. Dazu wird ein Scree-Diagramm konstruiert: Auf der X-Achse wird die Clusterzahl abgetragen, auf der Y-Achse das Verschmelzungsniveau. Im Unterschied zu den unvollständigen Clusteranalyseverfahren wird beim Lesen des Scree-Diagramms nicht mit der Clusterzahl 1 begonnen, sondern mit der höchsten Clusterzahl. Deshalb wird von einem *inversen Scree-Test* gesprochen. Im Unterschied zum Scree-Test der Faktorenanalyse ist die Clusterzahl gleich K , wenn bei K der Knickpunkt auftritt, und nicht wie bei der Faktorenanalyse gleich $K - 1$. Für das Verschmelzungsschema des Complete-Linkage ergibt sich das in der Abbildung 9.3 auf der nächsten Seite dargestellte Scree-Diagramm, wobei die Verschmelzungsniveaus auf den Zahlenbereich von 1 (erstes Verschmelzungsniveau) bis 7 (letztes Verschmelzungsniveau) normiert wurden. Die Clusterzahl wird dort festgelegt, wo ein (erster) deutlicher Knick vorliegt. Beim Complete-Linkage ist dies bei drei Clustern der Fall. Bei der Auswertung wird man daher annehmen, dass drei Cluster vorliegen. Diesen Knick erkennt man auch im Verschmelzungsschema durch eine deutliche Zunahme. Eine graphische Darstellung ist nicht unbedingt erforderlich.

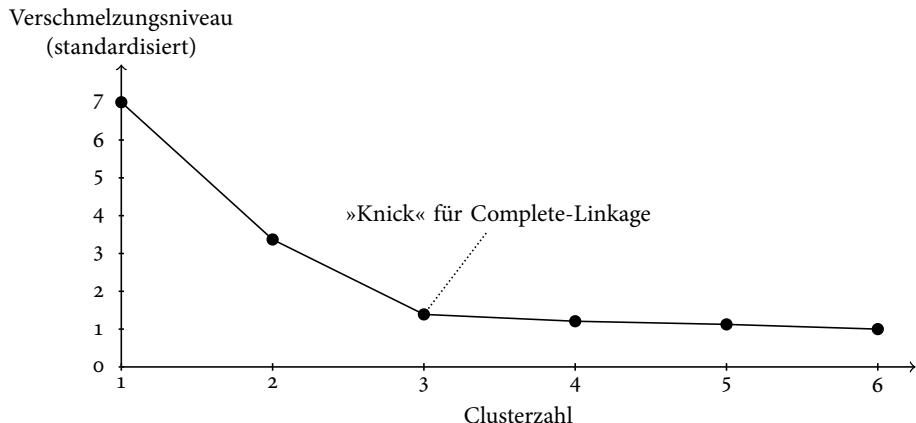


Abb. 9.3: Scree-Diagramm für den Complete-Linkage (auf das Intervall [0,7] reskalierte Werte aus Tabelle 9.2 auf Seite 236)

Eine weitere Methode ist, das Verschmelzungsschema von oben nach unten zu lesen und deutliche Zunahmen im Verschmelzungsniveau zu markieren. Die Clusterzahl beträgt dann $K + 1$, wenn die Zunahme bei K Clustern auftritt. Dieses Kriterium ist ähnlich jenem des Eigenwertabfalls (siehe Abschnitt 5.3.1). Im Unterschied zum Scree-Test wird nicht nur die höchste Clusterzahl mit einem Knick (erste Zunahme) ausgewählt, sondern mehrere Lösungen. Diese Methode wurde bereit in Abschnitt 6.6 angewendet. In dem Beispiel ergibt sich eine erste deutliche Zunahme beim Übergang von drei zu zwei Clustern, was auf eine 3-Clusterlösung hinweist, sowie beim Übergang von zwei zu einem Cluster, was auf eine 2-Clusterlösung hinweist.

Lathrop und Williams (1987, 1989, 1990) haben zahlreiche Studien über die Brauchbarkeit des inversen Scree-Tests für das Ward-Verfahren durchgeführt. Unter anderem haben sie die Verlaufsgestalt des Verschmelzungsniveaus untersucht, wenn keine Clusterstruktur vorliegt. Die Verlaufsgestalt für dieses Nullmodell ist kurvilinear und von der Größe der Cluster und der Nullverteilung der Variablen (Gleichverteilung oder Normalverteilung) relativ unabhängig. Diese Verlaufsgestalt tritt auch – wie unsere Erfahrungen mit Simulationen zeigen – bei den anderen hierarchisch-agglomerativen Verfahren auf. Die Abbildung 9.4 zeigt die typische Verlaufsgestalt des Verschmelzungsniveaus des Complete-Linkage, wenn eine Zufallsdatenmatrix mit 60 Objekten und 4 standardnormalverteilten Variablen untersucht wird. Die Verlaufskurven für die euklidische Distanz und die City-Block-Metrik sind beinahe vollkommen identisch. Bei der quadrierten euklidischen Distanz ist die Kurvilinearität stärker ausgeprägt. Ein oder mehrere Knickpunkte sind nicht zu erkennen. Zeigt das Verschmelzungsschema also keine Knickpunkte, so kann angenommen werden, dass keine Clusterstruktur den Daten zugrunde liegt.

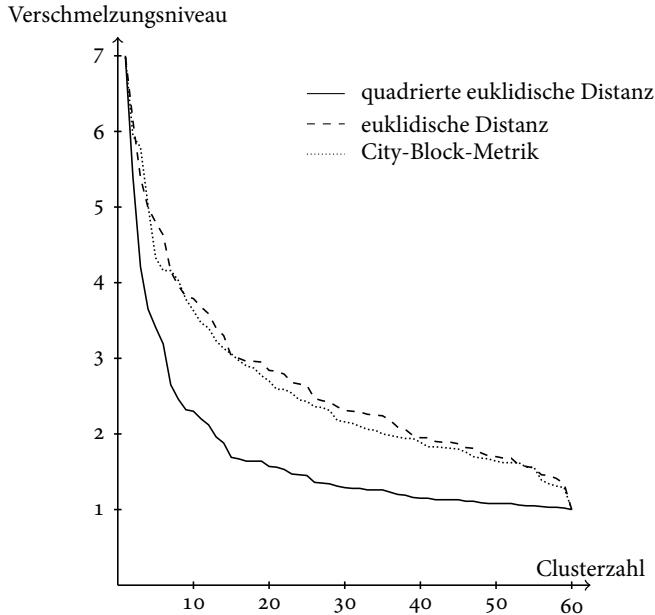


Abb. 9.4: Verlaufsgestalt des Verschmelzungsschemas für den Complete-Linkage bei Zufallsdaten

Dies kann mit dem in Abschnitt 9.1.4 beschriebenen Zufallstestverfahren untersucht werden.

Mojena (1977) hat das Vorgehen beim inversen Scree-Test formalisiert. Er geht von folgenden zwei Nullmodellen aus:

Nullmodell 1: Die Verschmelzungsniveaus ν_i ($i = 1, 2, \dots, n - 1$) sind bis zu einem bestimmten Schritt k normalverteilt mit dem Mittelwert

$$\nu_k = \frac{\sum_{i=1}^k \nu_i}{k}$$

und der Standardabweichung

$$s_k = \sqrt{\frac{1}{k-1} \sum_i (\nu_i - \bar{\nu}_k)^2}.$$

Im Schritt k wird nun geprüft, ob das Verschmelzungsniveau ν_{k+1} auf der Stufe $k + 1$ noch dieser Normalverteilung angehört. Ist dies nicht der Fall, liegt ein »signifikanter« Zuwachs des Verschmelzungsniveaus vor und die Clusterzahl wird gleich k gesetzt. Die entsprechende Teststatistik lautet:

$$\text{MOJENA-I} = \frac{\nu_{k+1} - \bar{\nu}_k}{s_k}$$

und sollte nach Mojena (1977) zwischen 2,75 und 3,50 liegen. Dies entspricht einem (einseitigen) Signifikanzniveau von mindestens 99,7 Prozent ($p < 0,003$), wenn eine Normalverteilung angenommen wird.

Nullmodell II: Die Verschmelzungsniveaus v_i ($i = 1, 2, \dots, k$) in einem Schritt k bilden eine Gerade mit

$$\hat{v}_k = a_k + b_k k ,$$

die mit Hilfe der einfachen linearen Regression geschätzt wird. Es wird geprüft, ob das $(k+1)$ -te Verschmelzungsniveau v_{k+1} noch innerhalb des aufgrund der Regressionsgerade prognostizierten Wertes $\hat{v}_{k+1} = a_k + b_k(k+1)$ und dessen Vertrauensintervalls liegt. Die entsprechende Teststatistik lautet:

$$\text{MOJENA-II} = \frac{v_{k+1} - \bar{v}_{k+1}}{s_k} .$$

Überschreitet die Teststatistik den Wert von 2,75 zum ersten Mal, liegt nach Mojena (1977) ein »signifikanter« Zuwachs vor, und die Clusterzahl sollte gleich k gesetzt werden. Der Schwellenwert von 2,75 entspricht wiederum einem (einseitigen) Signifikanzniveau von 99,7 Prozent ($p < 0,003$).

Ein alternatives Vorgehen besteht darin, dass jene Clusterlösung mit dem größten Testwert ausgewählt wird. Tritt das Maximum zum Beispiel bei $K = 5$ Cluster auf, wird diese Clusterzahl ausgewählt, auch wenn bereits vorher (zum Beispiel im zehnten Schritt) der Schwellenwert überschritten wurde. Die ist vor allem bei großen Objektmengen eine sinnvolle Strategie, da hier in einer frühen Verschmelzungsphase, also bei einer großen Clusterzahl, die Schwellenwerte überschritten werden.

In ALMO werden zusätzlich folgende Mojena-Teststatistiken berechnet:

$$\begin{aligned}\text{MOJENA-I}^* &= \frac{v_{k+1} - \bar{v}}{s} , \\ \text{MOJENA-II}^* &= \hat{v}_k = a + b k .\end{aligned}$$

Bei MOJENA-I* wird anstelle des gleitenden Mittelwertes \bar{v}_k der Gesamtmittelwert \bar{v} und anstelle der gleitenden Standardabweichung s_k die Gesamtstandardabweichung s verwendet. Bei MOJENA-II* wird die Regressionsgerade für alle Tests durch alle Verschmelzungsniveaus gelegt und nicht wie bei MOJENA-II schrittweise geschätzt.

Für das wahlsoziologische Beispiel ergeben sich die in Tabelle 9.7 dargestellten Werte: Beide Teststatistiken weisen die 3-Clusterlösung als »signifikant« aus. Das heißt, dass beim Übergang von drei zu zwei Clustern ein signifikanter Zuwachs im Verschmelzungsniveau auftritt.

Tab. 9.7: Teststatistiken von Mojena für das wahlsoziologische Beispiel

Clusterzahl	MOJENA-I		MOJENA-II	
	Teststatistik	Signifikanz	Teststatistik	Signifikanz
6	— ^{a)}	— ^{a)}	— ^{a)}	— ^{a)}
5	1,768	— ^{a)}	-0,354	— ^{a)}
4	2,800	88,649	0,807	71,707
3	13,858	99,821	11,947	99,755
2	5,450	99,470	3,938	98,632

a) nicht berechenbar

9.1.4 Zufallstestung des Verschmelzungsschemas

Die von Lathrop und Williams (1987, 1989, 1990) für das Ward-Verfahren durchgeführten Simulationsstudien können zur Entwicklung einer *Zufallstestung des Verschmelzungsschemas* für alle hierarchisch-agglomerativen Verfahren angewendet werden. Voraussetzung dafür ist, dass nicht eine direkt erhobene (Un-)Ähnlichkeitsmatrix analysiert wird, sondern eine Datenmatrix. In diesem Fall kann man wie folgt vorgehen:

Schritt 1: Berechne Mittelwerte und Standardabweichungen der untersuchten Variablen.

Schritt 2: Erzeuge eine Zufallsdatenmatrix unter der Annahme einer homogenen, normalverteilten Population mit den empirischen Verteilungskennwerten. Es wird angenommen, dass sich die Objekte in den Variablen normalverteilen mit den empirischen Mittelwerten und Standardabweichungen und voneinander unabhängig sind.

Schritt 3: Führe für die Zufallsdatenmatrix eine Clusteranalyse durch.

Schritt 4: Speichere das Verschmelzungsschema.

Schritt 5: Wiederhole die Schritte 2 bis 4 n -mal. Abhängig von der Größe der Datenmatrix wird man das Experiment 20- (bei großen Datenmatrizen), 50- (mittleren Datenmatrizen) oder 100-mal (bei kleinen Datenmatrizen) wiederholen.

Schritt 6: Berechne die Durchschnittswerte, Standardabweichungen und Teststatistiken aus den Verschmelzungsniveaus und stelle diese Werte den Verschmelzungswerten der ursprünglichen Datenmatrix gegenüber.

Schritt 7: Liegen keine deutlichen Abweichungen zwischen dem Nullmodell und dem Verschmelzungsniveau der ursprünglichen Datenmatrix vor, ist keine Clusterstruktur zu erkennen.

Anstelle einer Normalverteilung kann auch eine Gleichverteilung verwendet werden. Das Verfahren setzt voraus, dass eine Datenmatrix untersucht wird. Da dies für das wahlsoziologische Beispiel nicht der Fall ist, soll das Vorgehen am Beispiel der Entwicklungsländer-

Tab. 9.8: Empirisches Verschmelzungsschema und bei zufälligen Daten erwartetes Verschmelzungsschema für die Entwicklungsländerdaten

Schritt i	v_i	$E(v_i)$	$\sigma(v_i)$	$E(v_i) \pm 2\sigma(v_i)$
21	1,000	1,000	0,000	1,000
20	1,229	1,123	0,113	0,897
19	1,388	1,246	0,152	0,941
18	1,502	1,372	0,187	0,997
17	1,541	1,506	0,226	1,055
16	1,549	1,643	0,197	1,248
15	1,594	1,791	0,193	1,405
14	1,636	1,882	0,171	1,540
13	1,691*	2,048	0,167	1,714
12	1,695*	2,191	0,161	1,870
11	1,908*	2,417	0,200	2,018
10	2,219	2,649	0,267	2,116
9	2,351	2,824	0,257	2,310
8	2,511	2,992	0,229	2,535
7	2,768	3,234	0,249	2,735
6	2,773*	3,627	0,355	2,917
5	3,028*	3,904	0,407	3,089
4	3,665*	4,414	0,368	3,678
3	3,914*	4,943	0,484	3,976
2	4,497*	5,756	0,456	4,843
1	7,000	7,000	0,000	7,000

*: signifikante Abweichungen ($p < 0,05$)

grau hinterlegt: stärkste signifikante Abweichungen (siehe Text)

daten verdeutlicht werden (siehe Abschnitt 6.6). Der Complete-Linkage wird mit der City-Block-Metrik gerechnet. Die Ergebnisse sind in der Tabelle 9.8 dargestellt. Eingetragen wird v_i als standardisierter Wert des i -ten empirischen Verschmelzungsniveaus. Das Verschmelzungsschema wird skaliert auf das Intervall [1,7]. Daneben werden aufgenommen: $E(v_i)$ als bei zufälligen Daten erwarteter Wert für das i -te Verschmelzungsniveau, $\sigma(v_i)$ als Standardabweichung der bei zufälligen Daten erwarteten Verschmelzungsniveaus, $E(v_i) - 2\sigma(v_i)$ als untere Vertrauensschranke (95 Prozent) und $E(v_i) + 2\sigma(v_i)$ als obere Vertrauensschranke (95 Prozent). Signifikante Abweichungen sind in der Tabelle mit einem Stern (*) gekennzeichnet. Sie treten bei 13, 12, 11 sowie bei 6, 5, 4, 3 und 2 Clustern auf, wobei sie bei 12, 5 und 2 Clustern am stärksten ausgeprägt sind. Man wird daher diese Lösung als überzufällig betrachten. Die Vertrauengrenzen werden allerdings nur knapp unterschritten.

Allgemein lässt sich zu den bisher behandelten Testgrößen anmerken:

- Mögliche Clusterlösungen können mit dem Kriterium des Zuwachses im Verschmelzungsniveau gut erkannt werden.
- Nach einer bestimmten Anwendungspraxis bereitet das Erkennen von Knickpunkten im Verschmelzungsschema keine Probleme. Das gilt auch für das Erkennen von »Hügeln« im Dendrogramm.
- Besteht Unsicherheit dahingehend, ob überhaupt ein Knickpunkt vorliegt oder ein kontinuierlicher Verlauf, wird man eine Zufallstestung durchführen.
- Die Entscheidung für eine oder für mehrere Clusterlösungen wird man durch die Mojena-Kriterien absichern. Die Clusterlösungen sollten auch hohe Werte (Werte größer 2,75; Signifikanz 99,7 Prozent) in den Teststatistiken aufweisen. Bei der Verwendung von Signifikanztests ist allerdings Vorsicht angebracht. Unsere Erfahrungen zeigen, dass die Kriterien von Mojena bei großen Datensätzen oft bereits eine Zunahme bei einer sehr großen Clusterzahl (zum Beispiel 30 Cluster) als signifikant ausweisen. In dieser Situation ist es sinnvoll, als Auswahlkriterium das Maximum der Teststatistik zu verwenden.

Neben den hier behandelten Testgrößen wurden in der Literatur eine Reihe von weiteren Maßzahlen zur Bestimmung der Clusterzahl entwickelt (siehe zum Beispiel Arnold 1979; Jain und Dubes 1988, S. 177–188; Milligan 1981). Milligan (1981) untersuchte beispielsweise 30 Maßzahlen zur Bestimmung der Clusterzahl. Einige dieser Maßzahlen werden im nachfolgenden Abschnitt behandelt, da sie sich auch zur Beschreibung der Homogenität in den Clustern eignen.

9.1.5 Maßzahlen zur Beurteilung einer bestimmten Clusterlösung

Nach der Entscheidung für eine oder mehrere Clusterlösungen wird man in einem nächsten Schritt untersuchen, *in welchem Ausmaß die gebildeten Cluster die Vorstellungen der Homogenität in den Clustern und jene der Heterogenität zwischen den Clustern erfüllen*. Zur Beschreibung der Homogenität der Cluster wurden eine Reihe von Maßzahlen entwickelt, von denen hier nur folgende behandelt werden sollen:

1. Korrelationsmaße
2. Homogenitätsindizes

Tab. 9.9: Theoretische Unähnlichkeitsmatrix zwischen sieben politischen Parteien für den Complete-Linkage für die 3-Clusterlösung

	Cluster 1		Cluster 2		Cluster 3		
	KP	SP	AP	LIB	ZP	CVP	KON
Kommunistische Partei (KP)	0	0	1	1	1	1	1
Sozialistische Partei (SP)	0	0	1	1	1	1	1
Arbeiterpartei (AP)	1	1	0	0	1	1	1
Liberale (LIB)	1	1	0	0	1	1	1
Zentrumspartei (ZP)	1	1	1	1	0	0	0
Christliche Volkspartei (CVP)	1	1	1	1	0	0	0
Konservative (KON)	1	1	1	1	0	0	0

Korrelationsmaße: Für eine bestimmte Clusterlösung lässt sich nach folgender Regel eine theoretische Unähnlichkeitsmatrix bilden:

$$\hat{u}_{g,g^*} = \begin{cases} 0 & \text{wenn Objekt } g \text{ und } g^* \text{ demselben Cluster angehören,} \\ 1 & \text{wenn sie unterschiedlichen Clustern angehören.} \end{cases}$$

Für die 3-Clusterlösung des wahlsoziologischen Beispiels ergibt sich die theoretische Unähnlichkeitsmatrix in Tabelle 9.9: Die theoretische Unähnlichkeit zwischen der Kommunistischen Partei und der Sozialistischen Partei ist gleich 0, da beide Parteien demselben Cluster angehören. Die theoretische Unähnlichkeit zwischen der Kommunistischen Partei und der Arbeiterpartei ist dagegen gleich 1, da beide Parteien unterschiedlichen Clustern angehören.

Zur Messung der Übereinstimmung zwischen der theoretischen und empirischen Unähnlichkeitsmatrix kann wiederum der Koeffizient γ berechnet werden. Für ihn ergibt sich in dem Beispiel ein Wert von 1. Dies bedeutet: Alle Unähnlichkeiten innerhalb der Cluster sind kleiner als die Unähnlichkeiten zwischen den Clustern. Anstelle von γ können auch die in Kapitel 8.1 angeführten ordinalen Korrelationskoeffizienten verwendet werden. Bei den Mittelwertverfahren wird man anstelle von ordinalen Korrelationskoeffizienten den Produkt-Moment-Korrelationskoeffizienten (kophenetische Korrelation) verwenden. Für das wahlsoziologische Beispiel ergibt sich ein Wert von 0,601.

Homogenitätsindizes für eine Clusterlösung: Hier wird von der Überlegung ausgegangen, dass die Unähnlichkeiten *in* den Clustern kleiner als die Unähnlichkeiten *zwischen* den Clustern sein sollten, wenn die Cluster homogen sind. Berechnen wir für eine Clusterlösung die durchschnittliche paarweise Unähnlichkeit \bar{u}_{in} in den Clustern und die durchschnittliche paarweise Unähnlichkeit \bar{u}_{zw} zwischen den Clustern, so sollte entsprechend dieser Forderung \bar{u}_{zw} größer als \bar{u}_{in} sein. Zur Berechnung der *durchschnittlichen*

oder mittleren Unähnlichkeiten sind mehrere Ansätze möglich (Klastorin 1983). Wir wollen hier nur eine Möglichkeit darstellen, bei der die durchschnittlichen Unähnlichkeiten wie folgt berechnet werden:

$$\begin{aligned}\bar{u}_{\text{in}} &= \sum_k \bar{u}_{\text{in}}(k) \frac{1}{K}, \\ \bar{u}_{\text{zw}} &= \sum_k \sum_{k^* > k} \bar{u}_{\text{zw}}(k, k^*) \frac{2}{K(K-1)}, \\ \bar{u}_{\text{in}}(k) &= \sum_{g \in k} \sum_{g^* \in k} u_{g, g^*} \frac{2}{n_k(n_k - 1)}, \\ \bar{u}_{\text{zw}}(k, k^*) &= \sum_{g \in k} \sum_{g^* \in k^*} u_{g, g^*} \frac{1}{n_k n_{k^*}},\end{aligned}$$

wobei K die Zahl der Cluster und n_k die Größe des Clusters k ist. Bei der Berechnung wird angenommen, dass alle Cluster unabhängig von ihrer Größe dasselbe Gewicht haben sollen. Die Berechnung der Größen ist für das wahlsoziologische Beispiel in der Tabelle 9.10 auf der nächsten Seite dargestellt. In dem Beispiel ist die mittlere Unähnlichkeit in den Clustern gleich 8,8 und jene zwischen den Clustern gleich 28,1. Zur Charakterisierung der Homogenität lässt sich als Homogenitätsindex G die Differenz

$$G = \bar{u}_{\text{zw}} - \bar{u}_{\text{in}} \tag{9.2}$$

oder auch das Verhältnis

$$G = \frac{\bar{u}_{\text{in}}}{\bar{u}_{\text{zw}}}$$

verwenden. Der Vorteil der Verwendung der Differenz liegt darin, dass sich dann eine einfache Teststatistik konstruieren lässt.

Signifikanztests für die Koeffizienten: Die Signifikanz des in Gleichung 9.2 definierten Homogenitätsindex G lässt sich wie folgt prüfen: Unter der Annahme, dass die Clustergrößen n_k konstante Größen sind und die gefundene Clusterlösung rein zufällig ist, lässt sich der Erwartungswert und die Standardabweichung berechnen (Klastorin 1983, S. 95)². Aus dem berechneten Erwartungswert und der Varianz lässt sich eine z-Teststatistik konstruieren mit $z = (G - E(G))/\sigma(G)$. Für diese kann geprüft werden, ob sie signifikant größer 0 ist, wenn eine Standardnormalverteilung angenommen wird. Ist z größer 2, kann die Nullhypothese des Vorliegens einer zufälligen Partition zu einem Fehlerniveau von 2,5 Prozent verworfen werden, da eine einseitige Fragestellung ($G > E(G)$)

² Die von Klastorin (1983) angeführte Formel zur Varianzberechnung enthält einen Fehler. Die richtige Formel kann in Hubert und Levin (1977) nachgelesen werden.

Tab. 9.10: Berechnung der mittleren paarweisen Unähnlichkeiten in und zwischen den Clustern

Unähnlichkeiten innerhalb der Cluster			Unähnlichkeiten zwischen den Clustern		
<i>Cluster 1</i>	KP – SP $\bar{u}_{in}(1) =$	8,7 8,7	<i>Cluster 1 – Cluster 2</i>	KP – AP KP – LIB	25,3 33,7
<i>Cluster 2</i>	AP – LIB $\bar{u}_{in}(2) =$	10,0 10,0		SP – AP SP – LIB	14,8 19,0
<i>Cluster 3</i>	ZP – CVP ZP – KON CVP – KON $\bar{u}_{in}(3) = 23/3 =$	7,6 8,1 7,3 7,7		$\bar{u}_{zw}(1,2) = 92,8/4 =$ <i>Cluster 1 – Cluster 3</i>	23,2 37,9 49,3 50,2
	$\bar{u}_{in} = (8,7 + 10,0 + 7,7)/3 =$	8,8		SP – ZP SP – CVP SP – KON	33,2 50,5 40,0
				$\bar{u}_{zw}(1,3) = 261,1/6 =$ <i>Cluster 2 – Cluster 3</i>	43,5 17,8
				AP – ZP AP – CVP AP – KON	21,3 24,3
				LIB – ZP LIB – CVP LIB – KON	10,5 18,9 12,9
				$\bar{u}_{zw}(2,3) = 105,7/6 =$	17,6
				$\bar{u}_{zw} = (23,2 + 43,5 + 17,6)/3 =$	28,1

untersucht wird. Für das wahlsoziologische Beispiel ergibt sich für G ein Wert von 19,322. Der Erwartungswert ist 0 und die Varianz gleich 65,973. Die z-Teststatistik ist somit gleich 2,379 ($= (19,322 - 0) / \sqrt{65,973}$) und zu einem Niveau von 99,1 Prozent ($p < 0,01$ einseitig) von 0 verschieden. Unter Verwendung der konservativen Teststatistik von Chebychev (Chebychevsche Ungleichung) ergibt sich allerdings ein über 10 Prozent liegendes Fehlerniveau (einseitig) von 11,7 Prozent ($= \frac{100}{z^2} = \frac{100}{2,379^2}$).

Eine andere Möglichkeit der Bestimmung von Signifikanzschwellen ist die Durchführung von Simulationsrechnungen. Dabei kann von folgendem Nullmodell ausgegangen werden: Jede andere K -Clusterlösung liefert gleich gute Ergebnisse. Ist dies der Fall, wird man die gefundene Clusterlösung als zufällig betrachten. Das Vorgehen besteht aus folgenden Schritten:

Schritt 1: Ordne jedes Objekt zufällig einem der K Cluster zu.

Schritt 2: Berechne die entsprechende Maßzahl.

Schritt 3: Wiederhole die Schritte 1 und 2 r -mal (zum Beispiel $r = 100$).

Das Ergebnis der Simulationsrechnungen besteht darin, dass eine Wahrscheinlichkeitsverteilung für die verwendete Maßzahl berechnet wird. Aus dieser können zum einen

Tab. 9.11: Ergebnisse der Simulationsprüfung der 3-Clusterlösung (100 Simulationen)

	γ	KOPH
empirischer Wert	1,00	0,60
Erwartungswert	-0,01	-0,01
Standardabweichung	0,28	0,21
z-Teststatistik	3,61	2,90
Schwellenwert für eine Signifikanz von 90 Prozent	0,45	0,37
Fehlerniveau für die Chebychevsche Ungleichung in Prozent	7,67	11,90

Vertrauensintervalle und zum anderen eine z-Teststatistik berechnet werden. Unsere Erfahrungen zeigen, dass zur Berechnung der z-Teststatistik eine Zahl von 20 Simulationen ausreicht. Zur Berechnung von Signifikanzschwellen ist eine größere Zahl von Simulationen erforderlich. Die Zahl hängt hier von dem gewünschten Signifikanzniveau ab. Bei 100 Simulationen berechnet das Programmsystem ALMO die einseitige Signifikanzschwelle für 90 Prozent, bei 200 Simulationen für 95 Prozent usw.

Die Ergebnisse für das wahlsoziologische Beispiel sind in der Tabelle 9.11 wiedergegeben: Bei 100 Simulationen ergibt sich ein Erwartungswert von -0,01 für beide Maßzahlen. Die Standardabweichungen betragen 0,28 (γ -Koeffizient) und 0,21 (kophenetische Korrelation). Die z-Teststatistik ist folglich für γ gleich $(1 - (-0,01)) / 0,28 = 3,61$. Wird ein Schwellenwert von 2 für die z-Teststatistik angenommen ($p < 0,025$ für eine einseitige Fragestellung), wird man den berechneten z-Wert als signifikant und die gefundene Clusterlösung als überzufällig betrachten. Das heißt, dass eine Clusterstruktur in den Daten vorliegt. Der aus den Simulationen berechnete Schwellenwert für eine Signifikanz von 90 Prozent ($p < 0,10$ einseitig) ist für γ gleich 0,45. Der empirisch berechnete Wert liegt deutlich über diesem Wert. Auch das mit Hilfe der Chebychevschen Ungleichung berechnete Fehlerniveau liegt unter 10 Prozent (einseitig). Gleichermaßen gilt für die kophenetische Korrelation, wobei das Fehlerniveau der konservativen Chebychevschen Ungleichung knapp über 10 Prozent liegt. In dem Beispiel wird man daher die 3-Clusterlösung als überzufällig betrachten, da beim Complete-Linkage die kophenetische Korrelation nur bedingt zur Beurteilung der Modellanpassung geeignet ist.

9.2 Der Single-Linkage

Während der Complete-Linkage von einer sehr strengen Forderung hinsichtlich der Homogenität der Cluster ausgeht, stellt der *Single-Linkage* nur die schwache Anforderung, dass jedes Objekt mindestens einen nächsten Nachbarn im Cluster haben muss. Diese

Tab. 9.12: Verschmelzungsschema für das wahlsoziologische Beispiel der Tabelle 9.1 auf Seite 234 für den Single-Linkage

Verschmelzung der Cluster mit den Objekten	Zahl der Cluster	Verschmelzungsniveau	Zuwachs
{CVP} – {KON}	6	7,3	0,0
{ZP} – {CVP, KON}	5	7,7	0,3
{KP} – {SP}	4	8,7	1,1
{LIB} – {AP}	3	10,0	1,3
{AP, LIB} – {ZP, CVP, KON}	2	10,5	0,5
{KP, SP} – {AP, LIB, ZP, CVP, KON}	1	14,8	4,3

Forderung wird dadurch erreicht, dass im vierten Schritt des Algorithmus der hierarchisch-agglomerativen Verfahren (siehe Abschnitt 9.1.1) die neuen Unähnlichkeiten wie folgt berechnet werden:

$$u_{(p+q),i}^{\text{neu}} = \min(u_{pi}, u_{qi}) \quad \text{bzw.} \quad \ddot{a}_{(p+q),i}^{\text{neu}} = \max(\ddot{a}_{pi}, \ddot{a}_{qi})$$

Tabelle 9.12 zeigt das Verschmelzungsschema für die wahlsoziologischen Daten: Die Verschmelzungsniveaus sind wie folgt zu interpretieren: Jedes Objekt hat in dem Cluster, dem es angehört, mindestens einen nächsten Nachbarn mit einer Unähnlichkeit kleiner/gleich dem angegebenen Verschmelzungsniveau von v_i . Betrachten wir beispielsweise die 2-Clusterlösung: Die beiden Clusterlösungen sind: $C_1 = \{KP, SP\}$ und $\{AP, LIB, ZP, CVP, KON\}$. Für die 2-Clusterlösung gilt: Innerhalb jedes Clusters hat jedes Objekt mindestens einen nächsten Nachbarn mit einem Unähnlichkeitsniveau von kleiner/gleich 10,5. Von dieser Tatsache kann man sich leicht überzeugen, wenn wir alle paarweisen Unähnlichkeiten zwischen den Objekten innerhalb der Cluster betrachten und auszählen, wie viele nächste Nachbarn jedes Objekt zum Niveau von 10,5 hat. Es ergibt sich das in der Tabelle 9.13 dargestellte Bild. Jedes Objekt hat mindestens einen nächsten Nachbarn. Betrachten wir Cluster 2: Objekt {AP} hat zum Niveau 10,5 einen nächsten Nachbarn, nämlich {LIB}. Objekt {LIB} hat zwei nächste Nachbarn, nämlich die Objekte {AP} und {ZP} usw.

Wenn wir das Verschmelzungsschema des Single-Linkage (Tabelle 9.13) mit jenem des Complete-Linkage (Tabelle 9.2 auf Seite 236) vergleichen, so tritt erst beim Übergang von zwei zu einem Cluster eine deutliche Zunahme im Verschmelzungsniveau auf. Beim Complete-Linkage war dies dagegen beim Übergang von drei zu zwei Clustern der Fall. Die Ursache hierfür ist der bereits erwähnte *Verkettungseffekt* des Single-Linkage (siehe Abschnitt 6.3). Betrachten wir dazu die paarweisen Unähnlichkeiten der Objekte jener Cluster ($C_p = \{AP, LIB\}$, $C_q = \{ZP, CVP, KON\}$), die beim Übergang von drei zu zwei Clustern verschmolzen werden. Die paarweisen Unähnlichkeiten sind entsprechend

Tab. 9.13: Zahl nächster Nachbarn zum Niveau von 10,5 in der 2-Clusterlösung

Cluster 1	nächste Nachbarn	Zahl	Cluster 2	nächste Nachbarn	Zahl
KP	SP	1	AP	LIB	1
SP	KP	1	LIB	AP, ZP	2
			ZP	ZP, CVP, KON	3
			CVP	ZP, KON	2
			KON	ZP, CVP	2

Tabelle 9.1 auf Seite 234: $u_{AP,ZP} = 17,8$, $u_{AP,CVP} = 21,3$, $u_{AP,KON} = 24,3$, $u_{LIB,ZP} = 10,5$, $u_{LIB,CVP} = 18,9$ und $u_{LIB,KON} = 12,9$. Beim Single-Linkage wird nur die kleinste der sechs Unähnlichkeiten (10,5) verwendet. Dies führt dazu, dass kein Zuwachs gegenüber dem vorausgehenden Verschmelzungsniveau (10,0) auftritt. Hätten wir dagegen die zweitkleinste Unähnlichkeit (12,9) verwendet, wäre der Zuwachs bereits deutlicher ausgefallen (von 10,0 auf 12,9). Für die drittkleinste Unähnlichkeit hätte sich ein Zuwachs von 7,8 ergeben.

Wegen des Verkettungseffekts wird man den Single-Linkage in der Forschungspraxis nicht zum Auffinden einer homogenen Klassifikation einsetzen. Ein Einsatz zum Auffinden einer hierarchischen Darstellung ist dagegen möglich. Der Single-Linkage kann auch zum Auffinden von Ausreißern eingesetzt werden (siehe Abschnitt 12.6). Ferner kann mit ihm geprüft werden, ob überhaupt eine Clusterstruktur vorliegt (Hartigan und Mohanty 1992). Die dem Test zugrunde liegende Überlegung ist sehr einfach: Liegt eine eingipflige Verteilung, also eine homogene Population, vor, führt der Single-Linkage dazu, dass in jedem Verschmelzungsschritt ein sehr großes und ein sehr kleines Cluster fusioniert wird. Die Teststatistik von Hartigan und Mohanty basiert auf dieser Überlegung und verwendet den aus der Prüfung von Zufallszahlen bekannten RUNT-Test. Die Teststatistik wird wie folgt berechnet: In jedem Verschmelzungsschritt V_i werden zwei Cluster C_p und C_q mit einer bestimmten Clustergröße von n_p und n_q verschmolzen. Mit n_i soll das Minimum aus n_p und n_q bezeichnet werden. Die RUNT-Teststatistik ist nun wie folgt definiert:

$$\text{RUNT} = \max_i(n_i).$$

Im eindimensionalen Fall – nur eine Variable wird in die Clusteranalyse einbezogen – kann der kritische Tabellenwert für eine Signifikanz von 95 Prozent mit $n/2,1 + 0,5$ berechnet werden (n entspricht der Zahl der Objekte). Bei mehr als einer Variablen ist ein Nachschlagen in den Tabellenwerten von Hartigan und Mohanty (1992) erforderlich. Strenggenommen setzt das Testverfahren das Vorliegen einer Datenmatrix voraus, da die Signifikanzwerte von der Zahl der Variablen abhängen. Wird eine direkt erhobene

Tab. 9.14: Berechnung der RUNT-Teststatistik für die Entwicklungsländerdaten (Single-Linkage mit euklidischer Distanz)

Schritt	Verschmelzung der Cluster $\{p\}$ und $\{q\}$ mit den Objekten	n_p	n_q	n_i
1	5 – 18	1	1	1
2	1 – 14	1	1	1
3	13 – 17	1	1	1
4	6 – 13	1	2	1
5	1 – 21	2	1	1
6	5 – 6	2	2	2
7	8 – 11	1	1	1
8	1 – 12	3	1	1
9	5 – 8	3	2	2
10	3 – 15	1	1	1
11	1 – 4	4	1	1
12	5 – 16	4	1	1
13	7 – 22	1	1	1
14	2 – 19	1	1	1
15	3 – 5	2	8	2
16	3 – 7	10	2	2
17	2 – 3	12	2	2
18	1 – 2	5	14	5
19	1 – 9	19	1	1
20	1 – 20	20	1	1
21	1 – 10	21	1	1

grau hinterlegt: RUNT-Teststatistik

(Un-)Ähnlichkeitsmatrix untersucht, muss eine Schätzung für die zugrunde liegende Variablenzahl durchgeführt werden.

Wir wollen das Testverfahren am Beispiel der Entwicklungsländerdaten darstellen. Die entsprechenden Werte zur Berechnung der RUNT-Teststatistik enthält die Tabelle 9.14. Im ersten Schritt werden die aus den Objekten {5} und {18} bestehenden Cluster verschmolzen, n_p und n_q sind daher gleich 1. Im zweiten Schritt werden die aus den Objekten {1} und {14} bestehenden Cluster fusioniert, n_p und n_q sind wiederum 1 usw. Im 18. Schritt werden die aus den Objekten {1, 4, 12, 14, 21} und {2, 3, 5, 6, 7, 8, 11, 13, 15, 16, 17, 18, 19, 22} bestehenden Clustern verschmolzen, $n_p = 5$ und $n_q = 14$. Das Minimum ist gleich 5 und gleich dem Wert der RUNT-Teststatistik, da in keinem anderen Verschmelzungsschritt ein größeres Minimum auftritt.

Für die untersuchte Datenkonstellation (22 Objekte, 4 Variablen) ergibt sich entsprechend den Tabellenwerten von Hartigan und Mohanty (1992, S. 66) bei einem Signifikanzniveau

von 95 Prozent ein kritischer Wert von 8, wenn eine Normalverteilung für 20 Objekte als Nullmodell angenommen wird. Die Nullhypothese, dass eine eingipflige Verteilung vorliegt, kann nicht verworfen werden, da der empirische Wert von 5 kleiner dem kritischen Wert von 8 ist. Daraus zu schließen, dass keine Clusterstruktur vorliegt, wäre aber voreilig. Erstens ist der Test nicht besonders mächtig (Hartigan und Mohanty 1992). Zweitens ist das Vorliegen einer mehrgipfligen Verteilung nur eine hinreichende Bedingung für das Vorliegen einer Clusterstruktur. Werden beispielsweise zwei eindimensionale Normalverteilungen (zwei Cluster) mit einer Varianz von 1 gemischt, so ist die bei einem Mischungsverhältnis von 1 : 1 entstehende Mischverteilung eingipflig, wenn die euklidische Distanz zwischen den Mittelwerten μ_1 und μ_2 kleiner/gleich 2 ist (Behboodian 1970, zitiert in Kaufmann und Pape 1984, S. 433). Der von Hartigan und Mohanty entwickelte Test sollte daher zum Nachweis einer hinreichenden Bedingung für eine Clusterstruktur eingesetzt werden: Ein Überschreiten des kritischen Tabellenwertes kann als Hinweis für das Vorliegen einer Clusterstruktur betrachtet werden, ein Nichtüberschreiten bedeutet dagegen nur, dass keine klar erkennbare mehrgipflige Verteilung, nicht aber, dass überhaupt keine Clusterstruktur vorliegt.

9.3 Complete-Linkage für überlappende Cluster

Opitz und Wiedemann (1989) haben das Modell des Complete-Linkage zum Auffinden von *überlappenden Clusterstrukturen* erweitert. Dazu wird der Grundalgorithmus der hierarchisch-agglomerativen Verfahren wie folgt geändert:

Schritt 1: Jedes Objekt g bildet zu Beginn wiederum ein selbständiges Cluster $C_g = \{i\}$ mit $g = 1, 2, \dots, n$. Die Clusterzahl K wird also gleich n (Zahl der Objekte) gesetzt. Ferner wird für jedes Objektpaar (g, g^*) eine Dummy-Variablen C_{g,g^*} eingeführt. Diese besitzt den Wert 0, wenn die Unähnlichkeit des Objektpaares (g, g^*) noch nicht zur Verschmelzung ausgewählt wurde, und den Wert 1, wenn dies der Fall ist. Zu Beginn sind somit alle Dummy-Variablen C_{g,g^*} gleich 0.

Schritt 2: Suche in der empirischen Unähnlichkeitsmatrix unter den Objektpaaren, deren Unähnlichkeit noch nicht ausgewählt wurde ($C_{g,g^*} = 0$), das Objektpaar mit der kleinsten Unähnlichkeit. Dieses Objektpaar soll mit (p, q) und das entsprechende Unähnlichkeitsniveau (Verschmelzungsniveau) mit $u_{p,q}$ bezeichnet werden.

Schritt 3: Bestimme alle Cluster C_i , denen das Objekt p angehört. Prüfe, ob das Objekt q zu allen Objekten g des Clusters C_i eine Unähnlichkeit kleiner/gleich $u_{p,q}$ hat. Bei »ja« ordne das Objekt q dem Cluster C_i zu. Formal ausgedrückt wird das Objekt q dann dem Cluster C_i zugeordnet wenn gilt

$$p \in C_i \cap u_{q,g} \leq u_{p,q} \quad \forall g \in C_i \Rightarrow q \rightarrow C_i ,$$

wobei das Zeichen → für »ordne zu« steht.

Schritt 4: Führe Schritt 3 für das Objekt q aus. Das Objekt p wird dann einem Cluster C_i zugeordnet, wenn gilt

$$q \in C_i \cap u_{p,g} \leq u_{p,q} \quad \forall g \in C_i \Rightarrow p \rightarrow C_i .$$

Schritt 5: Prüfe alle Clusterpaare (C_i, C_j) mit $j > i$ auf Gleichheit. Sind zwei Cluster C_i und C_j gleich, dann streiche Cluster j und reduziere den Clusterzähler um 1: $K = K - 1$. Gleichheit bedeutet, dass die beiden Cluster aus denselben Objekten bestehen.

Schritt 6: Setze die Dummy-Variable $C_{p,q}$ gleich 1. Die Unähnlichkeit zwischen den Objekten p und q wird als ausgewählt markiert.

Schritt 7: Prüfe, ob alle Unähnlichkeiten bereits ausgewählt wurden. Bei »nein« gehe zu Schritt 2, ansonsten beende das Verfahren.

Im Unterschied zum Grundalgorithmus der hierarchisch-agglomerativen Verfahren wird der Algorithmus $\binom{n(n-1)}{2}$ -mal durchlaufen. Es entstehen also $\binom{n(n-1)}{2}$ Clusterlösungen, wobei zwei oder mehrere aufeinanderfolgende Clusterlösungen identisch sein können. Die Verschmelzungsniveaus V_i der einzelnen Schritte sind gleich den Werten der untersuchten Unähnlichkeitsmatrix.

Der Algorithmus soll für das wahlsoziologische Beispiel verdeutlicht werden, wobei nur die ersten drei Verschmelzungsschritte durchgerechnet werden. Entsprechend dem ersten Schritt des Algorithmus (Initialisierungsschritt) werden sieben Cluster gebildet ($K = 7$). Das Cluster C_1 besteht aus dem ersten Objekt ($C_1 = \{\text{KP}\}$), das Cluster C_2 aus dem zweiten Objekt ($C_2 = \{\text{SP}\}$) usw. Die Dummy-Variablen C_{g,g^*} werden gleich 0 gesetzt.

Verschmelzungsschritt 1: Da alle Unähnlichkeiten noch nicht untersucht wurden, ist das Objektpaar mit der kleinsten Unähnlichkeit das Objektpaar ($p = \{\text{CVP}\}$, $q = \{\text{KON}\}$) mit $u_{p,q} = 7,3$. Das Objekt p (sechstes Objekt: $\{\text{CVP}\}$) gehört dem Cluster C_6 an. Das Objekt q ($\{\text{KON}\}$) besitzt zu dem einzigen Objekt $\{\text{CVP}\}$ des Clusters C_6 eine Unähnlichkeit kleiner/gleich 7,3 und wird daher dem Cluster C_6 zugeordnet. Das Objekt q (siebtes Objekt: $\{\text{KON}\}$) gehört dem Cluster C_7 an. Das Objekt p ($\{\text{CVP}\}$) besitzt zu dem einzigen Objekt $\{\text{KON}\}$ dieses Clusters wiederum eine Unähnlichkeit kleiner/gleich 7,3 und wird daher dem Cluster C_7 zugeordnet. Die neu berechneten Cluster C_6 und C_7 bestehen aus den Objekten $\{\text{CVP}, \text{KON}\}$ und sind gleich. C_7 wird daher gestrichen und $C_{6,7} = C_{\{\text{CVP}, \text{KON}\}}$ wird vor einer weiteren Verschmelzung gleich 1 gesetzt. Nach dem ersten Schritt liegt also folgende Clusterstruktur vor: $C_1 = \{\text{KP}\}$, $C_2 = \{\text{SP}\}$, $C_3 = \{\text{AP}\}$, $C_4 = \{\text{LIB}\}$, $C_5 = \{\text{ZP}\}$, $C_6 = \{\text{CVP}, \text{KON}\}$ und C_7 wird gestrichen.

Verschmelzungsschritt 2: Das Objektpaar unter den noch nicht ausgewählten Objektpaaren mit der kleinsten Unähnlichkeit ($u_{p,q} = 7,6$) ist das Objektpaar ($p = \{\text{ZP}\}$, $q = \{\text{CVP}\}$). Da die Unähnlichkeit von $q (\{\text{CVP}\})$ kleiner/gleich dem einzigen Objekt $p (\{\text{ZP}\})$ des Clusters C_5 ist, wird das Objekt q dem Cluster C_5 zugeordnet. Das Objekt q gehört dem Cluster $C_6 = \{\text{CVP}, \text{KON}\}$ an. Das Objekt p besitzt zwar zum Objekt $\{\text{CVP}\}$ eine Unähnlichkeit kleiner/gleich 7,6, nicht aber zum zweiten Objekt $\{\text{KON}\}$ des Clusters C_6 ($u_{\text{ZP},\text{KON}} = 8,1 > 7,6$). Eine Zuordnung zum Cluster C_6 findet daher nicht statt. Eine Streichung von Clustern ist nicht erforderlich, da keine gleichen Cluster vorliegen. Vor einem erneuten Durchlaufen der Schritte 2 bis 7 wird die Dummy-Variable $C_{5,6} = C_{\text{ZP},\text{CVP}} = 1$ gesetzt. Nach dem zweiten Schritt liegt folgende Clusterstruktur vor: $C_1 = \{\text{KP}\}$, $C_2 = \{\text{SP}\}$, $C_3 = \{\text{AP}\}$, $C_4 = \{\text{LIB}\}$, $C_5 = \{\text{ZP}, \text{CVP}\}$, $C_6 = \{\text{CVP}, \text{KON}\}$ und C_7 wird gestrichen.

Verschmelzungsschritt 3: Das Objektpaar unter den noch nicht ausgewählten Objektpaaren mit der kleinsten Unähnlichkeit ($u_{p,q} = 8,1$) ist das Objektpaar ($p = \{\text{ZP}\}$, $q = \{\text{KON}\}$). Das Objekt p gehört dem Cluster $C_5 = \{\text{ZP}, \text{CVP}\}$ an. Da die Unähnlichkeit von $q (\{\text{KON}\})$ zu den beiden Objekten $\{\text{ZP}\}$ und $\{\text{CVP}\}$ des Clusters C_5 kleiner/gleich 8,1 ($u_{\text{ZP},\text{KON}} = 8,1$, $u_{\text{CVP},\text{KON}} = 7,3$) ist, wird das Objekt q dem Cluster C_5 zugeordnet. Das Objekt q gehört dem Cluster $C_6 = \{\text{CVP}, \text{KON}\}$ an. Da das Objekt p zu den beiden Objekten des Clusters C_6 eine Unähnlichkeit kleiner/gleich 8,1 ($u_{\text{ZP},\text{KON}} = 8,1$, $u_{\text{ZP},\text{CVP}} = 7,6$) besitzt, wird es dem Cluster C_6 zugeordnet. C_5 und C_6 bestehen nun aus den Objekten $\{\text{ZP}, \text{CVP}, \text{KON}\}$ und sind somit identisch. Cluster C_6 wird gestrichen. Vor dem Schritt 8 wird die Dummy-Variable $C_{5,7} = C_{\text{ZP},\text{KON}}$ gleich 1 gesetzt. Führt man alle $7(7 - 1)/2 = 21$ Verschmelzungsschritte durch, ergibt sich das in der Tabelle 9.15 auf der nächsten Seite dargestellte Bild: Bei einem Verschmelzungsniveau von 7,3 liegen sechs Cluster vor. Überlappungen treten nicht auf. Beim zweiten Schritt liegen ebenfalls sechs Cluster vor. Es tritt eine Überlappung auf. Die Christliche Volkspartei gehört sowohl dem fünften als auch dem sechsten Cluster an. Die sechs Cluster kann man sich graphisch wie in der Abbildung 9.5 auf der nächsten Seite dargestellt vorstellen.

Die Darstellung der Ergebnisse in Form eines Dendrogramms ist nicht möglich. Eine graphische Darstellung kann wie folgt erreicht werden (Opitz und Wiedemann 1989): Für die untersuchte (Un-)Ähnlichkeitsmatrix wird eine zweidimensionale Darstellung mit Hilfe eines räumlichen Darstellungsverfahrens, zum Beispiel mit Hilfe der nicht-metrischen mehrdimensionalen Skalierung, berechnet. In diese kann die Clusterlösung eingezeichnet werden.³ Die Güte der Modellanpassung spielt dabei keine Rolle. Aus Gründen der Übersichtlichkeit wird man in die räumliche Darstellung nur die letzten Schritte oder die für die weitere Analyse ausgewählte Clusterlösung eintragen. In der Abbildung 9.5 auf der nächsten Seite wurde die 6-Clusterlösung mit einer Überlappung

³ Diese Technik ist selbstverständlich auch bei den anderen Clusteranalyseverfahren möglich.

Tab. 9.15: Verschmelzungsschema für das wahlsoziologische Beispiel für den Complete-Linkage für überlappende Cluster

Schritt	Zahl der Cluster	Zahl der Überlap-pungen	Verschmel-zungs-niveau	Schritt	Zahl der Cluster	Zahl der Überlap-pungen	Verschmel-zungs-niveau
1	6	0	7,3	12	3	3	21,3
2	6	1	7,6	13	2	0	24,3
3	5	0	8,1	14	2	1	25,3
4	4	0	8,7	15	2	1	33,2
5	3	0	10,0	16	2	2	33,7
6	3	0	10,5	17	2	3	37,9
7	3	0	12,5	18	2	4	40,0
8	3	0	14,8	19	2	4	49,3
9	3	0	17,8	20	2	6	50,2
10	3	2	18,9	21	1	0	50,5
11	3	2	19,0				

und die 3-Clusterlösung ohne Überlappungen eingezeichnet. Die räumliche Darstellung wurde mit der nichtmetrischen mehrdimensionalen Skalierung berechnet.

Da die Darstellung der Ergebnisse des Complete-Linkage für überlappende Cluster in Form eines hierarchischen Dendrogramms nicht möglich ist, kann der Complete-Linkage für überlappende Cluster auch nicht zum Auffinden einer hierarchischen Darstellung eingesetzt werden. Er ist aber dafür zum Auffinden einer (überlappenden) Clusterstruktur geeignet. Zur Beurteilung der Modellanpassung und der Beschreibung der Cluster können alle für den Complete-Linkage angeführten Teststatistiken und Strategien ein-

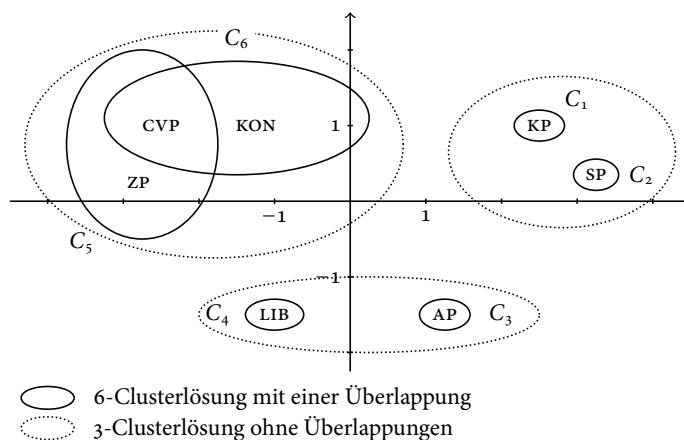


Abb. 9.5: Graphische Darstellung der Ergebnisse des Complete-Linkage für überlappende Cluster

gesetzt werden. Bei der Bestimmung der Clusterzahl kann man sich dabei neben dem Scree-Test oder dem Kriterium einer deutlichen Zunahme im Verschmelzungsschema an der Stabilität einer Lösung orientieren. Tritt eine Lösung mehrfach hintereinander auf, kann diese als relativ stabil bezeichnet und für weitere Analysen ausgewählt werden. In unserem Beispiel tritt die 3-Clusterlösung ohne Überlappungen fünfmal hintereinander auf. Man wird sie daher als stabil bezeichnen. Der Zuwachs zwischen dem Distanzniveau des ersten Auftretens der 3-Clusterlösung ohne Überlappungen und einer neuen Lösung (3-Clusterlösung mit zwei Überlappungen) beträgt 10,1.

9.4 Verallgemeinerte Nächste-Nachbarn-Verfahren

Der Complete- und Single-Linkage gehen von in der Forschungspraxis oft zu strengen bzw. zu schwachen Forderungen hinsichtlich der Homogenität in den Clustern aus. Der Complete-Linkage führt wegen des Dilatationseffekts häufig zur Bildung von Restclustern, während der Single-Linkage zu Verkettungen von gut getrennten Clustern führen kann. Diese beiden nachteiligen Effekte versuchen die *verallgemeinerten Nächste-Nachbarn-Verfahren* dadurch zu beseitigen, dass nicht nur der weitest entfernte Nachbar (Complete-Linkage) oder der nächste Nachbar (Single-Linkage) in die Clusterbildung eingeht. Dabei kann von zwei unterschiedlichen Modellen ausgegangen werden (siehe dazu Abschnitt 6.2). Im ersten Zugang wird gefordert, dass jedes Objekt eines Clusters eine bestimmte Anzahl von nächsten Nachbarn haben soll. Ein diesbezügliches Verfahren wurde von Denz (1983, 1989, S. 48–49) zur Analyse von soziometrischen Tests entwickelt. Hier können Cliques von Personen so bestimmt werden, dass ein bestimmter Prozentsatz der möglichen Paarbildungen zur Bildung einer Clique erfüllt sein muss. Überlappungen sind erlaubt. Weitere nach diesem Ansatz vorgehende Verfahren sind in Jain und Dubes (1988, S. 128–130) beschrieben. Der zweite Modellansatz der verallgemeinerten Nächste-Nachbarn-Verfahren nimmt an, dass innerhalb eines Clusters jedes Objekt einen B -ten Nachbarn haben soll.

Wir wollen hier das Verfahren von Gowda und Krishna (1978) darstellen (siehe dazu auch Jain und Dubes 1988, S. 129–130), das dem zweiten Modellansatz angehört. Das Verfahren geht von folgender Modellvorstellung aus: Jedes Objekt g hat $n - 1$ Nachbarn g^*, g^{**}, \dots , die im Hinblick auf g in eine Ordnung gebracht werden können: $u_{g,g^*} \leq u_{g,g^{**}} \leq u_{g,g^{***}}$ usw. Für ein Cluster C_k wird nun gefordert, dass jedes Objekt g in dem Cluster zumindest einen B -ten Nachbarn hat. Diese Modellvorstellung lässt sich durch folgenden Algorithmus realisieren:

Schritt 1: Berechnung der Nächsten-Nachbar-Funktion b_{g,g^*} für jedes Objektpaar (g,g^*) . Die Nächste-Nachbar-Funktion ist wie folgt definiert:

$$b_{g,g^*} = \begin{cases} b_{g|g^*} + b_{g^*|g} & \text{wenn } b_{g|g^*} < B \cap b_{g^*|g} < B \\ 0 & \text{sonst.} \end{cases}$$

Der Ausdruck $b_{g|g^*}$ gibt an, der wievielte nächste Nachbar des Objekts g^* das Objekt g ist. Umgekehrt gibt $b_{g^*|g}$ an, der wievielte nächste Nachbar des Objekts g das Objekt g^* ist. Ist $b_{g|g^*}$ oder $b_{g^*|g}$ größer dem vorgegebenen Schwellenwert B , wird die Nächste-Nachbar-Funktion auf eine beliebig wählbare große Zahl gesetzt. Zwei Objekte g und g^* werden also nur als nächste Nachbarn betrachtet, wenn g zu g^* und g^* zu g mindestens der B -te Nachbar ist. Gowda und Krishna sprechen daher vom gegenseitigen Nächste-Nachbarn-Verfahren (»mutual neighborhood clustering«). Durch dieses Vorgehen wird gewährleistet, dass in den gebildeten Clustern jedes Objekt zumindest einen B -ten Nachbarn hat.

Schritt 2: Ordne zu Beginn jedes Objekt g einem selbständigen Cluster $C_g = \{g\}$ zu. Führe einen weiteren Schwellenwert für die Nächste-Nachbar-Funktion ein und setze diesen gleich dem minimalen Wert der Nächste-Nachbar-Funktion von 2. Wir wollen diesen Schwellenwert mit b bezeichnen.

Schritt 3: Wähle alle Objektpaare (g,g^*) mit einem Wert von b in der Nächste-Nachbar-Funktion ($b_{g,g^*} = b$) aus.

Schritt 4: Suche unter den ausgewählten Objektpaaren das Objektpaar (g, g^*) mit der kleinsten Unähnlichkeit. Wir wollen dieses Objektpaar mit (p,q) bezeichnen. Bestimme die Clusterzugehörigkeit C_p und C_q der beiden Objekte ($p \in C_p, q \in C_q$). Gilt $C_p = C_q$ gehören die beiden Objekte bereits demselben Cluster an. Eine Fusionierung ist daher nicht erforderlich. Ist dagegen $C_p \neq C_q$, werden die beiden Cluster fusioniert und C_q gestrichen.

Schritt 5: Streiche das Objektpaar (g, g^*) aus der Liste der ausgewählten Objektpaare. Prüfe, ob die Liste der ausgewählten Objektpaare leer ist. Bei »ja« gehe zu Schritt 6, bei »nein« wiederhole Schritt 4.

Schritt 6: Erhöhe b um 1 ($b = b+1$). Ist b kleiner/gleich dem maximalen Wert der Nächste-Nachbar-Funktion von $2B$, gehe zu Schritt 3, ansonsten beende das Verfahren.

Der Algorithmus soll für das wahlsoziologische Beispiel veranschaulicht werden. Wir erzeugen uns dazu zunächst aus der Unähnlichkeitsmatrix der Tabelle 9.1 auf Seite 234 die Matrix der bedingten Nächste-Nachbar-Funktion $b_{g|g^*}$. Es ergeben sich die in der Tabelle 9.16 dargestellten Werte. Die Tabelle ist zeilenweise zu lesen: Das Objekt {SP} ist der erstmächtste Nachbar des Objekts {KP}, da gilt $u_{KP,SP} < u_{KP,AP} < \dots$ Das Objekt {AP} ist der zweitmächtste Nachbar des Objekts {KP} usw. Für einen Schwellenwert von $B = 2$ ergeben sich nun folgende Werte der Nächste-Nachbar-Funktion, wobei der Wert

Tab. 9.16: Werte der bedingten Nächste-Nachbar-Funktion $b_{g|g^*}$ für das wahlsoziologische Beispiel

g^*	Bezugsobjekte g für die Nächste-Nachbar-Funktion $b_{g g^*}$						
	KP	SP	AP	LIB	ZP	CVP	KON
KP	0	1	2	3	4	5	6
SP	1	0	2	3	4	5	6
AP	6	2	0	1	3	4	5
LIB	6	5	1	0	2	4	3
ZP	6	5	4	3	0	1	2
CVP	6	5	4	3	2	0	1
KON	6	5	4	3	2	1	0

10^{10} verwendet werden soll, wenn die (gegenseitige) Nächste-Nachbar-Funktion nicht berechnet werden darf:

$$b_{\text{KP}, \text{SP}} = b_{\text{KP}|\text{SP}} + b_{\text{SP}|\text{KP}} = 1 + 1 = 2 ,$$

$$b_{\text{KP}, \text{AP}} = b_{\text{KP}|\text{AP}} + b_{\text{AP}|\text{KP}} = (6) + 2 = 10^{10} ,$$

da der Wert von $b_{\text{KP}|\text{AP}}$ größer als der Schwellenwert 2 ist und

$$b_{\text{KP}, \text{LIB}} = b_{\text{LIB}|\text{KP}} + b_{\text{KP}|\text{LIB}} = (6) + (3) = 10^{10} ,$$

da beide Werte größer 2 sind usw.

In Matrixform dargestellt sind die Werte für die Nächste-Nachbar-Funktion b_{g,g^*} in der Tabelle 9.17 wiedergegeben. Für diese Matrix können nun die weiteren Schritte des Algorithmus durchgerechnet werden. Dazu wird zunächst jedes Objekt einem selbständigen Cluster zugewiesen: $C_1 = \{\text{KP}\}$, $C_2 = \{\text{SP}\}$, $C_3 = \{\text{AP}\}$ usw. Der zweite Schwellenwert b wird gleich 2 gesetzt. Im nächsten Schritt des Algorithmus werden alle Objektpaare mit einem Wert von 2 in der Nächste-Nachbar-Funktion b_{g,g^*} ausgewählt. Die Objektpaare,

Tab. 9.17: Werte der (gegenseitigen) Nächste-Nachbar-Funktion b_{g,g^*}

	KP	SP	AP	LIB	ZP	CVP	KON
KP	0	2	10^{10}	10^{10}	10^{10}	10^{10}	10^{10}
SP	2	0	4	10^{10}	10^{10}	10^{10}	10^{10}
AP	10^{10}	4	0	2	10^{10}	10^{10}	10^{10}
LIB	10^{10}	10^{10}	2	0	10^{10}	10^{10}	10^{10}
ZP	10^{10}	10^{10}	10^{10}	10^{10}	0	3	4
CVP	10^{10}	10^{10}	10^{10}	10^{10}	3	0	2
KON	10^{10}	10^{10}	10^{10}	10^{10}	4	2	0

die diese Bedingung erfüllen, sind: ($\{KP\}, \{SP\}$), ($\{AP\}, \{LIB\}$), ($\{CVP\}, \{KON\}$). Diese Objektpaare werden zu Clustern verschmolzen, wobei mit dem Objektpaar mit der kleinsten Unähnlichkeit ($\{CVP\}, \{KON\}$) begonnen wird. Das Objekt $\{CVP\}$ gehört dem Cluster $C_6 = \{CVP\}$ an, das Objekt $\{KON\}$ dem Cluster $C_7 = \{KON\}$. Die beiden Cluster werden verschmolzen. Es wird das neue Cluster $C_6 = \{CVP, KON\}$ gebildet und Cluster C_7 wird gestrichen. Das Objektpaar ($\{CVP\}, \{KON\}$) wird aus der Liste der ausgewählten Objektpaare entfernt, da es bereits untersucht wurde. In der verbleibenden Liste wird wiederum das Objektpaar mit der kleinsten Unähnlichkeit gesucht. Dies ist das Objektpaar ($\{KP\}, \{SP\}$). Da beide Objekte unterschiedlichen Clustern ($C_1 = \{KP\}$, $C_2 = \{SP\}$) angehören, werden die beiden Cluster fusioniert (neues $C_1 = \{KP, SP\}$), C_2 wird gestrichen. Auch für das letzte verbleibende Objektpaar (AP, LIB) findet eine Clusterverschmelzung statt: $C_3 = \{AP, LIB\}$, C_4 wird gestrichen. Damit sind alle Objektpaare mit einem Wert von 2 in der Nächste-Nachbar-Funktion abgearbeitet. Es liegt folgende Clusterstruktur vor: $C_1 = \{KP, SP\}$, C_2 wird gestrichen, $C_3 = \{AP, LIB\}$, C_4 wird gestrichen, $C_5 = \{ZP\}$, $C_6 = \{CVP, KON\}$, C_7 wird gestrichen.

Vor einem erneuten Verschmelzungsschritt wird b um 1 erhöht ($b = b + 1 = 2 + 1 = 3$). Da b kleiner/gleich dem maximalen Wert von $4 = 2B$ ist, werden die Schritte 3 bis 6 erneut mit einem Wert von $b = 3$ durchlaufen. Das einzige Objektpaar mit einem Wert von 3 in der Nächste-Nachbar-Funktion ist das Objektpaar ($\{ZP\}, \{CVP\}$). Da die beiden Objekte unterschiedlichen Clustern angehören ($\{ZP\} \in C_5$, $\{CVP\} \in C_6$), findet eine Verschmelzung der Cluster C_5 und C_6 statt. C_5 ist nach der Verschmelzung gleich $\{ZP, CVP, KON\}$. C_6 wird gestrichen. Es liegt nun folgende Clusterstruktur vor: $C_1 = \{KP, SP\}$, C_2 wird gestrichen, $C_3 = \{AP, LIB\}$, C_4 wird gestrichen, $C_5 = \{ZP, CVP, KON\}$, C_6 und C_7 werden gestrichen.

Wenn wir b um 1 erhöhen ($b = b + 1 = 3 + 1 = 4$) und zu Schritt 3 gehen, da $b \leq 2B = 4$ ist, werden zwei Objektpaare mit einem Wert von 4 in der Nächste-Nachbar-Funktion ausgewählt: ($\{SP\}, \{AP\}$) und ($\{ZP\}, \{KON\}$). Das Objektpaar mit der kleinsten Unähnlichkeit ist ($\{ZP\}, \{KON\}$). Da die beiden Objekte bereits demselben Cluster angehören, ist eine Verschmelzung von Clustern nicht erforderlich. Die Objekte $\{SP\}$ und $\{AP\}$ gehören dagegen unterschiedlichen Clustern an ($\{SP\} \in C_1$, $\{AP\} \in C_3$). Die Cluster C_1 und C_3 werden daher verschmolzen, C_3 wird gestrichen.

Ein erneutes Durchlaufen der Schritte 3 bis 5 ist nicht erforderlich, da für b bereits der maximale Wert von 4 untersucht wurde. Insgesamt führt das Verfahren zu folgender Clusterstruktur, wobei die gestrichenen Cluster nicht mehr angeführt und die verbleibenden Cluster neu nummeriert werden: $C_1 = \{KP, SP, AP, LIB\}$, $C_2 = \{ZP, CVP, KON\}$ und $B = 2$. Die Ergebnisse sind wie folgt zu lesen: Wird gefordert, dass jedes Objekt eines Clusters zumindest einen zweitnächsten Nachbarn hat ($B = 2$), entstehen zwei Cluster

Tab. 9.18: Clusterzahl und Verschmelzungsniveaus für unterschiedliche Werte von B für die Entwicklungsländerdaten

Schwellenwert B	Clusterzahl	Verschmelzungsniveau ^{a)}	Schwellenwert B	Clusterzahl	Verschmelzungsniveau ^{a)}
1	18	1,95	6	3	2,90
2	9	1,95	7	2	3,34
3	9	1,95	8	2	3,34
4	6	2,85	9	2	3,34
5	3	2,90	10	1	5,53

a) maximale Unähnlichkeit zum nächsten Nachbarn

C_1 und C_2 . Die maximale Unähnlichkeit zum nächsten Nachbarn in jedem Cluster ist für jedes Objekt gleich 14,38.

Vergleichen wir die Ergebnisse mit jenen des Complete- oder Single-Linkage, so unterscheiden sich die 2-Clusterlösungen dahingehend, dass beim Verfahren von Gowda und Krishna die Objekte {AP} und {LIB} dem ersten Cluster zugeordnet werden, beim Single- und Complete-Linkage dagegen dem zweiten Cluster. Die Ursache dafür ist darin zu sehen, dass nicht die Größenanordnung der empirischen Unähnlichkeiten den Verschmelzungsprozeß steuert, sondern die Rangordnung der nächsten Nachbarn für jedes Objekt. Die Unähnlichkeit zwischen den Objekten {LIB} und {ZP} (10,50) ist zwar kleiner 14,38, das Objekt {LIB} ist aber eben nur der drittäufigste Nachbar von {ZP}. Die Nächste-Nachbar-Funktion $b_{LIB|ZP}$ ist daher 10^{10} , wenn ein Schwellenwert von $B = 2$ gewählt wird.

Die Ergebnisse des gegenseitigen Nächste-Nachbarn-Verfahrens von Gowda und Krishna hängen entscheidend von dem Schwellenwert B ab. Wird B sehr klein gewählt, erhält man sehr viele, aber homogene Cluster, wird B sehr groß gewählt, entstehen sehr wenige, aber dafür sehr heterogene Cluster. Der Algorithmus ist somit sehr flexibel. In der Forschungspraxis wird man in der Regel mehrere Werte von B durchprobieren und jene Lösung bzw. Lösungen mit einer guten Modellanpassung auswählen. Für die Entwicklungsländerdaten ergibt sich beispielsweise bei Verwendung der City-Block-Metrik für unterschiedliche Werte von B das in der Tabelle 9.18 wiedergegebene Bild.

Wie beim Complete-Linkage für überlappende Cluster kann der Fall eintreten, dass mehrere aufeinanderfolgende Clusterlösungen identisch sind. Diese Lösungen kann man als stabil betrachten. In dem Beispiel ist dies für $B = 2$ und $B = 3$ sowie für $B = 5, 6$ und für $B = 7, 8, 9$ der Fall. In dem Verschmelzungsschema lassen sich zwei Sprünge erkennen: Beim Übergang von neun zu sechs Clustern (Zunahme von 1,95 auf 2,85) sowie beim Übergang von zwei zu einem Cluster (Zunahme von 3,34 auf 5,53). In dem Beispiel wird man somit die 9- und 2-Clusterlösung weiter untersuchen.

Allgemein empfehlen wir das Verfahren von Gowda und Krishna wegen der dargestellten Problematik (Wahl des Schwellenwertes von B , starke Abhängigkeit von der Anordnung der nächsten Nachbarn) nur dann anzuwenden, wenn

1. die direkt erhobene Unähnlichkeitsmatrix nichtmetrisch oder die gesuchte Klassifikation invariant gegenüber monotonen Transformationen sein soll und
2. die Homogenitätsforderung des Complete-Linkage zu streng und jene des Single-Linkage zu schwach ist und
3. der Complete-Linkage für überlappende Cluster nicht in Frage kommt, da eine überlappungsfreie Klassifikation gefunden werden soll. Ist dagegen Invarianz nicht erwünscht, empfiehlt sich ein Mittelwertverfahren.

Da für die Entwicklungsländerdaten die Distanzmatrix aus einer Datenmatrix berechnet wurde, ist sie metrisch. Man wird sich daher, wenn Punkt 2 zutrifft, eher für ein Mittelwertverfahren oder ein Verfahren zur Konstruktion von Clusterzentren entscheiden, wenn die gesuchte Klassifikation überlappungsfrei sein soll. Der Beseitigung der Nachteile des Complete- und Single-Linkage wird ein größeres Gewicht beigemessen als der Invarianz gegenüber monotonen Transformationen.

9.5 Mittelwertverfahren

In Abschnitt 6.3 haben wir bereits darauf hingewiesen, dass die *Mittelwertverfahren* (*Average-Linkage*, *Weighted-Average-Linkage*, *Within-Average-Linkage*) als Modifikation des Complete- und Single-Linkage betrachtet werden können. Bei der Neuberechnung der Unähnlichkeiten im vierten Schritt des Algorithmus der hierarchisch-agglomerativen Verfahren (siehe Abschnitt 9.1.1) findet eine Mittelwertbildung statt. Dadurch wird versucht, den Kontraktionseffekt des Single-Linkage und den Dilatationseffekt des Complete-Linkage zu vermeiden. Die Mittelwertbildung kann nach unterschiedlichen Strategien erfolgen. Wir wollen hier drei Verfahren darstellen.

Average-Linkage:

$$u_{(p+q),i}^{\text{neu}} = \frac{u_{p,i} + u_{q,i}}{2} .$$

Weighted-Average-Linkage:

$$u_{(p+q),i}^{\text{neu}} = \frac{n_p u_{p,i} + n_q u_{q,i}}{n_p + n_q} .$$

Within-Average-Linkage:

$$\begin{aligned} u_{(p+q),i}^{\text{neu}} &= \frac{n_p u_{p,i} + n_q u_{q,i}}{n_p + n_q} \\ u_{(p+q),(p+q)}^{\text{neu}} &= \frac{\frac{2}{n_p(n_p-1)} \cdot u_{p,p} + \frac{2}{n_q(n_q-1)} \cdot u_{q,q} + n_p n_q u_{p,q}}{(n_p+n_q)(n_p+n_q-1)/2}, \end{aligned}$$

mit n_p und n_q als Zahl der Fälle in Cluster p bzw. Cluster q .

Für den *Within-Average-Linkage* ist ferner eine Modifikation bei der Suche nach der kleinsten Unähnlichkeit (Schritt 2) des Algorithmus erforderlich. Es wird jenes Objekt-paar (p,q) ausgewählt, für das der Ausdruck

$$u_{p,q}^* = \frac{\frac{2}{n_p(n_p-1)} \cdot u_{p,p} + \frac{2}{n_q(n_q-1)} \cdot u_{q,q} + n_p n_q u_{p,q}}{(n_p+n_q) \cdot (n_p+n_q-1)/2}$$

ein Minimum ist. Diese Änderung ist erforderlich, um das Ziel des Within-Average-Linkage zu erreichen. Dieses besteht darin, dass das in einem Verschmelzungsschritt i ausgewiesene Verschmelzungsniveau V_i gleich der mittleren paarweisen Unähnlichkeit zwischen den Objekten des neu gebildeten Clusters $p + q$ ist. Es soll also gelten:

$$V_i = u_{p,q}^* = \sum_{\substack{g,g^* \in p+q \\ g \neq g^*}} u_{g,g^*} \cdot \frac{2}{n_{p+q} (n_{p+q} - 1)}.$$

Das Verschmelzungsniveau V_i lässt sich wie folgt interpretieren: Die paarweisen mittleren Unähnlichkeiten in den beim Verschmelzungsschritt i gebildeten Clustern ist kleiner/gleich V_i . Die beiden anderen dargestellten Mittelwertstrategien führen zu folgender Interpretation des Verschmelzungsniveaus:

Weighted-Average-Linkage: Das Verschmelzungsniveau V_i ist gleich der mittleren paarweisen Unähnlichkeit zwischen den Clustern p und q , die im Schritt i verschmolzen werden:

$$V_i = u_{p,q} = \sum_{\substack{g \in p \\ g^* \in q}} u_{g,g^*} \cdot \frac{1}{n_p n_q}.$$

Average-Linkage: Das Verschmelzungsniveau V_i ist gleich der mittleren paarweisen Unähnlichkeit zwischen den Clustern p und q , wenn die Cluster im gesamten vorausgehenden Verschmelzungsschritt als gleich groß betrachtet werden. Das Verschmelzungsniveau beim Average-Linkage besitzt somit keine direkte Beziehung zu den empirischen Unähnlichkeiten. Es ist nur schwer zu interpretieren, und der Average-Linkage sollte daher nur

Tab. 9.19: Verschmelzungsschemas der Mittelwertverfahren für das wahlsoziologische Beispiel der Tabelle 9.1 auf Seite 234

n_C	Within-Average		Weighted-Average		Average	
	Verschmelzung der Cluster mit den Objekten	V_i	Verschmelzung der Cluster mit den Objekten	V_i	Verschmelzung der Cluster mit den Objekten	V_i
6	6 – 7	7,300	6 – 7	7,300	6 – 7	7,300
5	5 – 6	7,667	5 – 6	7,850	5 – 6	7,850
4	1 – 2	8,700	1 – 2	8,700	1 – 2	8,700
3	3 – 4	10,000	3 – 4	10,000	3 – 4	10,000
2	3 – 5	13,870	3 – 5	17,617	3 – 5	16,750
1	1 – 3	23,871	1 – 3	35,390	1 – 3	32,363

Abkürzungen: n_C : Zahl der Cluster; V_i = Verschmelzungsniveau.

Anmerkung: zu den Objektbezeichnungen siehe Tabelle 9.5 auf Seite 239.

dann verwendet werden, wenn die gesuchte Klassifikation invariant gegenüber linearen Transformationen der berechneten oder erhobenen Unähnlichkeiten sein soll, da von den Mittelwertverfahren nur der Average-Linkage diese Eigenschaft besitzt. Die dargestellte Interpretation der Verschmelzungsniveaus soll an dem wahlsoziologischen Beispiel verdeutlicht werden. Die drei Verfahren führen zu den in der Tabelle 9.19 dargestellten Verschmelzungsschemas.

Für die berechneten 2-Clusterlösungen gilt nun:

Within-Average-Linkage: Die mittlere paarweise Unähnlichkeit in den beiden Clustern ist kleiner/gleich 13,870. Die beiden Cluster sind: $C_1 = \{1, 2\} = \{\text{KP, SP}\}$ und $C_2 = \{3, 4, 5, 6, 7\} = \{\text{AP, LIB, ZP, CVP, KON}\}$. Die mittlere paarweise Unähnlichkeit für das erste Cluster können wir unmittelbar aus der Unähnlichkeitsmatrix der Tabelle 9.1 auf Seite 234 oder dem Verschmelzungsschema (Tabelle 9.19) ablesen. Da das Cluster nur aus zwei Objekten besteht, ist sie gleich $u_{\text{KP,SP}} = 8,700$. Im zweiten Cluster ist das angegebene Verschmelzungsniveau von 13,870 gleich der mittleren Unähnlichkeit in dem neu gebildeten Cluster, wie man leicht nachrechnen kann:

$$\begin{aligned}
 u_{p+q} &= \frac{u_{3,4} + u_{3,5} + u_{3,6} + u_{3,7} + u_{4,5} + u_{4,6} + u_{4,7} + u_{5,6} + u_{5,7} + u_{6,7}}{5 \cdot (5-1)/2} \\
 &= \frac{10,0 + 17,8 + 21,3 + 24,3 + \dots + 7,3}{10} = 13,870 = V_i .
 \end{aligned}$$

Weighted-Average-Linkage: Die mittlere paarweise Unähnlichkeit für die beiden in diesem Schritt fusionierten Cluster ist gleich 17,617. Die beiden fusionierten Cluster sind wie

bei dem Within-Average-Linkage: $C_p = \{3,4\}$ und $C_q = \{5,6,7\}$. Die mittlere paarweise Unähnlichkeit zwischen diesen Clustern ist gleich

$$\begin{aligned} u_{p,q} &= \frac{u_{3,5} + u_{3,6} + u_{3,7} + u_{4,5} + u_{4,6} + u_{4,7}}{2 \cdot 3} = \frac{17,8 + 21,3 + 24,3 + \dots + 12,9}{6} \\ &= 17,617 = V_i \end{aligned}$$

und wird als Verschmelzungsniveau ausgewiesen.

Average-Linkage: Das für die 2-Clusterlösung ausgewiesene Verschmelzungsniveau lässt sich nur schwer interpretieren. Es ist die mittlere paarweise Unähnlichkeit zwischen den Clustern, wenn in dem gesamten Verschmelzungsprozeß die Cluster als gleich groß betrachtet werden. Das angegebene Verschmelzungsniveau kommt also wie folgt zustande. Im ersten Schritt werden die Cluster $C_p = \{6\}$ und $C_q = \{7\}$ verschmolzen. Die mittleren paarweisen Unähnlichkeiten zu den anderen Clustern sind:

$$\begin{aligned} u_{6+7,5} &= \frac{7,6 + 8,1}{2} = 7,85, \\ u_{6+7,4} &= \frac{18,9 + 12,9}{2} = 15,90, \\ u_{6+7,3} &= \frac{21,3 + 24,3}{2} = 22,8 \\ \text{usw.} \end{aligned}$$

Im zweiten Schritt werden die Cluster $C_p = \{5\}$ und $C_q = \{6,7\}$ verschmolzen. Die mittleren paarweisen Unähnlichkeiten zu den verbleibenden Clustern sind:

$$\begin{aligned} u_{5+6+7,4} &= \frac{15,9 + 10,5}{2} = 13,20, \\ u_{5+6+7,3} &= \frac{22,8 + 17,8}{2} = 20,30 \\ \text{usw.} \end{aligned}$$

Im dritten Schritt werden die Cluster $C_p = \{1\}$ und $C_q = \{2\}$ verschmolzen. Diese Verschmelzung hat keinen Einfluss auf die mittleren paarweisen Unähnlichkeiten zwischen den im dritten Schritt fusionierten Clustern. Im vierten Schritt werden die Cluster $C_p = \{3\}$ und $C_q = \{4\}$ verschmolzen. Die Unähnlichkeiten zwischen dem neu gebildeten Cluster und den verbleibenden Clustern sind:

$$\begin{aligned} u_{3+4,5+6+7} &= \frac{13,20 + 20,30}{2} = 16,75 \\ \text{usw.} \end{aligned}$$

Tab. 9.20: Verschmelzungsschema für den Within-Average-Linkage

Verschmelzung der Cluster mit den Objekten	Clusterzahl	Verschmelzungs- niveau	Zunahme
2 – 7	6	4,000	0,000
5 – 6	5	7,000	3,700
3 – 4	4	10,000	2,300
2 – 3	3	14,167	4,167
1 – 2	2	20,290	6,123
1 – 5	1	20,257	-0,033

Im fünften Schritt werden die Cluster $C_p = \{3,4\}$ und $C_q = \{5,6,7\}$ zu einem Niveau von 16,75 verschmolzen.

Von den drei vorgestellten Verfahren sollte der Weighted-Average-Linkage verwendet werden, da sein Verschmelzungsschema im Unterschied zum Average-Linkage gut interpretierbar ist und er im Unterschied zum Within-Average-Linkage Inversionen vermeidet. Eine Inversion bedeutet, dass in einem späteren Verschmelzungsschritt kleinere Verschmelzungsniveaus auftreten als in den vorausgehenden. Dafür hat der Within-Average-Linkage den Vorteil, dass die gefundene Klassifikation invariant gegenüber linearen Transformationen ist. Das in Tabelle 9.20 dargestellte Beispiel zeigt das Verschmelzungsschema für eine Clusteranalyse mit dem Within-Average-Linkage, das zu Inversionen führt. Beim Übergang von zwei zu einem Cluster reduziert sich das Distanzniveau. Es ergibt sich daher ein negativer Zuwachs.

Für die dargestellten Mittelwertverfahren werden zum Teil abweichende Namen verwendet, die in Tabelle 9.21 dokumentiert sind.

Als ein Anwendungsbeispiel sollen die bei der multiplen Korrespondenzanalyse zitierten Ergebnisse der Clusteranalyse mit dem Weighted-Average-Linkage (Abschnitt 3.1) etwas genauer betrachtet werden. Dabei soll auch die – bisher noch nicht behandelte – Berechnung des RAND-Index zur Stabilitätsprüfung dargestellt werden. In Abschnitt 3.1 wurde eine multiple Korrespondenzanalyse für die sozialen Schichtungsvariablen »Einkommen von Müttern und deren Partnern« (EinkM und EinkP), »abgeschlossene Schulbildung der Mütter und deren Partner« (SchulbM und SchulbP) und »Beruf der Mütter und deren Partner« (BerufM und BerufP) durchgeführt. Ungültige Angaben (fehlende bzw. nicht valide Werte) wurden dabei als selbständige Ausprägungen betrachtet. Dadurch mussten auch die ordinalen Variablen (EinkM, EinkP, SchulbM und SchulbP) als nominalskaliert betrachtet werden.

Tab. 9.21: Unterschiedliche Namensgebungen für die Mittelwertverfahren

in dieser Arbeit verwendete Bezeichnung	synonyme Bezeichnungen in der Literatur
Average-Linkage (Steinhausen und Langer 1977)	Simple-Average-Linkage (Mojena 1977), Weighted-Average-Linkage (Sneath und Sokal 1973; Kaufman und Rousseeuw 1990; u. a.), Weighted Pair Group Method with Arithmetic Mean (WPGMA)
Weighted-Average-Linkage (Steinhausen und Langer 1977)	Group-Average (Kaufman und Rousseeuw 1990; Everitt 1980; Dunn und Everitt 1982), Unweighted Pair Group Method with Arithmetic Mean (UPGMA), Group (weighted) Average (Mojena 1977), Average-Linkage (Kaufmann und Pape 1984), Between Groups Method (SPSS Inc. 1990a)
Within-Average-Linkage	Average-Linkage Within Groups (SPSS Inc. 1990a, 2008)

Die Ergebnisse der multiplen Korrespondenzanalyse lassen sich dahingehend zusammenfassen (Abschnitt 3.1), dass für eine clusteranalytische Interpretation mindestens 10 Dimensionen erforderlich sind. Eine graphische Clusterbildung war daher nicht möglich. Aus diesem Grund wurde eine Clusteranalyse mit dem Weighted-Average-Linkage durchgeführt. Als Unähnlichkeitsmaß wurde die quadrierte euklidische Distanz verwendet, da die clusteranalytische Interpretation der multiplen Korrespondenzanalyse auf diesem Unähnlichkeitsmaß basiert. Wir wollen nun folgende Fragestellungen untersuchen:

1. Lassen sich die Unähnlichkeiten zwischen den Ausprägungen der sozialen Schichtungsmerkmale hierarchisch darstellen?
2. Lassen sich abgrenzbare homogene Schichten identifizieren?
3. Falls ja, wie stabil ist diese gefundene Schichtung (Clusterung)?

Hierarchische Darstellung der Unähnlichkeitsbeziehungen zwischen den Ausprägungen der sozialen Schichtungsvariablen: Da bei den Mittelwertverfahren das Vorliegen metrischer Ähnlichkeits- und Unähnlichkeitsmaße vorausgesetzt wird, kann neben dem γ -Koeffizienten auch der kophenetische Korrelationskoeffizient KOPH verwendet werden. Tabelle 9.22 enthält die Werte beider Maßzahlen für den Weighted-Average-Linkage. Aus

Tab. 9.22: Modellprüfgrößen γ und KOPH für die hierarchische Darstellung von Ähnlichkeitsbeziehungen sozialer Schichtungsmerkmale

	γ	KOPH
Average-Linkage	0,741	0,847
Weighted-Average-Linkage	0,733	0,856
Within-Average-Linkage	0,354	0,390

Tab. 9.23: Verschmelzungsschema für den Weighted-Average-Linkage für die 10-dimensionale Lösung der multiplen Korrespondenzanalyse (Abschnitt 3.1, die letzten 18 Schritte)

Verschmelzung der Cluster mit den Objekten	Clusterzahl	Verschmelzungs- niveau	Zunahme
8 – 16	18	2,953	0,105
2 – 4	17	3,369	0,415
8 – 23	16	3,663	0,294
2 – 20	15	3,667	0,004
17 – 40	14	4,003	0,336
2 – 5	13	5,126	1,123
2 – 7	12	5,839	0,713
2 – 38	11	6,722	0,883
11 – 37	10	6,942	0,220
2 – 10	9	7,191	0,250
2 – 33	8	7,300	0,108
2 – 8	7	7,455	0,155
1 – 2	6	7,982	0,527
11 – 12	5	10,342	2,360
1 – 39	4	11,606	1,265
1 – 17	3	15,658	4,051
1 – 11	2	17,300	1,642
1 – 34	1	22,268	4,969

grau hinterlegt: deutliche Zunahme im Verschmelzungsniveau

Vergleichsgründen wurden auch die entsprechenden Werte für den Average-Linkage und den Within-Average-Linkage aufgenommen. Wenn wir die Modellprüfgrößen betrachten, so ergeben sich für den Within-Average-Linkage die schlechtesten Werte. Dies ist darauf zurückzuführen, dass in dem Beispiel Inversionen auftreten. Die hierarchische Darstellung ist nicht monoton: In einem späteren Verschmelzungsschritt treten kleinere Verschmelzungsniveaus auf als in den vorausgehenden. Der Within-Average-Linkage ist somit für eine hierarchische Darstellung nicht geeignet, während die beiden anderen Verfahren zu annähernd gleich guten Ergebnissen führen. Das entsprechende Dendrogramm für den Weighted-Average-Linkage wurde bereits in Abschnitt 3.1 dargestellt. Es lassen sich die dort beschriebenen Ähnlichkeitsstrukturen erkennen.

Auffinden einer Klassifikation: Die zweite Fragestellung bezog sich darauf, ob abgrenzbare soziale Schichten erkennbar sind. Betrachten wir dazu das Verschmelzungsschema des Weighted-Average-Linkage, so lässt sich unter anderem beim Übergang von sechs zu fünf Clustern ein deutlicher Zuwachs erkennen (siehe Tabelle 9.23).

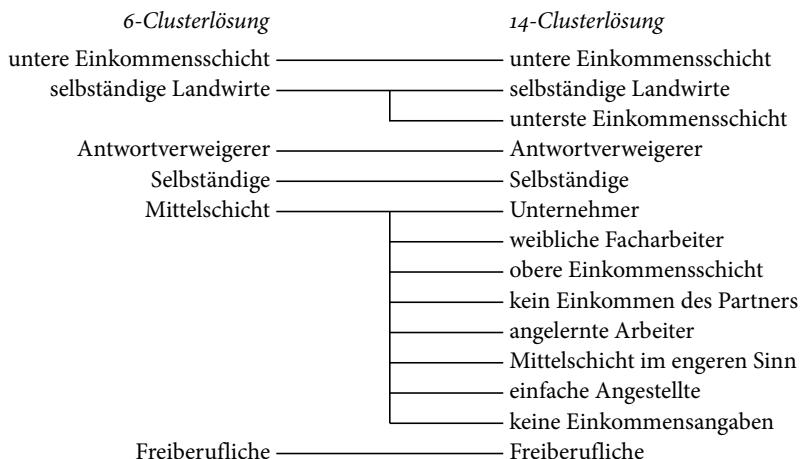


Abb. 9.6: Beziehung zwischen der 14- und 6-Clusterlösung

Die sechs Cluster lassen sich wie folgt beschreiben:

Cluster 1: Untere Einkommensschichten. Das Einkommen des Partners liegt zwischen 5 500 und 7 900 öS.

Cluster 2: Selbständige Landwirte. Die Landwirte gaben ebenfalls ein geringes Einkommen an.

Cluster 3: Cluster der Antwortverweigerungen.

Cluster 4: Selbständig Erwerbstätige. Beide Elternteile sind selbständig erwerbstätig.

Cluster 5: Mittelschicht.

Cluster 6: Freiberuflich Erwerbstätige.

Betrachten wir die 14-Clusterlösung, bei der zum ersten Mal ein deutlicher Zuwachs beim Übergang zur nachfolgenden Clusterlösung auftritt, so wird die Mittelschicht und das Cluster der selbständigen Landwirte aufgespalten. Es ergibt sich folgende Schichtungsstruktur (siehe Abbildung 9.6): Sowohl die 14- als auch die 6-Clusterlösung sind inhaltlich interpretierbar. Die 6-Clusterlösung ist den Daten etwas besser angemessen. Der kophenetische Korrelationskoeffizient zwischen der 6-Clusterlösung und den empirischen Daten beträgt 0,771, während er für die 14-Clusterlösung nur 0,492 ist. Hinsichtlich des γ -Koeffizienten liegt sowohl für die 14- als auch für die 6-Clusterlösung eine gute Modellanpassung vor (für die 14-Clusterlösung: $\gamma = 0,893$; für die 6-Clusterlösung $\gamma = 0,923$). Auch der γ -Koeffizient fällt für die 6-Clusterlösung etwas höher aus.

Die Ursachen für die Unterschiede zwischen γ und KOPH sind:

- Bei γ wird nur die ordinale Information der empirischen und berechneten Unähnlichkeiten verwendet.

2. Beide Koeffizienten sind unterschiedlich normiert. In KOPH geht die Standardabweichung der berechneten (theoretischen) Unähnlichkeiten ein. Da diese nur Werte von 0 und 1 haben können, ist die Standardabweichung ein Maximum, wenn der Anteil der Einsen gleich 0,5 ist. Eine solche Konstellation tritt mit einer größeren Wahrscheinlichkeit in der Mitte des Verschmelzungsschemas auf als am Ende oder am Beginn. In unserem Beispiel liegt die 14-Clusterlösung näher an der Mitte des Verschmelzungsschemas als die 6-Clusterlösung.

Aufgrund des dargestellten Effektes wird man die 14-Clusterlösung nicht von vornherein wegen der geringen kophenetischen Korrelation verwerfen. Als weitere Entscheidungskriterien für eine 14- oder 6-Clusterlösung können neben inhaltlichen Überlegungen die Ergebnisse einer Stabilitätsprüfung verwendet werden. Eine stabile Clusterlösung ist eine formale Voraussetzung für deren Gültigkeit. Die Grundlogik der Stabilitätsprüfung besteht darin, dass die Einflüsse von »unsicherer« Entscheidungen, die bei der Clusterlösung getroffen wurden, untersucht werden. In unserem Beispiel unter anderem:

1. Sind zehn oder mehr Dimensionen erforderlich? Bei der Analyse haben wir uns für die Mindestzahl an Dimensionen entschieden, die für eine clusteranalytische Interpretation erforderlich ist. Bei der Stabilitätsprüfung wird nun untersucht, ob sich die Ergebnisse ändern, wenn anstelle der zehn Dimensionen elf, zwölf oder mehr Dimensionen in die Analyse eingehen.
2. Wie ändern sich die Ergebnisse, wenn die ungültigen Angaben aus der Analyse eliminiert werden?
3. Welches der Mittelwertverfahren soll verwendet werden? Aufgrund der bisherigen Ergebnisse wissen wir bereits, dass der Within-Average-Linkage für eine hierarchische Darstellung ungeeignet ist.
4. Ist die 14- oder die 6-Clusterlösung stabiler?

Die Logik der Stabilitätsprüfung besteht – wie bereits in dem Anwendungsbeispiel des Abschnitts 6.6 erwähnt – darin, dass mehrere Analysen mit unterschiedlichen Modellparametern durchgeführt und die Übereinstimmungen zwischen den einzelnen Lösungen berechnet werden. Als eine Maßzahl der Übereinstimmung kann der RAND-Index (Rand 1971) verwendet werden. Er ist wie folgt definiert:

$$\text{RAND}(\text{CL}_i, \text{CL}_j) = \frac{2}{n(n-1)} \sum_g \sum_{g^* > g} r_{g,g^*}, \quad (9.3)$$

wobei $r_{g,g^*} = 1$, wenn Objekt g und g^* in beiden Clusterlösungen CL_i und CL_j demselben Cluster angehören oder in beiden Clusterlösungen unterschiedlichen Clustern angehören. Der RAND-Index misst den Anteil übereinstimmender Zuordnungen. Bei einer perfekten Übereinstimmung liegen $n(n-1)/2$ Übereinstimmungen vor und es gilt $\text{RAND} = 1$. Werte größer 0,7 können als ausreichende Übereinstimmung interpretiert werden (Dreger 1986;

Tab. 9.24: Ergebnisse der Stabilitätsprüfung für die 6- und 14-Clusterlösung (RAND-Index)

	14-Clusterlösung			6-Clusterlösung		
	Within	Average	Weight.	Within	Average	Weight.
Within-Average	1,000			1,000		
Average	0,796	1,000		0,512	1,000	
Weighted-Average	0,817	0,961	1,000	0,512	1,000	1,000
Mittelwert		0,858			0,786	

Fraboni und Saltstone 1992). Bei rein zufälligen Übereinstimmungen ergibt sich ein Wert von 0,5. Modifikationen des RAND-Index wurden unter anderem von Morey und Agresti (1984) entwickelt.

Zur Beantwortung der Frage, ob eine 6- oder 14-Clusterlösung stabiler ist, wurden die RAND-Indexwerte für 6- und 14-Clusterlösungen berechnet (siehe Tabelle 9.24). Bei der 14-Clusterlösung stimmen die Ergebnisse für alle drei Verfahren weitgehend überein. Der durchschnittliche RAND-Index beträgt 0,858 ($= (0,796 + 0,817 + 0,961)/3$). Für die 6-Clusterlösung ergibt sich zwar eine perfekte Übereinstimmung zwischen dem Average-Linkage und Weighted-Average-Linkage, aber nur eine sehr geringe zwischen diesen beiden Verfahren und dem Within-Average-Linkage. In dem Beispiel wird man die mangelnde Stabilität nicht auf die 6-Clusterlösung, sondern auf das Within-Verfahren zurückführen, da dieses zu Inversionen führt.

Zusammenfassend wird man aus den bisherigen Analysen schlussfolgern:

- Der Within-Average-Linkage ist für die untersuchten Daten nicht geeignet. Er führt zu Inversionen und instabilen Ergebnissen.
- Eine hierarchische Darstellung durch den Weighted-Average- und Average-Linkage ist möglich. Die Modellanpassung der hierarchischen Darstellung ist sehr gut.
- Den Daten liegt eine 14- bzw. 6-Clusterlösung zugrunde. Die Modellanpassung – gemessen durch den γ -Koeffizienten – ist für beide Lösungen sehr gut. Für die 14-Clusterlösung ergibt sich dagegen bei Verwendung der kophenetischen Korrelation eine schlechtere Modellanpassung. Diese ist aber zum Teil auf die Art der Berechnung des kophenetischen Korrelationskoeffizienten zurückzuführen.
- Hinsichtlich des Weighted-Average- und des Average-Linkage sind die 14- und 6-Clusterlösung stabil.
- Beide Clusterlösungen sind inhaltlich interpretierbar.

Im Anschluss an diese Zwischenbilanz wird man die Analyse fortsetzen und Validitätstests durchführen (siehe Beispiel in Abschnitt 6.6). Erst aufgrund dieser Ergebnisse kann entschieden werden, ob die gefundenen Clusterlösungen brauchbar sind. In der Literatur

(Everitt, Landau u. a. 2001, S. 182; Gordon 1999, S. 198) wird der adjustierte RAND-Index RAND^* von Hubert und Arabie (1985) empfohlen, da der gewöhnliche RAND-Index von den Randhäufigkeiten und der Zahl der Cluster abhängt und damit rein zufällig korrekte Klassifikationen nicht korrigiert. Der adjustierte RAND-Index ist definiert als:

$$\text{RAND}^*(\text{CL}_i, \text{CL}_j) = \frac{A - E(A)}{\max(A) - E(A)}, \quad (9.4)$$

mit

$$\begin{aligned} A &= \sum_{i=1}^I \sum_{j=1}^J \binom{n_{ij}}{2}, \\ E(A) &= \sum_{i=1}^I \binom{n_{i\cdot}}{2} \sum_{j=1}^J \binom{n_{\cdot j}}{2} / \binom{n}{2}, \\ \max(A) &= \left[\sum_{i=1}^I \binom{n_{i\cdot}}{2} + \sum_{j=1}^J \binom{n_{\cdot j}}{2} \right] / \binom{n}{2}. \end{aligned}$$

Steinley (2004) konnte mithilfe von Simulationsrechnungen zeigen, dass zwischen dem RAND-Index und dem adjustierten RAND-Index von Hubert und Arabie eine lineare Beziehung der folgenden Art besteht:

$$\text{RAND}^* = -1,05 + 1,82 \cdot \text{RAND}.$$

Der adjustierte RAND-Index hat eine größere Steigung und kann damit gute Lösungen besser voneinander trennen. Als Schwellenwert ergibt sich für $\text{RAND} = 0,700$ ein adjustierter RAND-Index von $\text{RAND}^* = 0,224$. Bei einer rein zufälligen Übereinstimmung ($\text{RAND} = 0,50$) ergibt sich entsprechend der Umrechnungsformel ein RAND^* -Wert von $-0,14$, der etwas kleiner als der Erwartungswert von 0 ist.

9.6 Anwendungsempfehlungen

1. Im Allgemeinen empfehlen wir Nächste-Nachbarn- oder Mittelwertverfahren für:
 - Eine Analyse einer direkt erhobenen (Un-)Ähnlichkeitsmatrix (Alternativverfahren: mehrdimensionale Skalierung, siehe Kapitel 4).
 - Eine hierarchische Darstellung von (Un-)Ähnlichkeitsbeziehungen (alternative räumliche Darstellung mittels mehrdimensionaler Skalierung oder anderer unvollständiger Clusteranalyseverfahren, siehe Teil 1).

- Eine variablenorientierte Clusteranalyse (Alternativverfahren: multiple Korrespondenzanalyse, siehe Kapitel 4, Faktorenanalyse, siehe Kapitel 5 oder nominale Faktorenanalyse nach McDonald, siehe ebenfalls Kapitel 5).
 - Eine objektorientierte Analyse, sofern die Anwendungsvoraussetzungen für das Ward- oder K-Means-Verfahren nicht erfüllt sind. Dies ist dann der Fall, wenn der Einsatz der quadrierten euklidischen Distanz nicht sinnvoll ist (Modifikationen dazu siehe Abschnitt 12.14). Beispielsweise wenn größeren Distanzen nicht ein größeres Gewicht als vielen kleinen Distanzen gegeben werden soll, wenn Korrelationen als Ähnlichkeitsmaße verwendet oder wenn, wie bei einer Inhaltsanalyse, die Nullausprägungen anders behandelt werden sollen als die Einsausprägungen (zur Robustheit siehe Abschnitt 8.1).
2. Von den Nächste-Nachbarn- oder Mittelwertverfahren empfehlen wir den Weighted-Average-Linkage (siehe Abschnitt 9.5) aus folgenden Gründen:
 - Er vermeidet die negativen Effekte der Alternativverfahren Complete- und Single-Linkage (siehe Abschnitte 9.1 und 9.2), nämlich Dilatations- und Kontraktionseffekte (siehe Abschnitt 6.3).
 - Er steht im Unterschied zum Complete-Linkage für überlappende Cluster (siehe Abschnitt 9.3) und den verallgemeinerten Nächste-Nachbarn-Verfahren (siehe Abschnitt 9.4) in den meisten Softwareanwendungen zur Verfügung.
 - Sein Verschmelzungsschema ermöglicht im Unterschied zum Average-Linkage eine klare Interpretation (siehe Abschnitt 9.5).
 - Er vermeidet zudem Inversionen im Verschmelzungsschema, die beim Within-Average-Linkage auftreten können (siehe Abschnitt 9.5).
 3. Wir empfehlen den Weighted-Average-Linkage nicht, wenn
 - die Matrix nur ordinalen Informationsgehalt hat und Invarianz gegenüber monotonen Transformationen somit erwünscht ist, oder
 - wenn Überlappungen vermutet werden.
 4. Bei erwünschter Invarianz gegenüber monotonen Transformationen empfehlen wir den Complete-Linkage (siehe Abschnitt 9.1), da die Alternative, der Single-Linkage (siehe Abschnitt 9.2), zu Verkettungen führt.
 5. Für Überlappungen kann der Complete-Linkage für Überlappungen eingesetzt werden (siehe Abschnitt 9.3).
 6. Für die Bestimmung der Clusterzahl haben wir mit dem Kriterium der Zunahme im Verschmelzungsschema die besten Erfahrungen gemacht (siehe Abschnitt 9.1.3). Ergänzend sollten die Mojena-Kriterien und eine Zufallstestung verwendet werden (siehe Abschnitte 9.1.3 und 9.1.4).
 7. Zur Prüfung der Modellanpassung sollte bei den Mittelwertverfahren die kophenetische Korrelation (siehe Abschnitt 9.5) und bei den Nächste-Nachbarn-Verfahren der

- γ -Koeffizient (siehe Abschnitt 9.1.5) eingesetzt werden (siehe auch Abschnitt 9.1.2). Zur Interpretation sei auf die Schwellenwerte aus Tabelle 3.20 auf Seite 74 verwiesen.
8. Untersucht werden sollte auch die Stabilität einer gefundenen Clusterlösung. Dabei können entweder der RAND-Index oder der adjustierte RAND-Index eingesetzt werden, da sie linear ineinander überführbar sind. Der adjustierte RAND-Index ermöglicht eine feinere Differenzierung von sehr guten Lösungen. Als brauchbar können Lösungen mit einem Wert beim RAND-Index von über 0,700 und beim adjustierten RAND-Index von über 0,224 erachtet werden (siehe Abschnitt 9.5).

10 Repräsentanten-Verfahren

10.1 Modellansatz

Das Ziel der *Repräsentanten-Verfahren* ist, K Cluster durch jeweils ein typisches (repräsentatives) Objekt zu charakterisieren. Dieses typische Objekt wird in der Literatur als *Repräsentant*, *Pivotelement*, *Clusterkern*, »*centrotype*«, »*medoid*« oder »*leading case*« bezeichnet (Hartigan 1975, S. 74–83; Kaufman und Rousseeuw 1990, S. 69; Lorr und Radhakrishnan 1967; Sawrey, Keller u. a. 1960; Tryon 1958). Wir werden in diesem Kapitel ein einfaches Repräsentanten-Verfahren¹ darstellen, das von folgenden Modellannahmen ausgeht:

1. Es gibt K Cluster, die der Größe nach geordnet sind:

$$n_1 \geq n_2 \geq \dots \quad (n_k = \text{Größe des Clusters } k).$$

2. Ein Objekt g soll einem Cluster k angehören, wenn es zu dem Repräsentanten r_k des Clusters k eine Unähnlichkeit kleiner/gleich bzw. eine Ähnlichkeit größer/gleich einem bestimmten Schwellenwert für die Clusterhomogenität besitzt. Dieser Schwellenwert soll mit U_{homo} für Unähnlichkeiten bzw. mit \ddot{A}_{homo} für Ähnlichkeiten bezeichnet werden. Ein Objekt g muss somit folgende Bedingung erfüllen, damit es einem Cluster k angehört:

$$u_{g,r_k} \leq U_{\text{homo}} \quad \text{bzw.} \quad \ddot{a}_{g,r_k} \geq \ddot{A}_{\text{homo}}, \quad (10.1)$$

mit u_{g,r_k} als Unähnlichkeit zwischen Objekt g und dem Repräsentanten r_k des Clusters k bzw. \ddot{a}_{g,r_k} als Ähnlichkeit zwischen Objekt g und dem Repräsentanten r_k .

3. Der Repräsentant r_k eines Clusters k soll jenes Objekt mit den meisten nächsten Nachbarn sein. Ein nächster Nachbar soll dann vorliegen, wenn die Unähnlichkeit kleiner/gleich bzw. die Ähnlichkeit größer/gleich dem Schwellenwert für die Clusterhomogenität ist. Ferner wird gefordert, dass die Repräsentanten voneinander verschieden sind: Die Unähnlichkeit bzw. Ähnlichkeit zwischen den Repräsentanten soll größer

¹ Weitere Repräsentanten-Verfahren werden in Abschnitt 10.4 behandelt.

bzw. kleiner einem weiteren Schwellenwert für die Clusterheterogenität sein. Bezeichnen wir diese Schwellenwerte mit $U_{\text{hetero}} (\geq U_{\text{homo}})$ bzw. $\ddot{A}_{\text{hetero}} (\leq \ddot{A}_{\text{homo}})$, so soll gelten:

$$u_{r_k, r_{k^*}} > U_{\text{hetero}} \quad \text{bzw.} \quad \ddot{a}_{r_k, r_{k^*}} < \ddot{A}_{\text{hetero}} \quad \forall r_k, r_{k^*}. \quad (10.2)$$

Die Schwellenwerte für die Clusterheterogenität haben die Funktion, die Zahl der Überlappungen zu steuern. Je geringer der Unterschied zwischen den Schwellenwerten U_{hetero} und U_{homo} bzw. \ddot{A}_{hetero} und \ddot{A}_{homo} ist, desto größer ist der Anteil der Überlappungen.

4. Grundsätzlich sind aber Überlappungen und Nichtklassifikationen erlaubt.

Das Verfahren unterscheidet sich von den bisher behandelten Verfahren unter anderem darin, dass nicht das Auffinden einer Klassifikation im Vordergrund steht, sondern das Auffinden eines »typischen« Objekts (Repräsentant) für jedes Cluster.

Der Algorithmus zum Auffinden der Repräsentanten ist:

Schritt 1: Berechnung der Zahl nn_g nächster Nachbarn für jedes Objekt g . Zusätzlich wird für jedes Objekt g die durchschnittliche (mittlere) Unähnlichkeit uu_g bzw. die durchschnittliche (mittlere) Ähnlichkeit \ddot{a}_g der nächsten Nachbarn zum Objekt g berechnet. Die mittleren Ähnlichkeiten bzw. Unähnlichkeiten werden für die Größenanordnung benötigt, wenn zwei Objekte dieselbe Anzahl von nächsten Nachbarn haben. In die Berechnung beider Kennwerte wird auch die Ähnlichkeit bzw. Unähnlichkeit des Objekts zu sich selbst einbezogen. Dadurch können auch einzelne Objekte als Cluster identifiziert werden. Jedes Objekt besitzt somit immer mindestens einen nächsten Nachbarn, nämlich sich selbst.

Schritt 2: Größenanordnung der Objekte. Die Objekte g werden entsprechend der Zahl nächster Nachbarn absteigend mit $nn_g \geq nn_{g^*} \geq nn_{g^{**}} \geq \dots$ angeordnet. Bei Gleichheit ($nn_g = nn_{g^*}$) werden die Objekte entsprechend den durchschnittlichen Unähnlichkeiten oder Ähnlichkeiten angeordnet. Gilt also zum Beispiel $nn_g = nn_{g^*}$ und $uu_g < uu_{g^*}$, wird das Objekt g vor dem Objekt g^* gereiht, da g als möglicher Repräsentant homogener ist. Die erwähnte Invarianzeigenschaft des Repräsentanten-Verfahrens gegenüber monotonen Transformationen gilt also nur, wenn alle Objekte eine unterschiedliche Zahl von nächsten Nachbarn haben.

Schritt 3: Berechnung der Repräsentanten. Die Berechnung der Repräsentanten erfolgt hierarchisch, wobei das erste Objekt in der geordneten Objektliste immer ein Repräsentant (für das Cluster 1) ist. Die weiteren Repräsentanten werden wie folgt bestimmt: Es wird geprüft, ob das zweite Objekt der geordneten Objektliste ein Repräsentant ist, ob also die Bedingung 10.2 erfüllt ist. Ist dies der Fall, wird das zweite Objekt der Repräsentant eines neuen Clusters (Cluster 2). Die Prüfung wird

Tab. 10.1: Repräsentanten für das Entwicklungsländerbeispiel

Repräsentant	Objektnummer	ExGrundBed ^{a)}	wirtWachst ^{a)}	Indust. ^{a)}	Bildung ^{a)}
Panama	17	0,699	0,124	0,366	0,724
Honduras	11	-0,901	0,835	-0,806	-0,879
Ecuador	7	-0,671	1,785	0,231	0,133
Haiti	10	-2,041	-2,390	-1,264	-2,975
Trinidad	20	0,962	-0,233	2,913	0,143
Jamaika	12	0,816	-2,424	-0,110	-0,029

a) zur Bildung der Indikatoren siehe Abschnitt 6.6.

mit dem nächsten Objekt fortgesetzt. Ist das zweitgrößte Objekt kein Repräsentant, wird die Prüfung der Bedingung 10.2 unmittelbar mit dem nächsten Objekt fortgesetzt. Die Berechnung der Repräsentanten wird abgeschlossen, wenn alle Objekte geprüft sind.

Schritt 4: Zuordnung der Objekte. Nach der Bildung der Repräsentanten werden die Objekte entsprechend der Bedingung 10.1 auf Seite 277 den Clustern zugeordnet. Dabei können drei Ergebnisse eintreten: Das Objekt g gehört nur einem Cluster, zwei oder mehreren Clustern oder keinem Cluster an.

10.2 Anwendungsbeispiel

Der dargestellte Algorithmus soll für die Entwicklungsländerdaten veranschaulicht werden. Ziel der Analyse ist eine Clusterbildung für die 22 Länder Mittel- und Südamerikas aufgrund ihrer Faktorwerte in den vier berechneten Faktoren (siehe Abschnitt 6.6). Als Unähnlichkeitsmaß soll die mittlere City-Block-Metrik $\frac{1}{m} \sum |x_{gi} - x_{g^*i}|$ verwendet werden.² Für die Homogenität U_{homo} in den Clustern wurde ein Schwellenwert von 0,5 angenommen, für die Heterogenität U_{hetero} zwischen den Clustern ein Wert von 0,7. Die Unähnlichkeiten zwischen den Objekten eines Clusters k zu dem Clusterrepräsentanten sollen also kleiner/gleich 0,5 und die Unähnlichkeiten zwischen den Repräsentanten sollen größer 0,7 sein. Da mit Unähnlichkeiten gerechnet wird, soll zur Vereinfachung der Schreibweise nur von Unähnlichkeiten gesprochen werden. Bei diesen Vorgaben werden sechs Cluster und die in Tabelle 10.1 ausgewiesenen Repräsentanten ermittelt.

Das erste Cluster wird durch Panama repräsentiert. Panama als Repräsentant ist durch ein mittleres wirtschaftliches Wachstum und eine mittlere Industrialisierung gekenn-

² Dies hat für die Bestimmung der Schwellenwerte U_{homo} und U_{hetero} bzw. \ddot{A}_{homo} und \ddot{A}_{hetero} bestimmte Vorteile (siehe Abschnitt 10.3).

zeichnet. Die Befriedigung existentieller Bedürfnisse und das erreichte Bildungsniveau sind im Vergleich zu den anderen Ländern leicht überdurchschnittlich.³ Honduras als Repräsentant des Clusters 2 unterscheidet sich von Panama durch eine unter dem Gesamtdurchschnitt liegende Befriedigung existentieller Grundbedürfnisse. Die erreichte Industrialisierung und Bildung liegt ebenfalls unter dem Gesamtdurchschnitt, während das Wirtschaftswachstum über dem Gesamtdurchschnitt liegt. Die anderen Cluster sind analog zu interpretieren.

Im letzten Schritt des Algorithmus wird jedes Objekt g dem oder den Clustern zugeordnet, zu dessen Repräsentant bzw Repräsentanten die Unähnlichkeit bzw. Unähnlichkeiten kleiner 0,5 sind. Es ergibt sich das in der Tabelle 10.2 dargestellte Bild. Dem ersten Cluster gehören somit neben dem Repräsentanten (Panama) an: Argentinien, Brasilien, Chile, Costa Rica, Kolumbien, Kuba und Mexiko, dem zweiten Cluster die Dominikanische Republik, El Salvador, Guatemala und Honduras. Die anderen Cluster werden jeweils nur von ihren Repräsentanten gebildet.

Insgesamt können 16 der 22 Länder (62,7 Prozent) geclustert werden. Die Nichtklassifikationen (27,3 Prozent) entstehen dadurch, dass zum einen die nichtklassifizierten Objekte keinem Cluster zugeordnet werden können, da die Unähnlichkeit zu keinem Repräsentanten kleiner/gleich dem vorgegebenen Schwellenwert von 0,5 für die Clusterhomogenität ist. Auf der anderen Seite ist die Unähnlichkeit zu den Repräsentanten aber kleiner dem vorgegebenen Schwellenwert von 0,7 für die Clusterheterogenität, so dass diese Objekte auch kein eigenständiges Cluster bilden.

Zur graphischen Darstellung der Ergebnisse kann das beim Complete-Linkage für überlappende Cluster dargestellte Vorgehen verwendet werden: Mit einem dimensionalen Analyseverfahren (nichtmetrische mehrdimensionale Skalierung, Hauptkomponentenanalyse usw.) wird eine zweidimensionale Darstellung berechnet. In diese wird die gefundene Clusterlösung eingetragen (siehe Abschnitt 9.3 und Abbildung 9.5 auf Seite 258).

10.3 Die Wahl der Schwellenwerte

Die Ergebnisse des Repräsentanten-Verfahrens hängen von der Wahl der Schwellenwerte der Homogenität in den Clustern (U_{homo}) und der Heterogenität zwischen den Clustern (U_{hetero}) ab. Allgemein gilt: Je größer der Schwellenwert U_{homo} gewählt wird, desto

³ Zur Erinnerung: Da mit Faktorwerten gerechnet wird, sind nur relative Aussagen in Bezug auf die untersuchte Population möglich.

Tab. 10.2: Zuordnung der Objekte zu den Clustern

Land	Objektnr.	C_1	C_2	C_3	C_4	C_5	C_6
Argentinien	1	1	0	0	0	0	0
Bolivien	2	0	0	0	0	0	0
Brasilien	3	1	0	0	0	0	0
Chile	4	1	0	0	0	0	0
Costa Rica	5	1	0	0	0	0	0
Dom. Rep.	6	0	1	0	0	0	0
Ecuador	7	0	0	1*	0	0	0
El Salvador	8	0	1	0	0	0	0
Guatemala	9	0	1	0	0	0	0
Haiti	10	0	0	0	1*	0	0
Honduras	11	0	1*	0	0	0	0
Jamaika	12	0	0	0	0	0	1*
Kolumbien	13	1	0	0	0	0	0
Kuba	14	1	0	0	0	0	0
Mexiko	15	1	0	0	0	0	0
Nicaragua	16	0	0	0	0	0	0
Panama	17	1*	0	0	0	0	0
Paraguay	18	0	0	0	0	0	0
Peru	19	0	0	0	0	0	0
Trinidad	20	0	0	0	0	1*	0
Uruguay	21	0	0	0	0	0	0
Venezuela	22	0	0	0	0	0	0
n		8	4	1	1	1	1

»*«: Repräsentant des jeweiligen Clusters

grau hinterlegt: Zuordnungen

weniger, aber dafür inhomogenere Cluster werden berechnet. Je größer der Schwellenwert U_{hetero} im Vergleich zum Schwellenwert U_{homo} ist, desto mehr Nichtklassifikationen, aber weniger Überlappungen treten auf. In der Forschungspraxis wird man daher $U_{\text{homo}} = U_{\text{hetero}}$ wählen, wenn alle Objekte klassifiziert werden sollen. Ist man dagegen an gut getrennten Clustern interessiert, wird man für den Schwellenwert für die Clusterheterogenität einen größeren Wert wählen, zum Beispiel $U_{\text{hetero}} = 1,5 \cdot U_{\text{homo}}$. In der Regel wird man mehrere Konstellationen für die Schwellenwerte ausprobieren. Bei der Auswahl kann man sich an inhaltlichen und formalen Kriterien orientieren. So zum Beispiel kann gefordert werden, dass die mittlere Unähnlichkeit der Objekte zu ihren Repräsentanten nicht größer einer halben Skaleneinheit sein soll. Ist die Skaleneinheit beispielsweise – wie in unserem Beispiel – gleich 1, wird man für U_{homo} einen Wert von 0,5 wählen und abhängig davon, ob alle Objekte klassifiziert werden sollen, für den Schwellenwert U_{hetero} Werte von 0,5 (alle Objekte werden klassifiziert), 0,7 und 1,0 (Nichtklassifikationen, aber keine Überlappungen) wählen.

Neben der Beschaffenheit der Skala der untersuchten Variablen können Signifikanzüberlegungen, sofern sie für das ausgewählte Unähnlichkeitsmaß möglich sind, verwendet werden. So zum Beispiel wählen Lorr und Radhakrishnan (1967) die Schwellenwerte so, dass U_{homo} gleich dem kritischen Wert für eine Signifikanz von 95 Prozent und U_{hetero} gleich dem kritischen Wert für eine Signifikanz von 99 Prozent ist ($p < 0,05$ bzw $0,01$).

10.4 Weitere Repräsentanten-Verfahren

Das hier dargestellte Repräsentanten-Verfahren lässt sich in mehrfacher Hinsicht weiterentwickeln und modifizieren. Lorr und Radhakrishnan (1967) verwenden beispielsweise nicht die Unähnlichkeit eines Objekts zu seinem Repräsentanten als Zuordnungskriterium, sondern führen solange eine Zuordnung durch, bis die mittlere (Un-)Ähnlichkeit einen bestimmten Schwellenwert über- bzw. unterschreitet. Als Algorithmus dargestellt, sieht das Verfahren folgendermaßen aus:

- Schritt 1:* Betrachte alle Objekte als klassifizierbar und setze die Clusterzahl $k = 0$.
- Schritt 2:* Berechne die Zahl nächster Nachbarn für jedes klassifizierbare Objekt g .
- Schritt 3:* Wähle das Objekt g mit den meisten nächsten Nachbarn als Repräsentanten aus. Erhöhe den Clusterzähler um 1 ($k = k + 1$) und ordne das Objekt g dem Cluster k zu. Betrachte für die weiteren Schritte das Objekt g als nicht mehr klassifizierbar.
- Schritt 4:* Wähle das Objekt g^* mit der geringsten durchschnittlichen Unähnlichkeit zum Cluster k aus. Prüfe, ob dieses kleiner dem vorgegebenen Schwellenwert U_{homo} ist. Bei »nein« gehe zu Schritt 6. Bei »ja« ordne das Objekt g^* dem Cluster k zu und streiche es aus der Liste der klassifizierbaren Objekte. Gehe zu Schritt 5.
- Schritt 5:* Wiederhole Schritt 4.
- Schritt 6:* Streiche alle Objekte mit einer durchschnittlichen Unähnlichkeit zwischen U_{homo} und U_{hetero} zu den gebildeten Clustern aus der Liste der klassifizierbaren Objekte. Dieser Schritt wird durchgeführt, um Überlappungen zu vermeiden.
- Schritt 7:* Wiederhole die Schritte 2 bis 6 solange, bis ein Cluster mit weniger als vier Objekten entsteht.

Tryon (1958) geht in seiner »Cumulative Communality Cluster Analysis« ähnlich vor. Im Unterschied zu Lorr und Radhakrishnan verwendet er aber nur die Ähnlichkeit bzw. Unähnlichkeit der Objekte zu den berechneten Repräsentanten, die wie bei Lorr und Radhakrishnan als Pivotelemente bezeichnet werden. Gemeinsames Merkmal beider Verfahren ist, dass nur Nichtklassifikationen, aber keine Überlappungen auftreten können.

Die Grundlogik der Repräsentanten-Verfahren haben Kaufman und Rousseeuw (1990) zur Entwicklung eines umfassenden Klassifikationsansatzes verwendet und bezeichnen ihr Repräsentanten-Verfahren als Medoid-Verfahren. Es unterscheidet sich von den in diesem Abschnitt behandelten Verfahren dadurch, dass keine Schwellenwerte definiert werden müssen, sondern eine Clusterzahl vorgegeben wird. Das Verfahren berechnet eine überlappungsfreie Klassifikation aller Objekte. Einen Überblick über weitere Repräsentanten-Verfahren geben Jain und Dubes (1988, S. 128–129) sowie Kaufman und Rousseeuw (1990, S. 108–111).

10.5 Anwendungsempfehlungen

1. Repräsentanten-Verfahren sind attraktive Verfahren, da sie auch ohne großen mathematischen Sachverstand angewendet werden können.
2. Es liegen zwar Software-Implementierungen vor, wie zum Beispiel in ALMO oder R, unser Erfahrungswissen zur Brauchbarkeit ist allerdings für Praxisempfehlungen noch nicht ausreichend.
3. Problematisch ist, dass vom Anwender bzw. von der Anwenderin zumeist Schwellenwerte definiert werden müssen, für die derzeit noch keine formalen Begründungen vorliegen.

11 Hierarchische Verfahren zur Konstruktion von Clusterzentren

11.1 Modellansätze, Algorithmen und Ward-Verfahren

Während die im vorausgehenden Kapitel 10 behandelten Repräsentanten-Verfahren das Ziel verfolgen, ein typisches Objekt für jedes Cluster zu bestimmen, verwenden die in diesem und dem nachfolgenden Kapitel 12 dargestellten Verfahren *Clusterzentren* als Repräsentanten. Die *Clusterzentren* sind die Mittelwerte der Cluster, genauer der Objekte der Cluster, in den in die Analyse einbezogenen Klassifikationsvariablen. Da Mittelwerte berechnet werden, sind die Verfahren strenggenommen nur bei quantitativen Variablen anwendbar. Da ordinale und dichotome Variablen wie quantitative behandelt werden können¹, sind die Verfahren auch für ordinale und dichotome Variablen brauchbar (siehe Abschnitte 8.2 und 8.4). Nominale Variablen schließlich können durch eine Dummy-Auflösung »quantifiziert« werden (siehe dazu Abschnitte 7.5 und 8.3).

Bei den in diesem Abschnitt behandelten Verfahren werden die Clusterzentren hierarchisch mit folgendem Algorithmus konstruiert:

Schritt 1: Jedes Objekt g bildet zu Beginn ein selbständiges Cluster C_g mit den Clusterzentren $\bar{x}_{gj} = x_{gj}$ ($g = 1, 2, \dots, n; j = 1, 2, \dots, m$; x_{gj} entspricht der Ausprägung des Objekts g in der Variablen j).

Schritt 2: Es wird jenes Clusterpaar $\{p, q\}$ gesucht, für das a) die euklidische Distanz der Clusterzentren minimal ist oder b) das bei einer Verschmelzung zu einer minimalen Zunahme der Streuungsquadratsumme in den Clustern führt. Nach der Forderung a) gehen das *Median-* und *Zentroid-Verfahren* vor, nach der Forderung b) das *Ward-Verfahren*.

¹ Für dichotome Variablen ist dies formal zulässig, für ordinale aufgrund der Forschungspraxis.

Schritt 3: Das Clusterpaar $\{p, q\}$ wird verschmolzen, und die Clusterzentren werden neu berechnet. Beim Zentroid- und Ward-Verfahren erfolgt die Neuberechnung nach folgender Formel:

$$\bar{x}_{\{p+q\},j} = \sum_{g \in p+q} \frac{x_{gj}}{n_p + n_q},$$

wobei $\bar{x}_{(p+q),j}$ der Mittelwert des neu gebildeten Clusters $(p + q)$ in der Variablen j ist. Beim Median-Verfahren werden die neuen Clusterzentren dagegen als Mediane aus den alten Clusterzentren berechnet:

$$\bar{x}_{\{p+q\},j} = \frac{(\bar{x}_{pj} + \bar{x}_{qj})}{2}.$$

Schritt 4: Die Schritte werden so lange wiederholt, bis alle Objekte einem einzigen Cluster angehören.

Tatsächlich ist eine Neuberechnung der Clusterzentren in jedem Verschmelzungsschritt nicht erforderlich. Es kann der Algorithmus der hierarchisch-agglomerativen Verfahren angewendet werden, wenn quadrierte euklidische Distanzen verwendet und die Unähnlichkeiten im vierten Schritt wie folgt berechnet werden (siehe dazu zum Beispiel Steinhausen und Langer 1977, S. 77):

Median-Verfahren:

$$u_{\{p+q\},i}^{\text{neu}} = \frac{1}{2} \cdot u_{p,i} + \frac{1}{2} \cdot u_{q,i} - \frac{1}{4} \cdot u_{p,q}, \quad (11.1)$$

wobei $u_{p,i}$ die Unähnlichkeit (quadrierte euklidische Distanz) zwischen den Clustern p und i , $u_{q,i}$ jene zwischen den Clustern q und i sowie $u_{p,q}$ jene zwischen den Clustern p und q ist.

Zentroid-Verfahren:

$$u_{\{p+q\},i}^{\text{neu}} = \frac{n_p}{n_p + n_q} \cdot u_{p,i} + \frac{n_q}{n_p + n_q} \cdot u_{q,i} - \frac{n_p \cdot n_q}{(n_p + n_q)^2} \cdot u_{p,q}. \quad (11.2)$$

Ward-Verfahren:

$$u_{\{p+q\},i}^{\text{neu}} = \frac{1}{n_p + n_q + n_i} \cdot [(n_p + n_i) \cdot u_{p,i} + (n_q + n_i) \cdot u_{q,i} - n_i \cdot u_{p,q}]. \quad (11.3)$$

Die Beziehung zwischen den Neuberechnungsformeln 11.1 bis 11.3 zu den dargestellten Modellansätzen wird beispielsweise in Kaufman und Rousseeuw (1990, S. 230–233) nachgewiesen. Für das *Ward-Verfahren* gilt dabei, dass das angegebene Verschmelzungsniveau

Tab. 11.1: (Fiktive) Datenmatrix zur Veranschaulichung der Eigenschaften des Ward-Verfahrens

		Datenmatrix		Matrix der quadrierten euklidischen Distanzen								
		X_1	X_2	A	B	C	D	E	F	G	H	I
A	-2	1		0								
B	-1	2		2	0							
C	-1	-2		10	16	0						
D	0	-1		8	10	2	0					
E	1	-1		13	13	5	1	0				
F	2	2		17	9	25	13	10	0			
G	3	2		26	16	32	18	13	1	0		
H	4	2		37	25	41	25	18	4	1	0	
I	4	3		40	26	50	32	25	5	2	1	0

V_i gleich zweimal dem Zuwachs in der Streuungsquadratsumme innerhalb der Cluster beim Übergang vom $(i - 1)$ -ten zum i -ten Schritt ist:

$$V_i = u_{p,q} = 2 \cdot \Delta s_{\text{Qin}}(i),$$

wobei $\Delta s_{\text{Qin}}(i)$ der Zuwachs in der Streuungsquadratsumme innerhalb der Cluster ist.

Wir wollen uns diese Beziehung anhand der fiktiven Datenmatrix der Tabelle 11.1 verdeutlichen. Den Daten liegt eine 3-Clusterstruktur zugrunde. Das Cluster 1 wird von den Objekten {A} und {B}, das Cluster 2 von den Objekten {C}, {D} und {E} und das Cluster 3 von den verbleibenden Objekten {F} bis {I} gebildet. In die Tabelle wurde auch die Matrix der quadrierten euklidischen Distanzen aufgenommen. Entsprechend dem allgemeinen Grundprinzip der hierarchisch-agglomerativen Verfahren wird im ersten Verschmelzungsschritt das Clusterpaar (Objektpaar) mit der kleinsten Unähnlichkeit ausgewählt. In dem Beispiel kommen mehrere Paare in Frage. Es liegen sogenannte *Bindungen* vor, die das Ergebnis beeinflussen können. Nur beim Single-Linkage spielt die Behandlung der Bindungen keine Rolle. Wir wollen hier das letzte Objektpaar {H, I} auswählen. Das Verschmelzungsniveau ist gleich 1. Das Clusterzentrum des neugebildeten Clusters ist:

$$\bar{x}_{\{H+I\},1} = (4 + 4)/2 = 4 \quad \text{und} \quad \bar{x}_{\{H+I\},2} = (3 + 2)/2 = 2,5.$$

Die Streuungsquadratsumme innerhalb des neugebildeten Clusters ist gleich:

$$\begin{aligned} s_{\text{Qin}}(1) &= s_{\text{Qin}}(X_1|1) + s_{\text{Qin}}(X_2|1) \\ &= (4 - 4)^2 + (4 - 4)^2 + (3 - 2,5)^2 + (2 - 2,5)^2 = 0,5. \end{aligned}$$

Da die Streuungsquadratsumme in den Clustern zu Beginn des Verschmelzungsprozesses gleich 0 ist ($s_{\text{Qin}}(0) = 0$) – jedes Cluster besteht nur aus einem Objekt –, ist die Zunahme

Tab. 11.2: Verschmelzungsschema des Ward-Verfahrens für die fiktive Datenmatrix der Tabelle 11.1 auf der vorherigen Seite

Schritt	Anzahl Cluster	Verschmelzung der Cluster mit den Objekten	Verschmelzungsniveau	Streuungsquadratsumme in den Clustern	erklärte Streuung in % (η_k^2)
1	8	9 – 8	1,000	0,500	99
2	7	6 – 7	1,000	1,000	98
3	6	4 – 5	1,000	1,500	98
4	5	1 – 2	2,000	2,500	96
5	4	3 – 4	4,333	4,667	93
6	3	6 – 8	5,000	7,167	89
7	2	1 – 3	24,567	19,500	70
8	1	1 – 6	92,556	65,780	0

der Streuungsquadratsumme $\Delta SQ_{in}(1)$ gleich 0,5. Das angegebene Verschmelzungsniveau von 1 ist somit zweimal so hoch. Insgesamt ergibt sich für das Beispiel das in der Tabelle 11.2 dargestellte Verschmelzungsschema. Die Objekte sind wiederum fortlaufend mit 1 beginnend nummeriert (1 = A, 2 = B usw.).

Bei der 2-Clusterlösung ist das Verschmelzungsniveau gleich 24,567. Die Streuungsquadratsumme innerhalb der Cluster erhöht sich um 12,2835 (= 24,567/2) gegenüber der 3-Clusterlösung. Aus dem Verschmelzungsschema kann wegen der Beziehung

$$SQ_{in}(i) = SQ_{in}(i-1) + \Delta SQ_{in}(i) = SQ_{in}(i-1) + V_i/2$$

die Streuungsquadratsumme in den Clustern für jeden Verschmelzungsschritt bzw. für jede Clusterlösung berechnet werden. Da $SQ_{in}(0)$ gleich 0, ist $SQ_{in}(1) = 0 + 1/2 = 0,5$ für die 8-Clusterlösung (erster Schritt). Für den zweiten Schritt (7-Clusterlösung) ergibt sich ein Wert von $SQ_{in}(2) = 0,5 + 1/2 = 1$, für den dritten Schritt von $SQ_{in}(3) = 1,0 + 1/2 = 1,5$ usw. Die entsprechenden Werte sind in der Tabelle 11.2 eingetragen. Für den letzten Schritt, bei dem alle Objekte einem einzigen Cluster angehören, ergibt sich die Gesamtfehlerstreuung SQ_{ges} . In unserem Beispiel ist diese gleich 65,780. Setzen wir diese Größe mit den Streuungsquadratsummen der einzelnen Verschmelzungsschritte in Beziehung, kann die durch eine bestimmte Clusterlösung erklärte Streuung (η_k^2) mit

$$\eta_k^2 = 1 - \frac{SQ_{in}(k)}{SQ_{ges}}$$

berechnet werden. In unserem Beispiel erklärt die 8-Clusterlösung 99 Prozent, die 7-Clusterlösung 98 Prozent usw.

Tab. 11.3: Verschmelzungsschemas für das Median- und Zentroid-Verfahren für die fiktive Datenmatrix der Tabelle 11.1 auf Seite 287

Schritt	Anzahl Cluster	Verschmelzung der Cluster mit den Objekten	Verschmelzungsniveau	Zunahme
<i>Median-Verfahren</i>				
1	8	8 – 9	1,000	0,000
2	7	6 – 7	1,000	0,000
3	6	4 – 5	1,000	0,000
4	5	1 – 2	2,000	1,000
5	4	6 – 8	2,500	0,500
6	3	3 – 4	3,250	0,750
7	2	1 – 3	10,563	7,313
8	1	1 – 6	22,078	11,516
<i>Zentroid-Verfahren</i>				
1	8	8 – 9	1,000	0,000
2	7	6 – 7	1,000	0,000
3	6	4 – 5	1,000	0,000
4	5	1 – 2	2,000	1,000
5	4	6 – 8	2,500	0,500
6	3	3 – 4	3,250	0,750
7	2	1 – 3	10,278	7,028
8	1	1 – 6	20,825	10,547

Wird die fiktive Datenmatrix mit den anderen beiden Verfahren untersucht, ergeben sich die in Tabelle 11.3 dargestellten Verschmelzungsschemas. Die Verschmelzungsniveaus sind abhängig vom gewählten Verfahren unterschiedlich zu interpretieren:

Zentroid-Verfahren: Die quadrierte euklidische Distanz zwischen den Clusterzentren der im i -ten Schritt verschmolzenen Cluster ist gleich dem Verschmelzungsniveau V_i . Im vorletzten Schritt (2-Clusterlösung) werden also zwei Cluster mit einer quadrierten euklidischen Distanz von 10,278 zwischen den Clusterzentren verschmolzen. Im Vergleich zur vorausgehenden Verschmelzung stellt dies eine Zunahme um 7,028 dar.

Median-Verfahren: Das Verschmelzungsschema ist – ähnlich wie beim Average-Verfahren – schwieriger zu interpretieren. Es gibt die quadrierte euklidische Distanz zwischen den Clusterzentren an, wenn im gesamten Verschmelzungsprozess die Cluster als gleich groß betrachtet werden.

11.2 Bestimmung der Clusterzahl und Modellprüfgrößen

Zur Auswahl einer oder mehrerer bestimmter Clusterlösungen können wiederum die bereits bekannten Maßzahlen und Techniken (wie zum Beispiel inverser Scree-Test, siehe Abschnitt 9.1.3) eingesetzt werden. Dies gilt auch für Maßzahlen zur Beschreibung der Homogenität einer ausgewählten Clusterlösung. Darüber hinaus können varianzanalytische Maßzahlen zur Beschreibung der Homogenität einer Clusterlösung verwendet werden. Beim Ward-Verfahren können diese direkt aus dem Verschmelzungsschema berechnet werden, bei den anderen beiden Verfahren nachträglich aufgrund der Ergebnisse. Die Definition und Berechnung von varianzanalytischen Maßzahlen wird im Kapitel 12 ausführlich behandelt. Wir wollen nachfolgend am Beispiel des Ward-Verfahrens darstellen, wie sich die hierarchisch-agglomerativen Verfahren zur Analyse großer Datensätze einsetzen lassen. Die Ausführungen gelten allgemein für alle hierarchisch-agglomerativen Verfahren.

11.3 Analyse durchschnittlicher Befragter

Das in der Vergangenheit bedeutsame Problem, dass hierarchisch-agglomerative Verfahren zur Analyse großer Datensätze aufgrund mangelnder Rechenleistung ungeeignet waren², erscheint aus heutiger Sicht weniger gewichtig. Mittlerweile sind allerdings andere Probleme in den Fokus gerückt:

Bindungen: Bei großen Datensätzen können Bindungen dergestalt auftreten, dass zwei oder mehrere Objekt- bzw. Clusterpaare in einem Verschmelzungsschritt dieselbe Unähnlichkeit ($d_{ij} = d_{i^*j^*} \dots$) zueinander aufweisen. Der Algorithmus muss dann (willkürlich) entscheiden, welches Objekt- bzw. Clusterpaar ausgewählt wird. Erfolgt diese Entscheidung in einer frühen Stufe des Verschmelzungsprozesses, kann dies die Ergebnisse gravierend beeinflussen. Da die Verschmelzung der Objekte in einem Schritt »endgültig« ist, beeinflusst sie die nachfolgenden Verschmelzungsschritte. Die Ergebnisse werden sich also unterscheiden, abhängig davon, welches Objektpaar ausgewählt wird.

² Zwar muss die Unähnlichkeitsmatrix zwischen allen Objekten bzw. Variablen nach wie vor im Arbeitsspeicher gehalten werden, die heutigen »Rechenmaschinen« weisen jedoch zumeist eine ausreichende Speicherkapazität und Rechenleistung auf, um die Verfahren auch mit großen Datensätzen durchzuführen. Gleichwohl kann die Berechnung nach wie vor recht langwierig sein.

Unübersichtlichkeit der Ausgabe: Bieten Programme keine Möglichkeit an, das Verschmelzungsschema oder das Dendrogramm abzuschneiden und zum Beispiel nur die letzten x Schritte auszugeben, so ist der Output unübersichtlich und nicht lesbar.

Für große Datensätze sind daher folgende Strategien denkbar (siehe dazu auch Kaufman und Rousseeuw 1990, S. 144–146, 155–160):

1. Es wird mit einem anderen Verfahren gerechnet, zum Beispiel mit den im Kapitel 12 behandelten K-Means-Verfahren. Dieses Vorgehen wird man wählen, wenn keine »schwerwiegenden« inhaltlichen Gründe gegen die Verwendung der quadrierten euklidischen Distanzen sprechen. Ein schwerwiegender inhaltlicher Grund würde etwa vorliegen, wenn die Ausprägungen – wie beim JACCARD-Koeffizienten beispielsweise – unterschiedlich gewichtet werden sollen. Ein anderer schwerwiegender Grund würde vorliegen, wenn die gesuchte Klassifikation invariant gegenüber monotonen Transformationen oder überlappend sein soll. Kein schwerwiegender Grund würde vorliegen, wenn anstelle der quadrierten euklidischen Distanz die euklidische Distanz oder die City-Block-Metrik verwendet werden soll, da die K-Means-Verfahren³ relativ robust gegenüber dieser Modellverletzung sind.
2. Es wird mit durchschnittlichen Befragten gerechnet. Die Objektzahl wird dadurch reduziert, dass aufgrund bestimmter Merkmale durchschnittliche Befragte berechnet werden. Mitunter wird man sich für dieses Vorgehen entscheiden, wenn eine hierarchische Darstellung gefunden und/oder durch die Mittelwertbildung Messfehler ausgeglichen werden sollen.
3. Es wird mit einer Zufallsstichprobe gerechnet. Die Objektzahl wird durch das Ziehen einer Zufallsstichprobe reduziert. Wir haben dieses Vorgehen bereits beim Repräsentanten-Verfahren angewendet. Für dieses Vorgehen wird man sich nur dann entscheiden, wenn die gewünschten Anforderungen (zum Beispiel jene des Repräsentanten-Verfahrens) durch kein anderes Clusteranalyseverfahren und durch keine andere Lösungsstrategie (zum Beispiel durch durchschnittliche Befragte) erreicht werden kann.

Wir wollen hier die Methode der Verwendung von *durchschnittlichen Befragten* für das Beispiel der Analyse der Freizeitaktivitäten von Zehnjährigen darstellen. In die Analyse werden als Klassifikationsvariablen die dichotomen Freizeitaktivitäten einbezogen. Die durchschnittlichen Befragten werden aufgrund folgender sozialstruktureller Merkmale gebildet:

1. Höchste abgeschlossene Schulbildung der Eltern mit den Ausprägungen »niedrig« (Pflichtschule ohne Lehre), »mittel« (Pflichtschule mit Lehre, BMS) und »hoch«

³ Die Verwendung der euklidischen Distanzen anstelle von quadrierten euklidischen Distanzen hat keinen Einfluss auf die Ergebnisse.

(Matura und höher). Besitzt der Partner der Mutter eine höhere Schulbildung, wird die Schulbildung des Partners in die Analyse einbezogen, andernfalls jene der Mutter. Die Schulbildung der Mutter wurde ferner verwendet, wenn kein Partner vorhanden ist.

2. Pro-Kopf-Nettoeinkommen mit den Ausprägungen »niedrig« (bis 5 500 öS), »mittel« (5 500 bis 10 000 öS) und »hoch« (über 10 000 öS). Das Pro-Kopf-Nettoeinkommen ist das Nettohaushaltseinkommen (Nettoeinkommen aller Haushaltsmitglieder einschließlich von Transferzahlungen) dividiert durch eine gewichtete Personenzahl.
3. Beruf der Eltern mit den Ausprägungen »leitender Angestellter/Beamter, Selbständiger, Freiberuflicher«, »Facharbeiter«, »angelernter Arbeiter«, »mittlerer und einfacher Angestellter« und »Landwirt«. Es wurde der Beruf des Partners verwendet. Sofern dieser nicht bekannt war, wurde der Beruf der Mutter verwendet.
4. Gemeindegröße mit den Ausprägungen »klein« (bis 15 000 Einwohner), »mittel« (15 000 bis 160 000 Einwohner) und »groß« (Wien).
5. Geschlecht mit den Ausprägungen »Mädchen« und »Junge«.

Kombiniert man alle fünf Variablen, ergeben sich 270 mögliche Ausprägungen bzw. durchschnittliche Befragte: Der erste durchschnittliche Befragte ist durch die Merkmalskombination »geringe Schulbildung«, »geringes Pro-Kopf-Einkommen«, »leitender Angestellter«, »kleine Gemeinde« und »Mädchen« gekennzeichnet, der zweite durch »geringe Schulbildung«, »geringes Pro-Kopf-Nettoeinkommen«, »leitender Angestellter«, »kleine Gemeinde«, »Junge« usw.

Von den 270 möglichen durchschnittlichen Befragten treten nur 161 empirisch auf, da die Variablen nicht unabhängig sind. Bei der Analyse wird nun wie folgt vorgegangen:

1. Es werden die Mittelwerte der durchschnittlichen Befragten in den Freizeitaktivitäten berechnet.
2. Diese Mittelwerte werden als Klassifikationsvariablen in die Clusteranalyse einbezogen.
3. Als Objekte gehen die durchschnittlichen Befragten in die Analyse ein. Die nicht besetzten durchschnittlichen Befragten werden eliminiert.
4. Die Objekte werden mit ihren Auftrittshäufigkeiten gewichtet. Tritt also beispielsweise ein durchschnittlicher Befragter 30-mal auf, wird er 30-mal in die Analyse einbezogen.
5. Als Clusteranalyseverfahren soll das Ward-Verfahren verwendet werden. Es wird daher mit quadrierten euklidischen Distanzen gerechnet.
6. Als Deskriptionsvariablen, also als Variablen, die nicht in die Clusteranalyse eingehen, sondern nur zur Beschreibung der Cluster dienen, wurden die fünf sozialstrukturellen Variablen verwendet.

Die erste in einer hierarchischen Clusteranalyse zu lösende Aufgabe ist die Auswahl möglicher Clusterlösungen. Wir wollen dazu die Technik des Distanzzuwachses verwenden. Lesen wir das Verschmelzungsschema (siehe Tabelle 11.4 auf der nächsten Seite) von oben nach unten, so tritt ein erster deutlicher Zuwachs beim Übergang von sieben zu sechs Clustern auf. Wir entscheiden uns daher vorläufig für eine 7-Clusterlösung. Betrachten wir zur Absicherung der Entscheidungen die Teststatistiken von Mojena, so fällt auch hier die 7-Clusterlösung durch hohe Werte auf. Für die Teststatistik I (Mojena-I) besitzt sie eine Signifikanz von 99,9 Prozent ($p < 0,001$ einseitig). Allerdings treten bereits vorher Werte auf, die innerhalb des von Mojena empfohlenen Wertebereichs von 2,75 bis 3,50 liegen. In der Teststatistik II (Mojena-II) tritt bei der 7-Clusterlösung zum ersten Mal ein Wert mit einem Signifikanzniveau von 95 Prozent auf ($p < 0,05$ einseitig). Der kritische Wert von 2,75 in der Teststatistik wird aber erst bei der 3-Clusterlösung erreicht. Die Teststatistiken von Mojena bestätigen also – zumindest teilweise – die Wahl der 7-Clusterlösung. Mitunter wird man aber auch noch die 3-Clusterlösung untersuchen. Die erklärte Streuung für die 7 Clusterlösung ist 0,22 bzw. 22 Prozent. Sie ist somit nicht besonders hoch. Bei dieser Bewertung dürfen wir allerdings nicht vergessen, dass auch bei einem kausalanalytischen Vorgehen, zum Beispiel in Form einer Pfadanalyse, die erklärte Streuung oft nicht höher ist.

Für die 7-Clusterlösung ergeben sich die in Tabelle 11.5 auf der nächsten Seite dargestellten gewichteten Clustergrößen. Es werden drei Ausreißer (schwach besetzte Cluster) und vier relativ gut besetzte Cluster gebildet. Die zur Clusterbeschreibung (exemplarisch nur für Cluster 1) erforderlichen Informationen enthalten die Tabellen 11.6 bis 11.7 auf den Seiten 296–297. Die gebildeten Cluster können folgendermaßen beschrieben werden: Sozialstrukturell ist das Cluster 1 dadurch gekennzeichnet, dass es fast ausschließlich aus Jungen besteht, die überdurchschnittlich häufig aus einer mittleren Bildungsschicht kommen. Überdurchschnittlich häufig werden (wird) in der Freizeit Comics gelesen, ein Kino besucht, ferngesehen, Computer gespielt, Vereinsveranstaltungen und Parties besucht, Sport betrieben und radgefahren, aber auch alleine gespielt. Wir haben hier also ein typisches Jungen-Freizeitmuster und somit einen Hinweis auf geschlechtsspezifische Freizeitmuster, das vor allem für die mittlere Bildungsschicht charakteristisch ist. Das Cluster 2 ist dagegen für Mädchen aus unteren Bildungs- und Einkommensschichten, deren Väter Arbeiter oder Landwirte sind, charakteristisch. Alle Freizeitaktivitäten – mit Ausnahme des Spielens mit Haustieren und eines Kirchenbesuchs – werden mit einer geringeren Häufigkeit ausgeübt. Auch das Cluster C₄ ist für Mädchen typisch. Im Unterschied zum Cluster C₂ werden hier alle Freizeitaktivitäten mit Ausnahme von oft »als pädagogisch weniger wertvoll betrachteten« Freizeitaktivitäten (Comics lesen, Fernsehen, Computerspiele) und von organisierten Freizeitaktivitäten (Kinobesuch, Vereinsveranstaltungen, Besuch von Parties) überdurchschnittlich häufig ausgeübt. Sozialstrukturell rekrutiert sich dieses Cluster überdurchschnittlich aus der Schicht der leitenden Ange-

Tab. 11.4: Verschmelzungsschema für das Ward-Verfahren bei Verwendung von durchschnittlichen Befragten (nur die letzten 25 Schritte sind wiedergegeben)

Clusterzahl	Verschm.-niveau	Zunahme	erklärte Streuung	Mojena-I Testst.	Mojena-I Sign.	Mojena-II Testst.	Mojena-II Sign.
25	3,764	0,090	0,49	2,500	99,321	0,880	80,975
24	3,799	0,036	0,48	2,670	99,579	1,049	85,215
23	4,047	0,248	0,47	2,603	99,494	0,984	83,650
22	4,057	0,010	0,45	2,546	99,405	0,927	82,234
21	4,073	0,015	0,44	2,694	99,606	1,075	85,795
20	4,313	0,240	0,43	2,659	99,568	1,041	85,014
19	4,360	0,047	0,41	2,726	99,637	1,108	86,523
18	4,524	0,164	0,40	2,767	99,675	1,151	87,417
17	4,667	0,143	0,38	2,715	99,628	1,101	86,363
16	4,702	0,035	0,37	2,746	99,657	1,133	87,044
15	4,836	0,134	0,36	2,714	99,627	1,102	86,389
14	4,894	0,058	0,34	2,705	99,619	1,094	86,226
13	4,981	0,087	0,32	2,646	99,556	1,036	84,910
12	5,000	0,019	0,31	2,707	99,621	1,098	86,295
11	5,177	0,177	0,29	2,807	99,723	1,198	88,371
10	5,418	0,240	0,28	2,867	99,765	1,260	89,515
9	5,616	0,198	0,26	2,872	99,768	1,267	89,645
8	5,745	0,129	0,24	2,994	99,836	1,392	91,701
7	6,052	0,307	0,22	3,700	99,985	2,101	98,141
6	7,271	1,219	0,20	3,658	99,982	2,071	98,008
5	7,467	0,196	0,18	4,079	100,000	2,504	99,337
4	8,424	0,957	0,15	4,248	100,000	2,691	99,607
3	9,086	0,662	0,12	9,464	100,000	7,926	100,000
2	19,005	9,918	0,06	8,184	100,000	6,824	100,000
1	20,474	1,469	0,00	—	—	—	—

grau hinterlegt: empfohlener Wertebereich (Mojena I) bzw. Schwellenwert (Mojena II) für die Mojena-Teststatistiken; deutliche Zunahme im Verschmelzungsniveau

Tab. 11.5: Gewichtete Clustergrößen der 7-Clusterlösung

Cluster	n	Anteil in %
C_1	696	41,9
C_2	300	18,1
C_3	4	0,2
C_4	485	29,2
C_5	11	0,7
C_6	161	9,7
C_7	3	0,2
	1 660	100,0

stellten und aus kleinen Gemeinden. Analog kann C_6 beschrieben werden. Bei C_3 , C_5 und C_7 handelt es sich um kleine Restcluster.

Die Methode von durchschnittlichen Befragten ist vor allem zur Darstellung von hierarchischen Ähnlichkeitsbeziehungen im Rahmen einer Stimulusskalierung geeignet, also zum Beispiel wenn die Ähnlichkeit von Berufen, Regionen usw. untersucht werden soll. Das dabei vorliegende Datenmaterial lässt sich wie folgt skizzieren: In einer Befragung werden die Berufe oder eine andere zu skalierende Variable (zum Beispiel Region, soziale Schicht usw.) der Befragten erhoben. Darüber hinaus werden für jeden Beruf bzw. für die zu skalierende Variable weiter Informationen erhoben (zum Beispiel Qualifikationsprofil und Tätigkeitsprofil bei Berufen). Für die Analyse wird eine neue durchschnittliche Datenmatrix mit den Berufen oder zu skalierenden Variablen als Zeilen (Objekte) und den Durchschnittswerten in den zusätzlich erhobenen Variablen (Spalten) berechnet. Für diese neue Datenmatrix wird eine objektorientierte Clusteranalyse, zum Beispiel mit dem Ward-Verfahren, durchgeführt. Median- und Zentroid-Verfahren sind für eine hierarchische Darstellung nur bedingt geeignet, da Inversionen auftreten können. Anstelle des Ward-Verfahrens kann auch das Weighted-Average-Verfahren oder – falls sehr strenge Homogenitätsvorstellungen vorliegen – der Complete-Linkage verwendet werden.

11.4 Anwendungsempfehlungen

1. Bei kleineren Fallzahlen empfehlen wir bei einer objektorientierten Fragestellung den Einsatz des Ward-Verfahrens, wenn keine inhaltlichen Gründe gegen die Verwendung der quadrierten euklidischen Distanz sprechen (siehe Abschnitt 9.6).
2. Das Zentroid- und das Median-Verfahren sind allgemein weniger gut geeignet, da Inversionen auftreten können (siehe Abschnitt 9.5).

Tab. 11.6: Maßzahlen für die Deskriptionsvariablen für das Cluster 1

Variable	Ausprägung	n	Min	Max	\bar{x}	s	z-Wert
<i>Bildung</i>							
1	niedrig	696	0,00	1,00	0,04	0,19	-4,24
2	mittel	696	0,00	1,00	0,70	0,46	2,29
3	hoch	696	0,00	1,00	0,26	0,44	-0,52
<i>Einkommen</i>							
1	niedrig	696	0,00	1,00	0,16	0,37	-4,01
2	mittel	696	0,00	1,00	0,37	0,48	1,41
3	hoch	696	0,00	1,00	0,47	0,50	1,59
<i>Beruf der Eltern</i>							
1	leitAngBea	696	0,00	1,00	0,35	0,48	1,77
2	Facharb	696	0,00	1,00	0,21	0,40	0,51
3	angArbBea	696	0,00	1,00	0,11	0,31	0,08
4	mitAngBea	696	0,00	1,00	0,34	0,47	1,21
5	selbLandw	696	0,00	1,00	0,00	0,07	-25,17
<i>Gemeinde</i>							
1	klein	696	0,00	1,00	0,68	0,47	1,09
2	mittel	696	0,00	1,00	0,19	0,40	-1,57
3	groß	696	0,00	1,00	0,13	0,33	0,33
<i>Geschlecht</i>							
1	weiblich	696	0,00	1,00	0,04	0,20	-61,82
2	männlich	696	0,00	1,00	0,96	0,20	61,82

Anmerkung: Zur Interpretation der ausgewiesenen Maßzahlen (\bar{x} , s und z-Werte) siehe auch Abschnitt 12.4.

Zu den Variablenausprägungen siehe Text.

Tab. 11.7: Maßzahlen für die Klassifikationsvariablen für das Cluster 1

Variable	Ausprägung	n	Min	Max	\bar{x}	s	z-Wert
Ausruhen	ja	696	0,00	1,00	0,36	0,13	-2,64
Freunde	ja	696	0,00	1,00	0,77	0,10	-7,96
Familie	ja	696	0,00	1,00	0,62	0,15	1,72
Basteln	ja	696	0,00	1,00	0,37	0,12	-10,34
Comics	ja	696	0,00	1,00	0,55	0,15	15,49
Musizieren	ja	696	0,00	1,00	0,22	0,11	-17,93
Haustiere	ja	696	0,00	1,00	0,52	0,14	-8,00
Kino	ja	696	0,00	0,44	0,10	0,09	6,45
Konzert	ja	696	0,00	1,00	0,17	0,11	-2,28
Musikhören	ja	696	0,00	1,00	0,67	0,14	-8,86
Kirche	ja	696	0,00	1,00	0,50	0,16	-4,20
Fernsehen	ja	696	0,33	1,00	0,77	0,13	7,30
Computersp.	ja	696	0,00	1,00	0,57	0,14	33,31
Buch	ja	696	0,00	1,00	0,65	0,13	-15,60
Vereinsv.	ja	696	0,00	1,00	0,20	0,10	15,06
Radfahren	ja	696	0,60	1,00	0,91	0,08	10,65
Spazieren	ja	696	0,00	1,00	0,54	0,14	-7,77
alleine Sp.	ja	696	0,00	1,00	0,66	0,13	8,22
Parties	ja	696	0,00	1,00	0,20	0,10	6,36
Sport	ja	696	0,00	1,00	0,73	0,10	18,80

Anmerkung: Zur Interpretation der ausgewiesenen Maßzahlen (\bar{x} , s und z-Werte) siehe auch Abschnitt 12.4.

12 K-Means-Verfahren

12.1 Modellansatz und Algorithmus

Bei den *K-Means-Verfahren* werden – wie bei dem Ward-, Zentroid- und Median-Verfahren – Clusterzentren zur Bildung der Cluster konstruiert. Für K Cluster sollen die Clusterzentren so bestimmt werden, dass die *Streuungsquadratsumme in den Clustern* ein Minimum ist. Formal ausgedrückt: Es werden K Clusterzentren \bar{x}_{kj} ($k = 1, 2, \dots, K$; $j = 1, 2, \dots, m$; K = Anzahl der Cluster; m = Zahl der Variablen) so berechnet, dass die Streuungsquadratsumme in den Clustern

$$sQ_{in}(K) = \sum_k \sum_{g \in k} \sum_j (x_{gj} - \bar{x}_{kj})^2$$

ein Minimum annimmt. Da die Streuung der Variablenwerte eines Clusters von den Clusterzentren

$$\sum_k (x_{gj} - \bar{x}_{kj})^2 = d_{g,k}^2$$

gleich der quadrierten euklidischen Distanz d^2 zwischen dem Objekt g und dem Clusterzentrum k ist, kann die Minimierungsaufgabe geschrieben werden als

$$sQ_{in}(K) = \sum_j \sum_{g \in k} d_{g,k}^2 \rightarrow \min .$$

Die Streuungsquadratsumme in den Clustern lässt sich wie in der Varianzanalyse als *Fehlerstreuung* interpretieren. Sie ist jene Streuung in den Daten, die nicht durch die Cluster erklärt wird. Daher werden die Bezeichnung »Streuungsquadratsumme in den Clustern« und »Fehlerstreuung« synonym verwendet. Da die Gesamtstreuungsquadratsumme sQ_{ges} einer konstanten Größe entspricht, ist die Minimierung der Streuungsquadratsumme $sQ_{in}(K)$ innerhalb der Cluster gleich der Maximierung der Streuungsquadratsumme $sQ_{zw}(K) = sQ_{ges} - sQ_{in}(K)$ zwischen den Clustern.

In der Zielsetzung, die Fehlerstreuung bzw. die Streuungsquadratsumme in den Clustern zu minimieren, entspricht der Modellansatz jenem des Ward-Verfahrens (siehe Abschnitt 11.1). Im Unterschied zum Ward-Verfahren erfolgt die Clusterbildung aber nicht

hierarchisch, sondern nach einem partitionierenden Algorithmus, wobei die Clusterzahl K vorgegeben werden muss. Die Schritte des Algorithmus sind:

Schritt 1: Zufällige Zuordnung der Objekte zu K Clustern.

Schritt 2: Berechnung der Clusterzentren: Nach der Zuordnung aller Objekte zu den Clustern werden die Clusterzentren neu berechnet über

$$\bar{x}_{kj} = \frac{\sum_{g \in k} x_{gj}}{n_{kj}} \quad (12.1)$$

mit n_{kj} als Zahl der Objekte des Clusters k mit gültigen Angaben in der Variablen j . In die Summenbildung werden nur Ausprägungen mit gültigen Angaben einbezogen.

Schritt 3: Neuzuordnung der Klassifikationsobjekte: Die Klassifikationsobjekte g werden jenem Clusterzentrum k zugeordnet, zu dem die quadrierte euklidische Distanz minimal ist. Formal ausgedrückt:

$$g \in k \Leftrightarrow k = \min_{k^*=1,2,\dots,K} (d_{g,k^*}^2).$$

Dies führt dazu, dass die Streuungsquadratsumme in den Clustern

$$SQ_{in}(K) = \sum_k \sum_{g \in k} d_{g,k}^2 = \sum_g \min_{k^*=1,2,\dots,K} (d_{g,k^*}^2) \quad (12.2)$$

in jedem Iterationszyklus minimiert wird.

Schritt 4: Iteration: Es wird geprüft, ob sich im Schritt 3 die Zuordnung der Objekte ändert. Ist dies der Fall, werden die Schritte 2 und 3 erneut durchgeführt – ist dies nicht der Fall, wird der Algorithmus beendet.

Der hier angeführte Algorithmus wurde von Forgy (1965) entwickelt.¹ Er wird in der Literatur daher auch als »Forgys Methode« bezeichnet (Jain und Dubes 1988, S. 97; Kaufman und Rousseeuw 1990, S. 112–133; Punj und Stewart 1983). Für die K-Means-Verfahren existiert insgesamt keine allgemeinverbindliche Sprachkonvention. Steinhausen und Langer (1977) sprechen von »Verfahren zur Verbesserung einer Ausgangspartition«, Forgy's Methode wird von ihnen als »Minimaldistanzverfahren für das Varianzkriterium« bezeichnet. Jain und Dubes (1988) behandeln das K-Means-Verfahren als Submodell von partitionierenden Verfahren und sprechen von »Square-Error Clustering Methods« (Jain und Dubes 1988, S. 96). Kaufmann und Pape (1984) sprechen von »optimalen Partitionsverfahren« und dem dabei angewendeten »Varianzkriterium«. Auch die Bezeichnung

¹ Das K-Means-Verfahren lässt sich allerdings in mehrfacher Hinsicht modifizieren (siehe dazu Abschnitte 12.11 und 12.12).

Tab. 12.1: Konvergenzverhalten des K-Means-Verfahrens für unterschiedliche Stichprobengrößen

Stichprobengröße $n = n_1 + n_2$	Clusterzentren		Stichprobengröße $n = n_1 + n_2$	Clusterzentren	
	Cluster 1	Cluster 2		Cluster 1	Cluster 2
20	-0,67	1,75	1000	-1,17	1,18
50	-1,29	1,14	2000	-1,13	1,18
100	-1,07	1,26	5000	-1,14	1,15
200	-1,28	1,19	10 000	-1,15	1,18
500	-1,08	1,12			

»iteratives reallokatives Sum-of-Squares-Verfahren« wird verwendet (Gordon 1981, S. 44–46). Wir werden im Folgenden die Bezeichnung »K-Means-Verfahren« beibehalten.

Es lässt sich zeigen, dass der dargestellte Algorithmus für eine untersuchte Datenmatrix zumindest gegen ein *lokales Minimum* konvergiert (Jain und Dubes 1988, S. 99). Bock (1974, 1989), Bryant (1991), Jahnke (1988, S. 137–166) sowie Pollard (1981, 1982) haben das *asymptotische Verhalten des K-Means-Verfahrens* untersucht. Geht die Stichprobengröße n gegen unendlich ($n \rightarrow \infty$), gilt Folgendes (Bock 1989):

1. Die gefundene Partition nähert sich der optimalen Partition der »wahren« Clusterstruktur an.
2. Die berechneten Clusterzentren nähern sich den Clusterzentren für die optimale Partition an.

Voraussetzung für dieses asymptotische Verhalten ist, dass die untersuchten Variablen unabhängige Zufallsvariablen mit identischer Wahrscheinlichkeitsverteilung sind und dass in der Grundgesamtheit ($n \rightarrow \infty$) eine und nur eine optimale Partition mit einer minimalen Fehlerstreuung vorliegt. Bryant (1991) konnte aufbauend auf die Arbeiten von Windham (1986) zeigen, dass ein Großteil der asymptotischen Konvergenzbedingungen durch eine Analyse der verwendeten Zuordnungsfunktion untersucht werden kann.

Das Konvergenzverhalten soll eine Simulation verdeutlichen. Dabei wird von folgendem zugrunde liegenden Clustermodell ausgegangen: Es liegen zwei gleich große ($n_1 = n_2$), in einer Variablen X normalverteilte Cluster mit den Mittelwerten $\mu_1 = -1$ und $\mu_2 = 1$ und den Varianzen $\sigma_1^2 = \sigma_2^2 = 1$ vor. Wir erzeugen für unterschiedliche Stichprobengrößen n normalverteilte Zufallszahlen entsprechend den Modellparametern für jedes Cluster und führen anschließend eine Clusteranalyse mit dem K-Means-Verfahren durch. Die Ergebnisse dieser Simulation zeigt Tabelle 12.1: Es ist ersichtlich, dass ab $n \geq 1\,000$ die Ergebnisse weitgehend stabil sind. Die absoluten Abweichungen für die berechneten Mittelwerte sind kleiner/gleich 0,04. Das Verfahren konvergiert also – unter den getroffenen Modellannahmen – ab einer Stichprobengröße von 1 000, wobei zu beachten ist, dass die Cluster nicht besonders gut voneinander getrennt sind und ein beträchtlicher

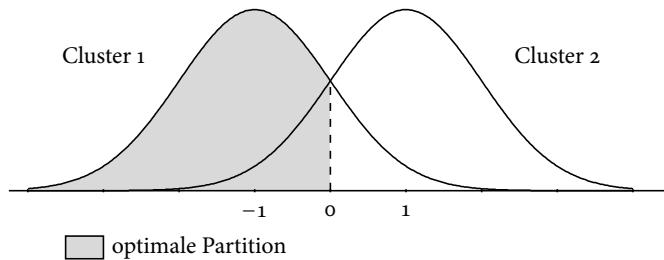


Abb. 12.1: Beziehung zwischen theoretischem Clustermodell und der optimalen Partition des theoretischen Modells

Anteil von Überlappungen vorliegt (siehe Abbildung 12.1). Die Überlappungen führen dazu, dass die Mittelwerte der optimalen Partition in der theoretischen Grundgesamtheit ($n \rightarrow \infty$) nicht gleich den Mittelwerten der gemischten Normalverteilungen sind. Die optimale Partition ist nämlich jene Partition, bei der alle Werte links dem Nullpunkt dem Cluster 1 zugeordnet werden und alle Werte rechts dem Cluster 2. Daher gilt: $\mu_1^{\text{opt}} < \mu_1$ und $\mu_2^{\text{opt}} > \mu_2$, wobei μ_1^{opt} und μ_2^{opt} die Mittelwerte der optimalen Partition in der theoretischen Grundgesamtheit sind (Kaufmann und Pape 1984, S. 451). Das Konvergenzverhalten ist allgemein besser, wenn ein geringer Überlappungsanteil vorliegt. Dies kann in unserem Simulationsbeispiel dadurch erreicht werden, dass der Abstand der vorgegebenen Mittelwerte μ_1 und μ_2 vergrößert und/oder mit mehreren voneinander unabhängigen Variablen mit denselben Verteilungsparametern gerechnet wird. Tabelle 12.2 zeigt das Konvergenzverhalten bei zwei und fünf Variablen. Die Erhöhung der Variablenzahl auf zwei Variablen führt zu keiner erkennbaren Verbesserung des Konvergenzverhaltens. Ab einer Stichprobengröße von 1 000 sind die Ergebnisse wiederum stabil. Bei fünf Variablen liegen dagegen ab einer Stichprobengröße von 500 stabile Ergebnisse vor. Da der Überlappungsanteil unter den getroffenen Modellannahmen mit der Variablenzahl abnimmt, nähern sich die Mittelwerte der optimalen Partition den Mittelwerten der angenommenen Normalverteilungen an.

Allgemein können wir festhalten: Liegt den Daten eine *fast überlappungsfreie* Clusterstruktur (»wahre« Clusterstruktur) zugrunde, so ist das K-Means-Verfahren auch bei mittleren Stichproben von $n > 500$ geeignet, die »wahren« Clusterstrukturen aufzufinden. Ist die zugrunde liegende Clusterstruktur *überlappungsfrei*, so ist auch eine kleinere Stichprobengröße ausreichend. Wenn beispielsweise angenommen wird, dass zwei Cluster vorliegen, die sich im Intervall $[-1,0]$ bzw. $[0,1]$ gleichverteilen, so werden bereits ab einem Stichprobenumfang von 50 Objekten stabile Ergebnisse erzielt. Die geschätzten Mittelwerte in einer Stichprobe von 50 Objekten sind $-0,53$ und $0,47$ und entsprechen den »wahren« Mittelwerten von $-0,5$ und $0,5$. Liegt dagegen eine *überlappende* Clusterstruktur vor, benötigt das K-Means-Verfahren eine größere Stichprobenzahl. Die

Tab. 12.2: Konvergenzverhalten des K-Means-Verfahrens bei zwei und fünf Variablen

Stichprobengröße $n = n_1 + n_2$	Clusterzentren bei zwei Variablen		Clusterzentren bei fünf Variablen	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
20	-0,83	1,26	-0,90	1,12
	-0,54	1,67	-0,46	1,37
			-1,05	1,11
			-1,22	1,68
			-1,03	1,06
50	-0,84	0,89	-0,97	0,94
	-1,54	1,40	-1,44	1,19
			-1,12	0,95
			-1,31	0,79
			-1,00	1,05
100	-0,71	1,02	-1,02	0,68
	-0,78	1,08	-0,93	0,60
			-0,95	0,93
			-0,78	1,03
			-1,07	1,18
200	-1,05	1,28	-1,02	1,20
	-1,11	1,26	-1,07	1,18
			-0,98	1,01
			-0,92	0,89
			-0,98	0,99
500	-1,00	1,10	-1,02	1,06
	-1,12	1,11	-1,09	1,00
			-1,01	0,98
			-1,01	0,97
			-1,01	1,03
1 000	-0,99	1,06	-0,96	1,03
	-1,09	1,00	-1,03	0,96
			-0,98	0,98
			-1,02	0,98
			-1,00	1,02
2 000	-1,07	1,07	-1,02	1,04
	-1,04	1,00	-1,00	0,97
			-1,03	1,04
			-0,99	0,98
			-0,97	1,02
5 000	-1,05	1,06	-0,99	1,01
	-1,05	1,03	-0,98	1,00
			-0,99	1,00
			-1,02	1,01
			-1,00	1,03

berechneten Mittelwerte entsprechen nicht den Mittelwerten der »wahren« Clusterstruktur, sondern jenen einer optimalen Partition der »wahren« Clusterstruktur. »Optimal« heißt, dass die Varianz in den Clustern ein Minimum ist. In der praktischen Anwendung ist nicht bekannt, welche Clusterstruktur den Daten zugrunde liegt. Erst aufgrund der Ergebnisse kann die Angemessenheit des K-Means-Verfahrens beurteilt werden. Für eine überlappende Clusterstruktur empfiehlt sich die Anwendung von probabilistischen Verfahren, wenn die Clusterzentren im Fokus der Analyse stehen (siehe Teil III).

Wir wollen nun den Algorithmus anhand der beim Ward-Verfahren verwendeten fiktiven Datenmatrix mit drei Clustern veranschaulichen (siehe Tabelle 11.1 auf Seite 287). Entsprechend Schritt 1 werden die Objekte den drei Clustern zufällig zugeordnet. Anschließend werden die Clusterzentren berechnet. Dem ersten Cluster gehören die Objekte A, D, F und I an. Die Clusterzentren für Cluster 1 sind daher (siehe Tabelle 12.3 auf Seite 306):

$$\begin{aligned}\bar{x}_{11} &= (-2 + 0 + 2 + 4)/4 = 1,00, \\ \bar{x}_{12} &= (1 - 1 + 2 + 3)/4 = 1,25.\end{aligned}$$

Analog werden die Clusterzentren für die anderen Cluster berechnet, die in der Tabelle 12.3 auf Seite 306 in den mit C_1 , C_2 und C_3 beschrifteten Zeilen stehen.

Im nächsten Schritt werden für jedes Objekt die quadrierten euklidischen Distanzen zu den Clusterzentren ermittelt. Für Objekt A ergeben sich folgende Werte:

$$\begin{aligned}d_{A,1} &= (-2 - 1,00)^2 + (1 - 1,25)^2 = 9,0625, \\ d_{A,2} &= (-2 - 1,3333)^2 + (1 - 1,00)^2 = 11,1109, \\ d_{A,3} &= (-2 - 1,00)^2 + (1 - 0,00)^2 = 10,0000.\end{aligned}$$

Die quadrierte euklidische Distanz von Objekt A zu Cluster 1 ist gleich 9,06. Zu Cluster 2 beträgt die quadrierte euklidische Distanz gleich 11,11 und zu Cluster 3 beträgt sie 10,00. Die geringste Distanz liegt zu Cluster 1 vor. Objekt A wird daher dem Cluster 1 zugeordnet. Diesem Cluster gehört es bereits an, eine Vertauschung der Clusterzuordnung ist nicht erforderlich. Für Objekt B ergibt sich eine Vertauschung. Es wurde im Startwertverfahren zufällig dem Cluster 2 zugeordnet. Die geringste Distanz wird aber für das Cluster 1 ausgewiesen, dem Objekt B dann auch zugeordnet wird. Insgesamt ändern sechs Objekte ihre Clusterzugehörigkeit. Da die Zahl der Vertauschungen größer 0 ist, wird der Algorithmus erneut durchlaufen (2. Iteration). Die Zuordnung der Objekte ändert sich nicht mehr, so dass der Algorithmus beendet werden kann. Die finalen drei Clusterzentren sind:

$$\bar{x}_{11} = -1,50, \quad \bar{x}_{12} = 1,50; \quad \bar{x}_{21} = 3,25, \quad \bar{x}_{22} = 2,25; \quad \bar{x}_{31} = 0,00 \quad \bar{x}_{32} = -1,33.$$

Die Streuungsquadratsumme in den Clustern kann wegen der Beziehung in Gleichung 12.2 auf Seite 300 aus den einzelnen Iterationsergebnissen berechnet werden. Sie ist für jede Iteration gleich der Summe der dunkelgrau hinterlegten Werte in Tabelle 12.3 auf der nächsten Seite. Für die erste Iteration ergibt sich ein Wert von: $sQ_{in}^{(1)}(K = 3) = 9,06 + 4,56 + 8,00 + 2,00 + 1,00 + 1,44 + 3,78 + 8,11 + 11,11 = 49,06$. Für die zweite Iteration nimmt die Streuungsquadratsumme in den Clustern folgenden Wert an: $sQ_{in}^{(2)}(K = 3) = 0,50 + 0,50 + 1,44 + 0,11 + 1,11 + 1,63 + 0,13 + 0,63 + 1,13 = 7,18$. Die gefundene 3-Clusterlösung besitzt demnach eine Fehlerstreuung von 7,18. Die Fehlerstreuung und daraus abgeleitete Maßzahlen bilden die Basis für die Modellprüfung des K-Means-Verfahrens.

12.2 Bestimmung der Clusterzahl

Voraussetzung für die *Bestimmung der Clusterzahl* ist, dass Clusterlösungen für eine unterschiedliche Clusterzahl berechnet werden. Das K-Means-Verfahren wird also für $K = 1, 2, 3$ usw. durchgerechnet. Die 1-Clusterlösung sollte dabei immer mitbetrachtet werden, um feststellen zu können, ob überhaupt eine Clusterstruktur vorliegt. Zur Vermeidung von *lokalen Minima* ist es dabei empfehlenswert, für jede Clusterzahl mehrere Lösungen mit unterschiedlichen zufälligen Startwerten zu ermitteln (Methode multipler zufälliger Startwerte; Steinley und Brusco 2007). Mitunter sind hierfür 1 000 oder mehr Versuche erforderlich. Je undeutlicher die Cluster voneinander getrennt sind, desto größer ist die Zahl der erforderlichen Versuche, um ein *globales Maximum bzw. Minimum* zu finden. Aber auch bei 1 000 und mehr Versuchen je Clusterzahl ist nicht vollständig gesichert, dass ein globales Minimum gefunden wurde, die Wahrscheinlichkeit dafür ist aber relativ hoch. Wie notwendig mehrere Rechenversuche je Clusterzahl sind, unterstreicht Tabelle 12.4 auf Seite 307. In die Klassifikation gingen zwei Variablen ein. Für die 1- und 2-Clusterlösung wird ein Minimum mit einem Versuch ermittelt. Für die 3-Clusterlösung ist ein Versuch nicht mehr ausreichend. Erst bei zehn und mehr Versuchen wird das Minimum ermittelt. Für die 4- bis 11-Clusterlösung sind 1 000 Versuche erforderlich, bei der 12-Clusterlösung 2 000. Obwohl sich oft nur die Zuordnung von wenigen Objekten ändert, während die Clusterzentren relativ stabil sind, sollte dennoch mit mehreren Versuchen gerechnet werden, um einen Anstieg der Fehlerstreuung bei einer höheren Clusterzahl zu vermeiden, wie dies noch in Bacher (1994, S. 319) der Fall war. In einem Vergleich von zwölf Startwertverfahren für das K-Means-Verfahren schnitt die Methode der multiplen zufälligen Startwerte am besten ab (Steinley und Brusco 2007). Damit konnten auch vorausgehende Simulationsstudien bestätigt werden. Wir empfehlen daher die Methode der multiplen zufälligen Startwerte. Eine für eine Clusterzahl gefundene

Tab. 12.3: Veranschaulichung des Algorithmus des K-Means-Verfahrens

X_1	X_2	Zuordnung	Distanzen zu Clustern			Neuzuordnung	Vertauschung
			1	2	3		
<i>erste Iteration</i>							
A	-2	1	9,06	11,11	10,00	1	0
B	-1	2	4,56	6,44	8,00	1	1
C	-1	-2	14,56	14,44	8,00	3	0
D	0	-1	6,06	5,78	2,00	3	1
E	1	-1	5,06	4,11	1,00	3	1
F	2	2	1,56	1,44	5,00	2	1
G	3	2	4,56	3,78	8,00	2	1
H	4	2	9,56	8,11	13,00	2	0
I	4	3	12,06	11,11	18,00	2	1
C_1	1,00	1,25	$SQ_{in}^{(1)} = 49,06$			gesamt	6
C_2	1,33	1,00					
C_3	1,00	0,00					
<i>zweite Iteration</i>							
A	-2	1	0,50	29,13	9,44	1	0
B	-1	2	0,50	18,13	12,11	1	0
C	-1	-2	12,50	36,13	1,44	3	0
D	0	-1	8,50	21,13	0,11	3	0
E	1	-1	12,50	15,63	1,11	3	0
F	2	2	12,50	1,63	15,11	2	0
G	3	2	20,50	0,13	20,11	2	0
H	4	2	30,50	0,63	27,11	2	0
I	4	3	32,50	1,13	34,78	2	0
C_1	-1,50	1,50	$SQ_{in}^{(2)} = 7,18$			gesamt	0
C_2	3,25	2,25					
C_3	0,00	-1,33					

grau hinterlegt: Objekte mit Vertauschungen nach dem ersten Iterationsschritt

dunkelgrau hinterlegte Werte: geringste Distanzen der Objekte zu den jeweiligen Clusterzentren

Lösung kann dadurch reproduziert werden, dass der entsprechende Startwert des Zufallzahlengenerators, der von ALMO dokumentiert wird, bei der erneuten Berechnung eingegeben und die Zahl der Versuche gleich 1 gesetzt wird.

Von den x Versuchen je Cluster wird für die weiteren Analysen jene Lösung mit der kleinsten Fehlerstreuung $SQ_{in}(K)$ ausgewählt. Diese stellt die Grundlage für die Berechnung von Modellprüfgrößen dar. Wie beim Ward-Verfahren kann die durch eine bestimmte Clusterlösung *erklärte Streuung* (η^2) mit

$$\eta_K^2 = 1 - \frac{SQ_{in}(K)}{SQ_{ges}} = 1 - \frac{SQ_{in}(K)}{SQ_{in}(1)} \quad (12.3)$$

Tab. 12.4: Minimum in Abhängigkeit von der Zahl der Versuche je Cluster

Cluster	Versuche je Clusterzahl				
	1	10	100	1 000	2 000
1	148,805	148,805	148,805	148,805	148,805
2	85,545	85,545	85,545	85,545	85,545
3	60,616	60,613	60,613	60,613	60,613
4	45,633	44,996	44,971	44,960	44,960
5	40,752	37,423	36,882	36,856	36,856
6	40,659	33,059	31,220	30,708	30,708
7	28,740	28,740	25,720	24,684	24,684
8	25,173	22,841	22,389	20,798	20,798
9	28,746	25,024	19,292	18,593	18,593
10	25,356	17,993	17,993	16,280	16,280
11	24,725	16,677	16,677	14,800	14,800
12	19,428	15,497	14,602	13,942	13,557

grau hinterlegt: stabile Minima

berechnet werden. $\text{SQ}_{\text{in}}(K)$ ist die Streuungsquadratsumme in den K Clustern, SQ_{ges} die Gesamtstreuungsquadratsumme. Die Gesamtstreuungsquadratsumme ist gleich der Streuungsquadratsumme in den Clustern für die 1-Clusterlösung $\text{SQ}_{\text{ges}} = \text{SQ}_{\text{in}}(1)$. Die prozentuelle Verbesserung gegenüber dem Nullmodell (es liegt kein Clusterstruktur vor) gibt $100 \cdot \eta^2$ an. Analog kann die *prozentuelle Verbesserung gegenüber einer vorausgehenden Clusterlösung* berechnet werden mit

$$\text{PRE}_K = 1 - \frac{\text{SQ}_{\text{in}}(K)}{\text{SQ}_{\text{in}}(K-1)}. \quad (12.4)$$

Diese Größe ist wie η_K^2 nach der Logik eines PRE-Koeffizienten (»proportional reduction of error«) konstruiert. Für die 2-Clusterlösung ($K = 2$) ist er gleich der erklärten Streuung.

Daneben kann noch in Anlehnung an die Varianzanalyse ein *F-Wert* berechnet werden mit

$$F_{\max_K} = \frac{\text{SQ}_{\text{zw}}(K)/(K-1)}{\text{SQ}_{\text{in}}(K)/(n-K)} = \frac{(\text{SQ}_{\text{ges}} - \text{SQ}_{\text{in}}(K))/(K-1)}{\text{SQ}_{\text{in}}(K)/(n-K)}. \quad (12.5)$$

Dieser F-Wert wird als maximale F-Statistik² bezeichnet, da bei der Clusteranalyse die Streuung zwischen den Clustern und damit auch der F-Wert maximiert wird. Im Unterschied zur erklärten Streuung und dem PRE-Koeffizienten wird der Tatsache Rechnung getragen, dass bei einer größeren Clusterzahl rein zufällig eine kleinere Fehlerstreuung

² Die F_{\max} -Statistik wird auch als Pseudo-F-Statistik nach Calinski und Harabasz (1974) bezeichnet.

auftritt. Dieser Effekt wird durch die Berücksichtigung der Freiheitsgrade beseitigt. Da mit der Minimierung der Streuungsquadratsumme innerhalb der Cluster der F_{\max} -Wert maximiert wird, besitzt die F_{\max} -Statistik im Unterschied zum F-Wert der Varianzanalyse keine F-Verteilung. Signifikanztests sind daher nicht möglich.³

Für die Fragestellung, ob die *Fehlerstreuung einer Clusterlösung mit K_2 Clustern kleiner ist als jene einer Clusterlösung mit K_1 Clustern ($K_2 > K_1$)*, lässt sich ebenfalls ein F-Wert konstruieren. Er wurde von Beale (Kendall 1980, S. 39–40) entwickelt. Daher soll hier von *Bealschen F-Werten* gesprochen werden. Sie sind wie folgt definiert:

$$\text{BEALE-F}_{K_2, K_1} = \left(\frac{\text{sq}_{\text{in}}(K_1) - \text{sq}_{\text{in}}(K_2)}{\text{sq}_{\text{in}}(K_2)} \right) \Bigg/ \left(\frac{n - K_1}{n - K_2} \cdot \left(\frac{K_2}{K_1} \right)^{2/m} - 1 \right), \quad (12.6)$$

wobei m die Variablenzahl ist.

Im Unterschied zur F_{\max} -Statistik können die Bealschen F-Werte auf Signifikanz geprüft werden. Unter der Annahme, dass die Variablen a) unabhängig sind, b) gleiche Skalen-einheiten besitzen und c) die gebildeten Cluster kugelförmige Gestalt haben, besitzen die F-Werte von Beale eine F-Verteilung mit $m \cdot (K_2 - K_1)$ und $m \cdot (n - K_2)$ Freiheitsgraden. Annahme b) ist eine Grundvoraussetzung. Mit Annahme a) ist gemeint, dass die Variablen innerhalb der Cluster unabhängig sind. Über alle Cluster hinweg muss keine Unabhängigkeit bestehen. Das K-Means-Verfahren hat allgemein die Tendenz, kugelförmige Cluster zu bilden (siehe zum Beispiel Everitt 1980, S. 69, 81–86), so dass die dritte Annahme in der Regel annähernd erfüllt ist. Die Bealschen F-Werte weisen allerdings nur dann Unterschiede als signifikant aus, wenn die Cluster sehr gut getrennt sind (Everitt 1980, S. 65). Tests mit ihnen sind – ähnlich wie der RUNT-Test von Hartigan und Mohanty (siehe Abschnitt 9.2) – sehr konservativ und mit einem großen β -Fehler behaftet. Werden zum Beispiel zwei Normalverteilungen mit den Mittelwerten $+2$ bzw. -2 und einer Varianz von 1 gemischt, ist der Bealsche F-Wert für die 1- und 2-Clusterlösung bei 400 Objekten nicht signifikant, obwohl die Mittelwerte deutlich getrennt sind.

Die Maßzahlen lassen sich nun wie folgt zur Bestimmung der Clusterzahl einsetzen:

1. Es wird jene Clusterlösung ausgewählt, für welche die erklärte Varianz über einem festgelegten Schwellenwert liegt. Die Festlegung eines Schwellenwertes ist aber schwierig und subjektiv. Diese Vorgehensweise ist daher nicht zu empfehlen. Abhilfe schafft hier die in Abschnitt 12.3 beschriebene Zufallstestung, welche die empirische Festlegung eines Schwellenwertes bei reinen Zufallsdaten ermöglicht.

³ Die asymptotische Verteilung der Testgröße $\text{sq}_{\text{zw}}(K)/\text{sq}_{\text{in}}(K)$ für den eindimensionalen Fall ist in Hartigan (1975, S. 97–100) angegeben.

2. Es wird (werden) jene Clusterlösung(en) ausgewählt, nach welche(n) der PRE-Koeffizient deutlich absinkt. Mit dieser Strategie haben wir die besten Erfahrungen gemacht.
3. Es wird die Lösung mit dem maximalen Wert in der F_{\max} -Statistik ausgewählt.
4. Es wird aufgrund der Bealschen F-Werte jene Lösung ausgewählt, die im Vergleich zu den vorausgehenden Clusterlösungen mit einer kleineren Clusterzahl zu einer signifikanten Reduktion der Fehlerstreuung führt, während bei den nachfolgenden Clusterlösungen keine signifikante Reduktion der Fehlerstreuung mehr eintritt.

Wir wollen diese Strategien am Beispiel der Wertedaten von Denz (1989) verdeutlichen. In die Analyse wurden als Klassifikationsvariablen die mittleren Gesamtpunktwerte in den postmaterialistischen und materialistischen Items einbezogen. Die Startwerte für die Clusterzentren wurden durch eine zufällige Zuordnung der Objekte zu den Clustern berechnet. Für jede Clusterzahl wurden 2 000 Versuche mit jeweils unterschiedlicher Startzahl gerechnet. Die Clusterzahl wurde zwischen 1 und 12 variiert. Es ergeben sich die in der Tabelle 12.5 auf der nächsten Seite dargestellten Kennwerte.

Die Modellprüfgrößen für jede Clusterlösung können entsprechend den Gleichungen 12.3 bis 12.6 auf den Seiten 306–308 aus den Streuungsquadratsummen innerhalb der Cluster berechnet werden. Für die F_{\max} -Statistik und die Bealschen F-Werte wird zusätzlich die Stichprobengröße benötigt. Für die 4-Clusterlösung ergeben sich folgende Werte: Die erklärte Streuung ist gleich

$$\eta_4^2 = 1 - 44,960 / 148,805 = 0,698$$

bzw. 69,8 Prozent. Die Werte zur Berechnung wurden der Tabelle 12.5 auf der nächsten Seite entnommen. Die vorausgehende Clusterlösung ($K = 3$) besitzt eine Streuungsquadratsumme in den Clustern von 60,613. Der PRE-Koeffizient ist daher gleich

$$PRE_6 = 1 - 44,960 / 60,613 = 0,258$$

bzw. 25,8 Prozent. Die 4-Clusterlösung verbessert die 3-Clusterlösung um 25,8 Prozent. Der F_{\max} -Wert der 4-Clusterlösung berechnet sich wie folgt: Die Streuungsquadratsumme $SQ_{zw}(4)$ zwischen den Clustern ist gleich der Gesamtstreuungsquadratsumme SQ_{ges} minus der Streuungsquadratsumme $SQ_{in}(4)$ in den Clustern:

$$SQ_{zw}(4) = 148,805 - 44,960 = 103,845 .$$

Insgesamt wurden 221 Befragte in die Analyse einbezogen, der F_{\max} -Wert ist daher für $K = 4$ gleich:

$$F_{\max_4} = (103,845 / (4 - 1)) / (44,960 / (221 - 4)) = 167,073 .$$

Tab. 12.5: Modellprüfgrößen für die Wertedaten von Denz für eine unterschiedliche Anzahl von Clustern

Clusterzahl <i>K</i>	Streuungsquadrat- summe in den Clustern sq _{in} (<i>K</i>)	erklärte Streuung η_K^2	proportionale Fehlerver- besserung PRE _K	F _{maxK}
1	148,805	0,000	—	0,000
2	85,545	0,425	0,425	161,949
3	60,613	0,593	0,291	158,598
4	44,960	0,698	0,258	167,073
5	36,856	0,752	0,180	164,025
6	30,708	0,794	0,167	165,369
7	24,684	0,834	0,196	179,345
8	20,798	0,860	0,157	187,287
9	18,593	0,875	0,106	185,588
10	16,280	0,891	0,124	190,840
11	14,800	0,901	0,091	190,145
12	13,942	0,906	0,058	183,786

grau hinterlegt: Schwellenwerte für ein deutliches Absinken des PRE-Koeffizienten; maximaler F_{max}-Wert

Der Bealsche F-Wert für die 3- und 4-Clusterlösung berechnet sich wie folgt: K_1 ist gleich 3 und K_2 gleich 4. Die entsprechenden Streuungsquadratsummen in den Clustern sind entsprechend Tabelle 12.5: $sq_{in}(3) = 60,613$ und $sq_{in}(4) = 44,960$. Da insgesamt $n = 221$ Objekte und $m = 2$ Variablen in die Analyse eingingen, ergibt sich bei Anwendung der Formel für die Bealschen F-Werte ein Wert von

$$\text{BEALE-F}_{3,4} = \left(\frac{60,613 - 44,960}{44,960} \right) \Bigg/ \left(\frac{221 - 3}{221 - 4} \cdot \left(\frac{4}{3} \right)^{2/2} - 1 \right) = 1,0256$$

mit 2 (= $2 \cdot (4 - 3)$) und 434 (= $2 \cdot (221 - 4)$) Freiheitsgraden.

Für die dargestellten Strategien zur Bestimmung der Clusterzahl ergeben sich nun folgende Ergebnisse:

- *Schwellenwert für η_K^2 :* Wird ein Schwellenwert von 0,70 bzw. 70 Prozent festgelegt, wird die 5-Clusterlösung ausgewählt. Allerdings liegt die erklärte Varianz für die 4-Clusterlösung mit 0,698 nur knapp unter dem Schwellenwert, so dass man auch die 4-Clusterlösung noch zulassen wird. Wird ein höherer Schwellenwert eingefordert, zum Beispiel 0,90, dann wird man sich für die 11-Clusterlösung entscheiden.
- *Deutlicher Abfall des PRE-Wertes:* Wir lesen die Tabelle 12.5 von oben nach unten. Ein erstes Absinken des PRE-Wertes kann nach zwei Clustern festgestellt werden. Der PRE-Koeffizient sinkt von 0,425 auf 0,291. Ein weiteres Absinken tritt nach vier Clustern ein. Der PRE-Koeffizient geht von 0,258 auf 0,180 zurück. Die 5-Clusterlösung

erbringt also im Vergleich zur 4-Clusterlösung eine deutlich geringere Verbesserung der Erklärungskraft. Ein erneuter Abfall ist bei acht Clustern und bei zehn Clustern zu erkennen. Auf der Basis der Entwicklung wird man sich für die 2-Clusterlösung, die 4-Clusterlösung und unter Umständen für die 8- oder 10-Clusterlösung entscheiden, da nach diesen Clusterlösungen der PRE-Koeffizient absinkt. Unterstützend kann auch ein Scree-Diagramm gezeichnet und nach einem Knickpunkt gesucht werden. Dieser ist aber graphisch oft schwer erkennbar.

- *Maximaler F_{\max} -Wert:* Die Lösung mit dem maximalen F_{\max} -Wert ist die 10-Clusterlösung ($F_{\max_{10}} = 190,840$). Der F_{\max} -Wert für die 11-Clusterlösung nimmt aber ebenfalls noch einen hohen Wert an. Knickpunkte mit geringeren Werten vorher und nachher sind noch für die 4-Clusterlösung und die 8-Clusterlösung beobachtbar.
- *Bealsche F-Werte:* Die Tabelle 12.6 auf der nächsten Seite enthält alle möglichen Bealschen F-Werte (unteres Dreieck). Der Bealsche F-Wert für die 2- und 4-Clusterlösung beispielsweise ist gleich 0,8864 (4. Zeile und 2. Spalte). Die Signifikanzen der Bealschen F-Werte stehen im oberen Dreieck. BEALE-F_{2,4} ist mit 52,6162 Prozent nicht signifikant (zweite Zeile und vierte Spalte). Das bedeutet, dass die erklärte Streuung der 4-Clusterlösung nicht signifikant größer ist als jene der 2-Clusterlösung.

Wir wollen nun exemplarisch für die 4-Clusterlösung prüfen, ob a) die Unterschiede zu allen vorausgehenden Clusterlösungen signifikant und b) zu allen nachfolgenden Clusterlösungen nicht signifikant sind. Im Idealfall müsste die Matrix der Bealschen F-Werte die in der Abbildung 12.2 auf Seite 313 dargestellte Struktur haben, wenn K Cluster vorliegen. Die Bealschen F-Werte aller vorausgehenden Clusterlösungen sind zu einem Niveau von 95 Prozent (oder einem anderen vorab definierten Schwellenwert) signifikant. Sie stehen oberhalb der Zeile 4 im oberen Dreieck der Tabelle 12.6 auf der nächsten Seite. Der erste Teil der Bedingung a) ist nicht erfüllt. Wenn wir alle berechneten Clusterlösungen prüfen, zeigt sich, dass keine Clusterlösung die geforderte Bedingung a) erfüllt, eine Prüfung von b) entfällt damit. Wir können dies auf die konservative Eigenschaft der Bealschen F-Werte zurückführen. Nur sehr gut getrennte Cluster werden als signifikant ausgewiesen. Inhaltlich bedeutet dies, dass in unserem Beispiel keine sehr gut getrennten Cluster vorliegen. Daraus darf aber nicht der Umkehrschluss gezogen werden, dass überhaupt keine Clusterstruktur vorliegt. Dies kann erst aufgrund weiterer Analyseschritte entschieden werden.

Die bisherige Analyse lässt sich wie folgt zusammenfassen:

1. Abhängig von einem gewählten Schwellenwert für η_K^2 (zum Beispiel 70 oder 90 Prozent) wird man sich für die 5- oder 11-Clusterlösung entscheiden. Allerdings liegen auch die 4- bzw. 10-Clusterlösung nahe dem jeweiligen Schwellenwert. Anzumerken ist, dass die Wahl eines Schwellenwerts schwierig ist. Abhilfe schafft hier die nachfolgend beschriebene Zufallstestung.

Tab. 12.6: Bealsche F-Werte

		Clusterzahl					
		1	2	3	4	5	6
1	0	51,4442	41,7075	39,3178	34,4120	31,9776	
2	0,7328	0	55,1461	52,6162	47,5104	45,6234	
3	0,7176	0,8115	0	63,9336	56,0844	53,8162	
4	0,7560	0,8864	1,0256	0	57,2814	53,6141	
5	0,7422	0,8608	0,9450	0,8596	0	61,9766	
6	0,7483	0,8686	0,9474	0,9030	0,9738	0	
7	0,8115	0,9550	1,0571	1,0605	1,1937	1,4179	
8	0,8475	1,0002	1,1071	1,1197	1,2403	1,3778	
9	0,8398	0,9869	1,0840	1,0883	1,1778	1,2501	
10	0,8635	1,0155	1,1142	1,1212	1,2066	1,2691	
11	0,8604	1,0094	1,1030	1,1065	1,1801	1,2257	
12	0,8316	0,9714	1,0552	1,0519	1,1102	1,1372	

unteres Dreieck: Bealsche F-Werte*oberes Dreieck:* Signifikanz der Bealschen F-Werte

(a) Teil 1

		Clusterzahl					
		7	8	9	10	11	12
1	36,0279	38,2176	35,9208	37,5568	36,0928	31,3103	
2	51,6977	55,1838	53,3948	56,1007	55,2123	50,3457	
3	60,7533	64,5271	62,8441	65,7757	65,0932	60,3753	
4	61,4257	65,1901	63,0614	65,9854	65,0866	59,9743	
5	68,7716	71,6415	68,9788	71,6241	70,5775	65,4114	
6	75,8044	76,0438	72,1308	74,2982	72,8528	67,2749	
7	0	71,6215	64,3892	66,8041	64,5627	57,4334	
8	1,2608	0	59,4150	62,3523	59,2493	50,7857	
9	1,1000	0,9099	0	70,4320	64,1876	53,7394	
10	1,1499	1,0596	1,2205	0	61,0640	46,7664	
11	1,1106	1,0269	1,0959	0,9506	0	46,7240	
12	1,0201	0,9300	0,9464	0,7930	0,6398	0	

unteres Dreieck: Bealsche F-Werte*oberes Dreieck:* Signifikanz der Bealschen F-Werte

(b) Teil 2

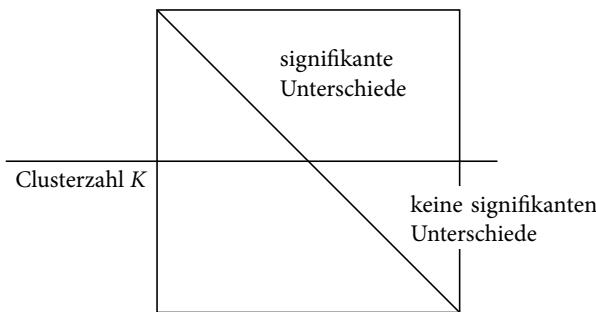


Abb. 12.2: Idealstruktur der Matrix der Bealschen F-Werte

2. Ein deutlicher Abfall des PRE-Koeffizienten ist bei zwei, vier, acht und zehn Clustern feststellbar.
3. Die 10-Clusterlösung ist die Lösung mit dem maximalen F-Wert. Ein »Peak« ist auch bei vier Clustern beobachtbar. Der F-Wert ist vorher und nachher geringer.
4. Die Bealschen F-Werte legen die Vermutung nahe, dass keine sehr deutlich getrennten Cluster vorliegen. Möglicherweise ist ein probabilistisches Clusteranalysemmodell den Daten angemessener (siehe Teil III).

Aus formalen Gründen sind – wenn das wegen der Festlegung von Schwellenwerten problematische Kriterium 1 außer Acht gelassen wird –, die 2-, 4-, 8- und 10-Clusterlösung geeignet. Dieser Befund, dass mehrere Clusterlösungen formal geeignet sind, ist für sozialwissenschaftliche Daten charakteristisch. Auch bei den hierarchisch-agglomerativen Verfahren kamen durchwegs mehrere Lösungen für eine weitere Analyse in Betracht. Welche Lösungen den Daten angemessen sind, kann erst nach den weiteren Analyse-schritten beurteilt werden. Nachfolgend werden die 4-, 8- und 10-Clusterlösung weiter untersucht. Die 2-Clusterlösung wird nicht weiter betrachtet, da sie zu heterogen ist.

12.3 Zufallstestung einer bestimmten Clusterlösung

Mit der *Zufallstestung* wird das Problem eines fehlenden Schwellenwertes für η^2 kompen-siert. Es wird eine Art Erwartungswert ermittelt, der deutlich kleiner sein sollte als der empirisch berechnete. Die Nullhypothese der Zufallstestung lautet: Auch bei Zufallsdaten liefert die gefundene Clusterlösung gleich gute Ergebnisse – gemessen durch η^2 .

Bei der Berechnung des Erwartungswertes für η^2 wird wie folgt vorgegangen: Es werden mit Rückgriff auf die empirischen Verteilungskennwerte der Variablen normalverteilte Zufallsdaten erzeugt. Daran anschließend wird untersucht, welches η^2 sich für diese

Zufallsdaten ergibt, wenn die Clusterstruktur als gegeben betrachtet wird. Das Vorgehen wird x -mal wiederholt. Die Zahl x wird von der Anwenderin spezifiziert. Hier empfiehlt sich eine iterative Vorgehensweise: Beginne zunächst mit einer kleinen Zahl, zum Beispiel mit 10, erhöhe dann diese Zahl und prüfe die Stabilität der Ergebnisse. Liegt keine Stabilität vor, erhöhe die Zahl erneut. Fahre so lange fort, bis stabile Ergebnisse vorliegen.

Für unser Beispiel sieht das Vorgehen konkret folgendermaßen aus: Die beiden Variablen besitzen Mittelwerte von 2,07 (Gesamtpunktwert für materialistische Items) bzw. von 1,60 (Gesamtpunktwert für postmaterialistische Items) und Standardabweichungen von 0,67 bzw. 0,48 (siehe Tabelle 12.8 auf Seite 316). Da 221 Befragte (Objekte) in die Analyse einbezogen wurden, werden Zufallsdaten für 221 fiktive Befragte in zwei normalverteilten Zufallsvariablen mit den obigen Verteilungskennwerten erzeugt. Für diese Zufallsdaten wird geprüft, wie gut sie durch die vorgegebene Clusterstruktur reproduziert werden können. Begonnen wird mit einer Startzahl von zehn Wiederholungen des Zufallsexperiments, ab 20 werden stabile Ergebnisse ermittelt. Die berechneten Erwartungswerte sind in Tabelle 12.7 wiedergegeben. Für die 4-Clusterlösung beträgt der Erwartungswert 0,572. Auch bei reinen Zufallsdaten würde die berechnete 4-Clusterlösung somit 57,2 Prozent der Varianz erklären. Der für die realen Daten ermittelte Wert ist 69,8 Prozent und somit um 12,6 Prozent höher. Die Standardabweichung ist 0,023 bzw. 2,3 Prozent. Aus den Kennwerten wird eine z-Teststatistik mit

$$z = \frac{t - E(t)}{\sigma(t)}$$

konstruiert, mit t als η^2 . Für die 4-Clusterlösung hat sie einen Wert von 5,528 (= $(0,698 - 0,572)/0,023$) und weicht signifikant von 0 ab ($p < 0,001$). Auch für die 8- und 10-Clusterlösung werden signifikante Abweichungen für den empirischen Wert berechnet. Für die 8-Clusterlösung ergibt sich der höchste z-Wert von allen geprüften Lösungen.

Zusammenfassend können somit alle drei Lösungen als »überzufällig« betrachtet werden. Der Befund, dass für die 8-Clusterlösung der höchste Wert der Teststatistik ermittelt wird, spricht für diese Lösung. Sie soll daher nachfolgend inhaltlich interpretiert werden.

12.4 Beschreibung und Interpretation der Cluster

Das vorrangige *Interpretationsziel des K-Means-Verfahrens* richtet sich auf eine Interpretation der Clusterzentren. Da die Aussagekraft der Mittelwerte (Clusterzentren) von der Clustergröße und den Standardabweichungen in den Clustern abhängt, müssen diese

Tab. 12.7: Erwartungswerte für die Zufallstestung

Kennwerte	4-Clusterlösung	8-Clusterlösung	10-Clusterlösung
empirischer Wert für η^2	0,698	0,860	0,891
Erwartungswert für η^2	0,572	0,704	0,776
Standardabw. des Erwartungswertes	0,023	0,014	0,017
z-Wert	5,528	11,139	6,894
Signifikanz $100 \cdot (1 - p)$	100,000	100,000	100,000
p bei Verwendung der Chebychevschen Ungleichung($1/z^2$)	0,033	0,008	0,021

Abkürzung: p : Fehlerniveau

Größen bei einer Interpretation mitberücksichtigt werden. Ziel der Interpretation ist es, Namen für die Cluster zu finden. Die Namensgebung kann sich stützen auf:

- *Interpretation der absoluten Mittelwerte.* In unserem Beispiel ist diese Strategie anwendbar. Es lässt sich feststellen, ob in einem Cluster Materialismus (oder Postmaterialismus) wichtig, weniger wichtig oder eher unwichtig ist. Voraussetzung für eine absolute Interpretation ist, dass die Ausprägungswerte inhaltlich sinnvoll interpretierbar sind. Werden statt der Gesamtpunktwerte Faktorwerte verwendet, ist eine absolute Interpretation nicht möglich. Aussagen der Art »In Cluster k liegt der Wert der Variablen x über bzw. unter dem Durchschnitt« sind möglich.
- *Vergleiche von (absoluten) Mittelwerten innerhalb eines Clusters.* In unserem Beispiel ist auch diese Strategie anwendbar. In jedem Cluster lässt sich analysieren, ob Materialismus oder Postmaterialismus wichtiger ist. Voraussetzung hierfür ist, dass die Variablen in derselben Skaleneinheit erfasst werden, nicht standardisiert sind und die Skaleneinheiten inhaltlich sinnvoll interpretierbar und vergleichbar sind.⁴
- *Unterschiede zwischen den Clustern in einer Variablen.* Dieser Vergleich ist immer zulässig. Für ihn ist das Vorliegen von absoluten Skalenwerten nicht erforderlich.

Interpretationsbasis für unser Beispiel ist Tabelle 12.8 auf der nächsten Seite. Eine grafische Darstellung der Ergebnisse ist in Form von Mittelwertprofilen möglich (siehe Abbildung 12.3 auf der nächsten Seite)⁵. Die Variablen variieren zwischen 1 und 5. Ein Wert von 1 bedeutet »sehr wichtig«, ein Wert von 2 »wichtig«, ein Wert von 3 »eher unwichtig«, ein Wert von 4 »unwichtig« und ein Wert von 5 »vollkommen unwichtig«. Die Cluster lassen sich wie folgt interpretieren:

⁴ Dies ist zum Beispiel nicht der Fall, wenn das Einkommen und das Berufsprestige verwendet werden. Eine Aussage der Art, dass im Cluster 1 das Einkommen höher ist als das Berufsprestige, ist nicht sinnvoll.

⁵ Mitunter wird auch eine andere Darstellung gewählt. Auf der Y-Achse werden dann die Variablen und nicht die Cluster abgetragen.

Tab. 12.8: Clusterzentren, Standardabweichungen, Besetzungszahlen und z-Werte für die 8-Clusterlösung der Wertedaten

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	gesamt
<i>Besetzungszahlen</i>									
n	23	54	5	5	40	18	42	34	221
in %	10,4	24,4	2,3	2,3	18,1	8,1	19,0	15,4	100,0
<i>Clustermittelwerte</i>									
gMat	2,03	1,98	2,53	4,30	2,54	3,22	1,43	1,46	2,07
gPmat	2,30	1,40	3,40	2,00	1,50	1,15	1,33	1,81	1,60
<i>Standardabweichungen</i>									
gMat	0,31	0,13	0,29	0,44	0,17	0,26	0,20	0,19	0,67
gPmat	0,24	0,23	0,58	0,28	0,23	0,17	0,17	0,17	0,48
<i>Homogenität in den Clustern</i>									
gMat	0,78	0,96	0,82	0,57	0,93	0,85	0,91	0,92	—
gPmat	0,76	0,77	-0,50	0,66	0,76	0,88	0,87	0,88	—
<i>z-Werte</i>									
gMat	-0,51	-5,33	3,24	10,15	17,22	18,52	-20,43	-17,99	—
gPmat	13,90	-6,32	6,18	2,88	-2,75	-11,15	-10,28	7,45	—

Abkürzungen: gMat: Gesamtpunktwerte Materialismus,

gPmat: Gesamtpunktwerte Postmaterialismus

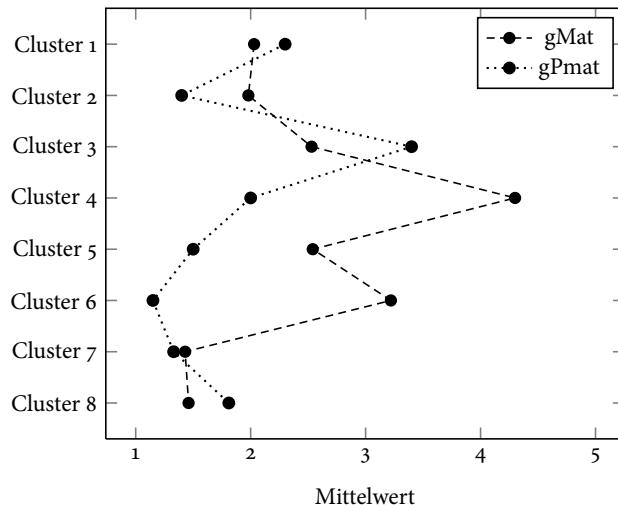


Abb. 12.3: Graphische Darstellung der Clustermittelwerte

Cluster 1: Ihm gehören 23 Schülerinnen und Schüler an. Der Mittelwert von 2,03 im Materialismus bedeutet, dass Materialismus wichtig ist. Der Mittelwert ist leicht kleiner als jener von 2,30 im Postmaterialismus. Da beide Werte jedoch nicht *sehr* wichtig sind (dies wären sie, wenn sie kleiner 1,5 lägen), soll das Cluster als »Nicht-Orientierte« bezeichnet werden (NO).

Cluster 2: Es wird von 54 Schülern gebildet. Der Mittelwert des Materialismus liegt mit 1,98 wieder nahe bei 2,00. Materialismus ist für dieses Cluster somit wichtig. Der Postmaterialismus ist mit 1,40 wichtiger. Dem Cluster wird der Name »gemäßigter Postmaterialismus« gegeben, da der Unterschied in den beiden Mittelwerten nur gering ist (GPMAT).

Cluster 3 und 4: Hier handelt es sich um kleine Restcluster, die jeweils von fünf Fällen gebildet werden. Sie sollen nicht interpretiert werden (Bezeichnung: R₁ und R₂).

Cluster 5: Ihm gehören 40 Personen an. Mit einem Mittelwert von 1,50 (Postmaterialismus) zu 2,54 (Materialismus) liegt eine klare Präferenz für Postmaterialismus vor. Diese ist noch stärker als in C₂. Daher soll dem Cluster der Name »Postmaterialismus« gegeben werden (PMAT).

Cluster 6: Diese Gruppierung hat mit einem Mittelwert von 1,15 eine noch stärkere Präferenz für den Postmaterialismus. Das Cluster soll daher als »extremer Postmaterialismus« bezeichnet werden. Ihm gehören 18 Fälle an (EPMAT).

Cluster 7: In diesem Cluster werden beide Werte (Mittelwerte von 1,43 und 1,33) für sehr wichtig gehalten. Das Cluster, das aus 42 Schüler und Schülerinnen besteht, wird als »Konsensotypus« interpretiert (KON).

Cluster 8: Cluster C₈ schließlich hat eine leichte Präferenz für materialistische Werte. Es wird als Cluster der »gemäßigten Materialisten« bezeichnet und von 34 Befragten gebildet (GMAT).

Für den Vergleich von Clustern in einer Variablen können auch die in Tabelle 12.8 ausgewiesenen z-Werte eingesetzt werden. Sie sind wie folgt definiert:

$$\bar{z}_{ki} = \frac{\bar{x}_{ki} - \bar{x}_i}{\bar{s}_{ki}}, \quad (12.7)$$

wobei \bar{x}_{ki} der Mittelwert bzw. Anteilswert des Clusters k in der Variablen i ist, \bar{s}_{ki} ist die Standardabweichung des Mittelwerts bzw. des Anteilswerts des Clusters k in der Variablen i und \bar{x}_i schließlich ist der Gesamtmittelwert. Dieser z-Wert unterscheidet sich von dem aus den empirisch standardisierten Werten berechneten z-Wert

$$\bar{z}_{ki} = \frac{\bar{x}_{ki} - \bar{x}_i}{s_i} \quad (12.8)$$

dadurch, dass im Nenner nicht die Gesamtstandardabweichung steht, sondern die Standardabweichung des Clusterzentrums. Dies hat den Vorteil, dass tatsächlich nur die

Streuung des Clusters in die Testung eingeht, während die Gesamtstandardabweichung auch die Varianzen in den anderen Clustern und die Varianzen zwischen den Clustern enthält. Werte größer 0 bedeuten, dass der Mittelwert größer als der Gesamtmittelwert ist, Werte kleiner 0 das Gegenteil. Betrachten wir beispielsweise die Werte für den Materialismus: C_1 mit einem z-Wert von $-0,51$ weicht kaum vom Gesamtmittelwert ab. Die Mittelwerte von C_7 und C_8 liegen im Materialismus mit z-Werten von $-20,43$ und $-17,99$ deutlich unter dem Gesamtmittel. Auf der Basis der z-Werte sind keine Aussagen möglich, ob innerhalb eines Clusters Materialismus oder Postmaterialismus wichtiger ist, da es sich um standardisierte Parameter handelt.

Bei der Interpretation der Clustermittelwerte ist auch die Standardabweichung zu berücksichtigen. Eine absolute Interpretation der Standardabweichungen ist aber oft schwierig. Daher wird in Tabelle 12.8 auf Seite 316 zusätzlich ein *Homogenitätsindex* je Variable und Cluster ausgewiesen. Er wird wie folgt berechnet:

$$\text{HOMO}_{kj} = 1 - \frac{s_{kj}^2}{s_j^2}.$$

Die Varianz der Variable j im Cluster k wird mit s_{kj}^2 , die Varianz der Variablen j mit s_j^2 bezeichnet. Ist ein Cluster in einer Variablen vollkommen homogen, ist $s_{kj}^2 = 0$ und $\text{HOMO}_{kj} = 1$. Kleinere Werte weisen auf eine geringere Homogenität hin. Ein Wert von 0 bedeutet, dass die Varianz im Cluster gleich der Gesamtvarianz der Variablen ist. Bis auf die Restcluster werden für alle anderen Variablen Werte größer 0,70 erzielt. Es fehlen zwar Schwellenwerte, aber man kann sich an denjenigen für Cronbachs α (Cronbach 1951) orientieren, wo üblicherweise ein Wert von $> 0,70$ gefordert wird (Schmitt 1996).

Bei der Interpretation der Cluster werden in der Regel Unterschiedshypothesen aufgestellt. Es ist sinnvoll, diese statistisch zu prüfen, dabei sind zwei Arten zu unterscheiden:

- *Unterschied zwischen zwei Clustern in einer Variablen:* Cluster x unterscheidet sich in der Variablen y von Cluster z . Zur Prüfung dieser Hypothese eignet sich der t-Test für unabhängige Stichproben (Student 1908; Bortz 2005, S. 140–143) bzw. ein äquivalenter Test, wie zum Beispiel der Median-Test (Bortz, Lienert u. a. 2008, S. 198–200) und der Mann-Whitney- bzw. U-Test (Mann und Whitney 1947; Wilcoxon 1945; Bortz, Lienert u. a. 2008, S. 200–212) für ordinale sowie der χ^2 -Unabhängigkeitstest (Pearson 1900; Bortz 2005, S. 154–177) für nominale Variablen.
- *Unterschied von Variablen innerhalb eines Clusters:* Cluster x hat in der Variablen y einen höheren Wert als in der Variablen z . Für die Prüfung dieser Hypothese kann der t-Test für abhängige Stichproben (Student 1908; Bortz 2005, S. 143–146) bzw. ein äquivalenter Test, wie zum Beispiel der Wilcoxon-Test (Wilcoxon 1945, 1947; Bortz, Lienert u. a. 2008, S. 259–267) für ordinale sowie der McNemar-Test (McNemar 1947; Bortz, Lienert u. a. 2008, S. 160–164) für nominale Variablen eingesetzt werden.

Zur Prüfung der Hypothesen wird als neue Variable die Clusterzugehörigkeit im Datensatz abgespeichert. Anschließend wird der entsprechende Test gerechnet. Das Vorgehen soll an zwei Beispielen verdeutlicht werden: Bei der Interpretation der Cluster C_2 , C_5 und C_6 wurde angenommen, dass in C_5 (Postmaterialisten) der Materialismus stärker abgelehnt wird als in C_2 (gemäßigte Postmaterialisten) und dass in C_6 (extreme Postmaterialisten) wiederum die Ablehnung des Materialismus größer ist als in C_5 . Dies kann durch zwei t-Tests für unabhängige Stichproben geprüft werden. Als zweites Beispiel soll untersucht werden, ob in C_8 tatsächlich eine Präferenz für Materialismus besteht, ob also der Mittelwert im Materialismus signifikant kleiner ist als jener im Postmaterialismus.

Die entsprechenden Tests resultieren für den Vergleich von C_2 und C_5 im Materialismus in:

$$\begin{array}{lll} \bar{x}_2 = 1,98 & \bar{x}_5 = 2,54 & d = \bar{x}_2 - \bar{x}_5 = -0,567 \\ s_2 = 0,13 & s_5 = 0,17 & t = -18,200 \\ n_2 = 54 & n_5 = 40 & p = 0,000 . \end{array}$$

Die Mittelwertdifferenz ist d , t ist der t-Testwert und p das Fehlerniveau. Die Interpretation kann bestätigt werden: In C_5 wird der Materialismus stärker abgelehnt als in C_2 .

Vergleich von C_5 und C_6 im Materialismus:

$$\begin{array}{lll} \bar{x}_5 = 2,54 & \bar{x}_6 = 3,22 & d = \bar{x}_5 - \bar{x}_6 = -0,677 \\ s_2 = 0,17 & s_6 = 0,26 & t = -11,602 \\ n_2 = 40 & n_6 = 18 & p = 0,000 . \end{array}$$

Auch diese Interpretation kann bestätigt werden: In C_6 wird der Materialismus stärker abgelehnt als in C_5 .

Vergleich von Materialismus (x) und Postmaterialismus (y) in C_8 (t-Test für abhängige Stichproben):

$$\begin{array}{lll} \bar{x}_8 = 1,46 & d = \bar{x}_8 - \bar{y}_8 = -0,35 \\ \bar{y}_8 = 1,81 & t = 7,582 \\ n_8 = 42 & p = 0,000 . \end{array}$$

Auch die Interpretation, dass in C_8 der Materialismus wichtiger ist als der Postmaterialismus, kann aufrecht erhalten werden.

Hingewiesen sei wiederum darauf, dass die Durchführung der Tests nicht ganz korrekt ist. Mit der Maximierung der Streuung zwischen den Clustern werden auch die Unterschiede in den Mittelwerten zwischen den Clustern maximiert. Der t-Test für Unterschiede

Tab. 12.9: Paarweise Differenzen für Cluster C_1 zu den anderen Clustern

Variable	C_2	C_3	C_4	C_5	C_6	C_7	C_8
gMat	=	=	<	<	<	>	>
gPmat	>	<	=	>	>	>	>

Abkürzungen: siehe Tabelle 12.8 auf Seite 316

zwischen den Clustern ist daher strenggenommen nicht zulässig. Da in die Maximierungsaufgabe aber viele paarweise Mittelwertunterschiede einfließen, ist dieser Effekt gering und die empfohlen Tests können durchgeführt werden. Die Anwenderin sollte sich aber des grundlegenden Problems bewusst sein.

Das Computerprogramm ALMO führt innerhalb der Clusteranalyse automatisch paarweise t-Tests für die Cluster durch. Die Testgröße ist definiert als (Kriz 1978, S. 134–137)

$$t = \frac{\bar{x}_{kj} - \bar{x}_{k^*j^*}}{\sqrt{\left(n_{kj} \cdot s_{kj}^2 + n_{k^*j^*} \cdot s_{k^*j^*}^2 \right) / (n_{kj} + n_{k^*j^*} - 2)}} \cdot \sqrt{\frac{n_{kj} \cdot n_{k^*j^*}}{n_{kj} + n_{k^*j^*}}},$$

wobei s_{kj}^2 die mit n_{kj} normierte Varianz im Cluster k in der Variablen j ist. Die Zahl der Fälle im Cluster k in der Variablen j wird mit n_{kj} bezeichnet. Analog sind für das Cluster k^* und die Variable j^* die Größen $s_{k^*j^*}^2$ und $n_{k^*j^*}$ definiert.

In Tabelle 12.9 werden beispielhaft die paarweisen Differenzen für das Cluster C_1 angeführt. Die Ergebnisse sind wie folgt zu lesen: C_1 unterscheidet sich im Materialismus nicht signifikant von C_2 und C_3 (»=«) und hat im Vergleich zu den Clustern C_4 , C_5 und C_6 einen kleineren Mittelwert (»<«) im Materialismus. Im Unterschied dazu ist der Mittelwert im Materialismus im Vergleich zu C_7 und C_8 größer (»>«). Bezuglich des Postmaterialismus treten folgende Unterschied auf: C_1 hat im Vergleich zu C_3 einen kleineren Mittelwert, unterscheidet sich jedoch nicht von C_4 . Im Vergleich zu allen anderen Clustern hat C_1 einen höheren Wert. Als Schwellenwert wurde eine Signifikanz von 95 Prozent festgelegt ($p < 0,05$ zweiseitig, $p < 0,025$ einseitig). Da Mehrfachtestungen durchgeführt werden, nimmt das Programm automatisch eine Bonferroni-Korrektur vor (Miller 1981, zitiert in SAS Institute 1991, S. 595).

Für Fragestellungen, die sich auf Unterschiede innerhalb eines Clusters beziehen, führt ALMO eine Art hierarchische Clusteranalyse durch und verschmilzt jene Variablen, die sich nicht signifikant unterscheiden, zu einer Variablengruppe.

12.5 Formale Beschreibung der Cluster

Häufig ist auch eine Beurteilung des *Beitrags der Klassifikationsmerkmale zur Trennung der Cluster* von Interesse. Dazu lassen sich für jede Variable die erklärte Varianz $\eta_{j|K}^2$ und ein F-Wert berechnen mit:

$$\eta_{j|K}^2 = 1 - \frac{\text{SQ}_{\text{in}}(j|K)}{\text{SQ}_{\text{ges}}(j)} \quad \text{und}$$

$$F_{j|K} = \frac{\text{SQ}_{\text{zw}}(j|K)/(K-1)}{\text{SQ}_{\text{ges}}(j)/(n_j-1)}.$$

$\text{SQ}_{\text{in}}(j|K)$ ist die Streuungsquadratsumme in der Variablen j bei K gegebenen Clustern, $\text{SQ}_{\text{zw}}(j|K)$ ist die Streuungsquadratsumme in der Variablen j zwischen den K Clustern und $\text{SQ}_{\text{ges}}(j)$ ist die Gesamtstreuungssquaretsumme in der Variablen j .

Der F-Wert wird häufig auf Signifikanz geprüft. Die Annahme einer F-Verteilung als Nullmodell ist allerdings strenggenommen nicht zulässig, da mit der Minimierung der Streuungsquadratsumme in den Clustern auch die Streuungsquadratsumme in jeder Variablen minimiert wird. Werden aber sehr viele Variablen in die Clusteranalyse einbezogen, hat jede Variable ein geringeres Gewicht. Die Annahme der F-Verteilung als Nullmodell ist dann annähernd erfüllt. Ein Schwellenwert, ab dem eine Variablenzahl als ausreichend groß zu betrachten ist, liegt leider nicht vor. Da wir nur zwei Variablen untersucht haben, trifft diese Argumentation in unserem Fall jedoch nicht zu. Die Signifikanzen sind daher mit Vorsicht zu interpretieren. Leisten eine oder mehrere Variablen keinen signifikanten Beitrag zur Trennung der Cluster, kann bzw. können sie zur Prüfung der Stabilität als Klassifikationsvariable eliminiert werden (siehe Abschnitt 12.7). Stabilität liegt dann vor, wenn das Entfernen von nicht signifikanten Variablen zu keiner Änderung der Clusterergebnisse führt. In unserem Beispiel ergeben sich die Werte der Tabelle 12.10 auf der nächsten Seite. Die erklärte Streuung des Materialismus ist mit 0,901 bzw. 90,1 Prozent größer als jener von 0,789 bzw. 78,9 Prozent für den Postmaterialismus. Materialistische Wertorientierungen leisten somit einen höheren Beitrag zur Trennung der Cluster, dies bringt auch der größere F-Wert zum Ausdruck. Da nur zwei Variablen untersucht werden, werden die Signifikanzen in diesem Beispiel nicht genutzt.

Für einen Detailblick auf die Cluster berechnet ALMO zusätzlich einen Homogenitäts- und einen Heterogenitätsindex:

$$\text{HOMO}_k = 1 - \frac{s_k^2}{s_{\text{ges}}^2} = 1 - \frac{\frac{1}{m} \sum_j s_{kj}^2}{\frac{1}{m} \sum_j s_j^2} \quad \text{und}$$

$$\text{HETERO}_k = 1 - \frac{s_{\text{ges}}^2}{d_{kk}^2} = 1 - \frac{\frac{1}{m} \sum_j s_j^2}{\sum_{k^* \neq k} n_k \cdot n_{k^*} \cdot d_{kk^*}^2 / \sum_{k^* \neq k} n_k \cdot n_{k^*}},$$

Tab. 12.10: Gesamtstatistiken für Klassifikationsmerkmale

Variable	F-Wert	Signifikanz $(1 - p) \cdot 100$	η^2
gMat	276,055	100,0	0,901
gPmat	108,177	100,0	0,780

Abkürzungen: siehe Tabelle 12.8 auf Seite 316

mit s_k^2 als Varianz im Cluster k , $s_{kj}^2 = \frac{1}{n_k} \sum_g (x_{gj} - \bar{x}_{kj})^2$ als Varianz des Merkmals j im Cluster k und $d_{kk^*}^2 = \sum_j (\bar{x}_{kj} - \bar{x}_{k^*j})^2$ als quadrierte euklidische Distanz der Clusterzentren von Cluster k und Cluster k^* .

Beim *Homogenitätsindex* $HOMO_k$ des Clusters k wird die Streuung innerhalb des Clusters k in Beziehung zur Gesamtstreuung gesetzt. Ist die Streuung innerhalb des Clusters 0 – sind die Homogenitätsvorstellungen also perfekt erfüllt –, nimmt der Index einen Wert von 1 an. Ist die Streuung innerhalb der Cluster dagegen größer 0, ist der Index kleiner 1. Je kleiner der Index ist, desto schlechter ist die Homogenitätsvorstellung in den Cluster erfüllt.

Beim *Heterogenitätsindex* $HETERO_k$ des Clusters k werden die Distanzen des Clusters k zu den anderen Clusterzentren berechnet und zur Gesamtstreuung in Beziehung gesetzt. Wenn die durchschnittlichen Distanzen zu den anderen Clustern sehr groß sind, wird durch eine große Zahl dividiert, so dass der zweite Ausdruck nahe 0 und der Index nahe 1 ist.

Tab. 12.11: Ergebnisse der formalen Gültigkeitsprüfung für die 8-Clusterlösung

Cluster	n	Varianz in den Clustern	Homogenität in den Clustern $HOMO_k$	durchschnittl. Distanz zu den anderen Clusterzentren	Heterogenität zw. Clustern $HETERO_k$
C_1	23	0,076	0,774	1,192	0,716
C_2	54	0,034	0,900	0,829	0,592
C_3	5	0,211	0,376	4,132	0,981
C_4	5	0,136	0,599	5,851	0,942
C_5	40	0,042	0,875	0,995	0,660
C_6	18	0,047	0,862	2,294	0,853
C_7	42	0,034	0,898	1,307	0,741
C_8	34	0,033	0,904	1,176	0,712
gesamt	221	0,400	0,861	1,349	0,701

Für die 8-Clusterlösung ergeben sich die in Tabelle 12.11 dargestellten Werte: Die 8-Clusterlösung weist zwei Cluster mit einer geringen Homogenität aus, nämlich die beiden Restcluster C_3 und C_4 , die jeweils nur aus fünf Fällen bestehen. Umgekehrt sind diese beiden Cluster durch eine große Heterogenität zu den anderen Clustern gekennzeichnet. Die Heterogenitätsindizes sind deutlich größer 0,900, es ist daher zu vermuten, dass es sich bei den Restclustern um Ausreißer handelt.

Weitere Maßzahlen zur formalen Beschreibung der Cluster lassen sich auf Basis der Distanzen der Objekte zu dem Clusterzentrum, dem das Objekt angehört, berechnen. Die Distanz eines Objekts g zu seinem Clusterzentrum ist

$$d_{g,k}^2 = \sum_i (x_{gj} - \bar{x}_{kj})^2 .$$

Es ist sinnvoll, die Distanz zu normieren. Als Normierungsgröße kann die mittlere Gesamtstreuungssquaratsumme (SQ_{ges}/n) verwendet werden. In unserem Beispiel ist diese Bezugsgröße gleich 0,6733 (= 148,805/221). Sie soll mit d_o^2 bezeichnet werden. Sie ist gleich der mittleren quadrierten euklidischen Distanz eines Objekts g in der 1-Clusterlösung. Mittels d_o^2 lassen sich nun *standardisierte Distanzen* berechnen mit

$$d_g^{\text{stand}} = \sqrt{\frac{d_{g,k}^2}{d_o^2}} .$$

Für eine brauchbare Clusterlösung ist wünschenswert, dass es eine bestimmte Anzahl von Objekten gibt, die nahe dem Clusterzentrum liegen und deren standardisierte Distanz daher deutlich kleiner 1,0 sind. Umgekehrt sollte es wenige Objekte mit großer standardisierter Distanz geben. Sind die Cluster deutlich voneinander getrennt, sollte es schließlich auch wenige Objekte geben, die gleich weit oder gleich nahe zu zwei oder mehreren Clusterzentren liegen. Objekte, die sehr nahe am Clusterzentrum liegen, sollen als *Repräsentanten* bezeichnet werden. Objekte, die weit von ihrem Clusterzentrum entfernt sind, als *Ausreißer*, und Objekte, die gleich weit zu zwei Clusterzentren entfernt sind, als *Objekte im Überlappungsbereich bzw. Überlappungen*.

Bei der 8-Clusterlösung ergibt sich das in der Tabelle 12.12 auf der nächsten Seite dargestellte Bild: Für Cluster C_1 werden drei Repräsentanten gefunden, Ausreißer liegen nicht vor und drei Objekte liegen im Überlappungsbereich. Für Cluster C_2 liegen 20 Repräsentanten und ebenfalls kein Ausreißer vor, zwei Objekte weisen Überlappungen auf. Repräsentanten, Ausreißer und Überlappungen wurden wie folgt definiert:

Repräsentant: $d_g^{\text{stand}} < 0,25$,

Ausreißer: $d_g^{\text{stand}} > 1,25$,

Überlappung: $d_g^{\text{stand}} / d_{g^*}^{\text{stand}} < 1,25$ und $d_g^{\text{stand}} < 1,25$.

Tab. 12.12: Formale Beschreibung der Cluster

Cluster	Repräsentanten		Ausreißer		Überlappungen	
	Anzahl	Prozent	Anzahl	Prozent	Anzahl	Prozent
C_1	3	13,043	0	0,000	3	13,043
C_2	20	37,037	0	0,000	2	3,704
C_3	0	0,000	1	20,000	0	0,000
C_4	1	20,000	0	0,000	0	0,000
C_5	13	32,500	0	0,000	2	5,000
C_6	5	27,778	0	0,000	2	11,111
C_7	22	52,381	0	0,000	5	11,905
C_8	15	44,118	0	0,000	1	2,941

Die absoluten Ergebnisse sind schwer zu interpretieren. Sie können aber als formale Kriterien in eine formale Gültigkeitsprüfung einbezogen werden, bei der mehrere formal zulässige Lösungen verglichen werden (siehe Kapitel 20).

Zur Beschreibung und Beurteilung einer Clusterlösung wird mitunter eine *Diskriminanzanalyse* (Fisher 1936) bzw. äquivalent ein *allgemeines lineares Modell* (Cohen 1968; Overall und Spiegel 1969) oder alternativ eine *multinominale logistische Regression* (Agresti 2002) gerechnet. Untersucht wird, wie gut die gefundene Klassifikation ist und welchen partiellen Beitrag dazu jede Variable leistet. Im Unterschied zu Tabelle 12.10 auf Seite 322 wird nicht der pauschale Beitrag einer Variablen zur Trennung der Cluster betrachtet, sondern der isolierte Beitrag der Variablen unter der Annahme, dass die anderen Variablen konstant sind. Das Vorgehen besteht darin, dass die Clusterzugehörigkeit im Datensatz abgespeichert und anschließend als nominale abhängige Variable in eine Diskriminanzanalyse (oder äquivalent in ein allgemeines lineares Modell) oder eine multinomiale logistische Regression einfließt. Die Klassifikationsmerkmale werden als unabhängige Variablen in die Analysen einzbezogen.

Exemplarisch sollen die Ergebnisse für ein allgemeines lineares Modell (Holm 1979) auszugsweise wiedergegeben werden. Die Clusterzugehörigkeit ist die abhängige Variable, die Klassifikationsvariablen die unabhängigen. Die erklärte Varianz des Gesamtmodells beträgt 84,1 Prozent (kein Output wiedergegeben).⁶ Dem Materialismus (gMat) kommt eine größere Erklärungskraft zu als dem Postmaterialismus (gPmat, siehe Tabelle 12.13). Die ermittelten Effekte sind in der Tabelle 12.14 auf Seite 326 abgebildet. Der positive

⁶ Dieser Wert ist etwas kleiner als die durch die 6-Clusterlösung erklärte Varianz von 86,0 Prozent, da sich im allgemeinen linearen Modell die unabhängigen Variablen gegenseitig Erklärungskraft »wegnehmen«, wenn sie korrelieren. In dem Beispiel korrelieren der Materialismus und Postmaterialismus nur schwach, so dass die erklärte Varianz des allgemeinen linearen Modells nur gering von der beim K-Means-Verfahren ausgewiesenen Varianz abweicht.

Tab. 12.13: Ergebnisse des allgemeinen linearen Modells – Erklärungskraft der Variablen

	partielle Korrelation	Wilks Lambda	F-Wert	df ₁	df ₂	p
gMat	0,9491	0,0993	274,7500	7	212	0,000
gPmat	0,8834	0,2195	107,6653	7	212	0,000

Abkürzungen: siehe Tabelle 12.8 auf Seite 316

Effekt von 0,3206 für das Cluster C_1 im Postmaterialismus bedeutet, dass der Wert des Clusters über dem Gesamtmittel liegt. Der Mittelwert im Materialismus unterscheidet sich nicht vom Gesamtmittel (Effekt liegt bei -0,0100). Die anderen Werte sind analog zu interpretieren.

Das grundlegende Problem, dass strenggenommen statistische Signifikanztests nicht möglich sind, da die Streuungsquadratsumme zwischen den Clustern maximiert wird, bleibt dabei natürlich ungelöst und zwar unabhängig davon, welches Verfahren ausgewählt wird. Ob die Entscheidung auf die Diskriminanzanalyse oder die multinomiale logistische Regression fällt, hängt zum Teil von der Vorliebe des Anwenders ab. Die multinomiale logistische Regression ist das »neuere« (moderne) Verfahren. Sie weist häufig einen höheren Anteil korrekter Klassifikationen aus als die Diskriminanzanalyse, wenn die Clustergrößen streuen und es wenige Cluster mit großen Besetzungszahlen und viele kleine Cluster gibt. Bei dieser Datenkonstellation werden die Prognosewahrscheinlichkeiten primär für die großen Cluster maximiert, so dass sich ein höherer Anteil an korrekten Klassifikationen bei der logistischen Regression ergibt.

12.6 Analyse von Ausreißern

Werden auf der Grundlage der Homogenitäts- und Heterogenitätsindizes *Ausreißer* vermutet, ist die Durchführung einer *Ausreißeranalyse* hilfreich. Dazu können die im vorausgehenden Abschnitt definierten standardisierten Distanzen verwendet werden. Für die 8-Clusterlösung gibt es nur ein Objekt mit einer standardisierten Distanz größer 1,25 (siehe Tabelle 12.15 auf Seite 327). Es handelt sich um das Objekt 216 mit $d_g^{\text{stand}} = \sqrt{d_{g,k}^2/d_0^2} = 1,27$. Es gehört Cluster 3 an, das mit 0,376 auch den geringsten Wert des Homogenitätsindex hat. Es folgt ein Objekt (173) aus Cluster 4, das den zweitgeringsten Wert des Homogenitätsindex hat, gefolgt von drei Objekten (46, 172, 170) des Clusters 1. Die kleinsten Distanzen werden für die Objekte (77) bis (188) zu den Clustern 5, 8, 1 und 6 ausgewiesen. Insgesamt wird man den Schluss ziehen, dass für die 8-Clusterlösung

Tab. 12.14: Effekte der Klassifikationsvariablen auf die Clusterzugehörigkeit

	stand. Effekt	Effekt	part. Korrelation	<i>p</i>
<i>C₁: Nicht-Orientierte (NO)</i>				
gMat	-0,0218	-0,0100	-0,025	0,710
gPmat	0,4999	0,3206	0,500	0,000
<i>C₂: gemäßiger Postmaterialismus (GPMAT)</i>				
gMat	-0,0769	-0,0494	-0,079	0,242
gPmat	-0,2349	-0,2120	-0,236	0,000
<i>C₃: Rest 1 (R₁)</i>				
gMat	0,1006	0,0224	0,123	0,068
gPmat	0,5746	0,1795	0,578	0,000
<i>C₄: Rest 2 (R₂)</i>				
gMat	0,5068	0,1128	0,511	0,000
gPmat	0,1236	0,0386	0,143	0,033
<i>C₅: Postmaterialismus (PMAT)</i>				
gMat	0,3347	0,1928	0,336	0,000
gPmat	-0,1050	-0,0849	-0,111	0,099
<i>C₆: extreme Postmaterialisten (EPMAT)</i>				
gMat	0,5157	0,2110	0,537	0,000
gPmat	-0,2855	-0,1640	-0,333	0,000
<i>C₇: Konsenstypus (Kon)</i>				
gMat	-0,4587	-0,2693	-0,478	0,000
gPmat	-0,2743	-0,2261	-0,309	0,000
<i>C₈: gemäßige Materialisten (GMAT)</i>				
gMat	-0,3897	-0,2104	-0,397	0,000
gPmat	0,1957	0,1483	0,212	0,002

Abkürzungen: siehe Tabelle 12.8 auf Seite 316

keine bzw. wenige Ausreißer vorliegen, da diese in den kleinen Restclustern versammelt sind.

Zur Ermittlung von Ausreißern kann auch der *Single-Linkage* eingesetzt werden (siehe Abschnitt 9.2). Ausreißer sind dadurch erkennbar, dass sie im Verschmelzungsschema sehr lange »alleine« bleiben – also erst in späten Verschmelzungsschritten mit anderen Clustern verknüpft werden. Eine Anwendung des Single-Linkage für das Beispiel liefert das Ergebnis der Abbildung 12.4. Das Objekt 216 ist wiederum klar als Ausreißer erkennbar. Es wird erst im letzten Schritt mit den anderen Clustern fusioniert. Ebenfalls als Ausreißer erkennbar sind die Objekte 218 und 173. Das Objekt 164 ist hingegen kein Ausreißer, da es bereits früher mit Objekt 186 verschmolzen wird. Der Single-Linkage identifiziert somit drei Ausreißer – auch diese Zahl ist gering. Die Ergebnisse der Ausrei-

Tab. 12.15: Standardisierte Entferungen der Objekte zu den Clustern

Objekt	Cluster	Distanz	stand. Distanz	Objekt	Cluster	Distanz	stand. Distanz
216	3	1,09	1,27	usw.			
173	4	0,75	1,06	158	6	0	0,07
46	1	0,57	0,92	215	6	0	0,07
172	1	0,42	0,79	139	1	0	0,06
170	1	0,40	0,77	77	8	0	0,05
64	8	0,36	0,73	80	5	0	0,05
221	6	0,36	0,73	85	5	0	0,05
21	1	0,35	0,72	111	8	0	0,05
115	3	0,34	0,71	125	8	0	0,05
180	5	0,32	0,69	128	8	0	0,05
218	3	0,32	0,69	171	5	0	0,05
usw.				188	5	0	0,05

grau hinterlegt: Ausreißerobjekt 216

ßeranalyse mittels Single-Linkage und K-Means-Verfahren weichen voneinander ab, da verschieden vorgegangen wird und unterschiedliche Modellvorstellungen vorliegen.

Manche Computerprogramme, wie zum Beispiel CLUSTAN (Wishart 2003) oder auch die Prozedur »TwoStep« in SPSS (siehe Abschnitt 17.2) haben in der Zwischenzeit Optionen zur Ausreißerbehandlung implementiert. Dabei kann die Benutzerin einen bestimmten Schwellenwert definieren. Bei der Bildung der Cluster werden die Ausreißer ausgeschlossen, können aber dann wieder zugeordnet werden. Das Erfahrungswissen über diese Methoden ist noch gering, so dass Empfehlungen schwer zu geben sind. Ein automatisches Vorgehen hat zweifelsohne seinen Charme, ist aber oft auch problematisch. Werden die oben genannten Optionen eingesetzt, empfehlen wir, auch eine Analyse ohne diese Optionen zu rechnen und die Ergebnisse zu vergleichen.

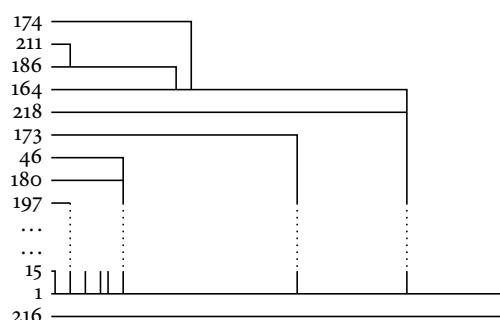


Abb. 12.4: Ausreißerdiagnose mittels Single-Linkage

12.7 Stabilitätsprüfung

Eine Clusterlösung sollte gegenüber geringfügigen Modifikationen der Daten sowie der getroffenen Spezifikationen stabil sein. Die *Stabilitätsprüfung* erfordert daher eine Festlegung, welche (geringfügigen) Modifikationen untersucht werden, sowie ein Kriterium zur Beurteilung der Stabilität. Das Kriterium kann sich beziehen auf die:

- *Clusterzahl*: Eine Lösung wird als stabil betrachtet, wenn geringfügige Modifikationen die Clusterzahl nicht ändern.
- *Clusterzentren* für eine bestimmte Clusterzahl: Eine Lösung wird als stabil betrachtet, wenn geringfügige Modifikationen die Clusterzentren nicht ändern.
- *Zuordnung der Objekte* zu den Clusterzentren: Eine Lösung wird als stabil bezeichnet, wenn geringfügige Modifikationen die Zuordnung der Objekte nicht ändern.

Das letzte Kriterium ist das strengste. Ändert sich die Zuordnung der Objekte nicht, so ändern sich die Clusterzentren ebenfalls nicht. Umgekehrt können die Clusterzentren stabil bleiben, auch wenn sich ein bestimmter Prozentsatz an Zuordnungen ändert. Das letzte Kriterium der Stabilität der Zuordnung der Fälle lässt sich technisch mit allen Computerprogrammen zur Clusteranalyse realisieren, wenn diese das Speichern der Clusterzugehörigkeit zulassen. Das Vorgehen besteht darin, dass die untersuchten Clusterlösungen, anhand derer die Stabilität beurteilt werden soll, miteinander kreuztabelliert werden. In den Tabellen 12.16 und 12.17 wird die Stabilität hinsichtlich der 8-Clusterlösung untersucht. Es wird der Frage nachgegangen, ob sich die 8-Clusterlösung auch in der 7- und in der 9-Clusterlösung wiederauffinden lässt. Die Stabilität der 8-Clusterlösung ist gut erkennbar. Bei der 7-Clusterlösung bleiben sechs der acht Cluster weitgehend bestehen. Nur wenige Objekte ändern ihre Zuordnung. Cluster 7 und 8 werden zu einem Cluster verschmolzen. Umgekehrt wird bei der 9-Clusterlösung das Cluster 2 der 8-Clusterlösung in zwei Cluster aufgespalten. Die anderen Cluster bleiben weitgehend erhalten; dies ist ein deutlicher Hinweis auf Stabilität.

Neben einer Inspektion der Kreuztabellen kann auch der *Koeffizient κ* , der *RAND-Index* oder der *adjustierte RAND-Index* berechnet werden (siehe Abschnitte 8.3 und 9.5).

Nach der Festlegung des Kriteriums ist zu entscheiden, welche Modifikationen untersucht werden. Modifikationen zur Prüfung der Stabilität können sein:

1. *Stabilität bezüglich der Variablenauswahl*:

- Entfernen irrelevanter Variablen: Die Ergebnisse sollten sich wiederum nicht ändern, wenn Variablen, die keinen Beitrag zur Trennung der Cluster leisten, entfernt werden.

Tab. 12.16: Zusammenhang der 7-Clusterlösung mit der 8-Clusterlösung

Cluster (CL ₇)	Cluster (CL ₈)								gesamt
	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	C ₈	
C ₅	0	52	0	0	0	0	6	0	58
C ₇	0	0	0	0	0	0	36	28	64
C ₁	1	0	5	0	0	0	0	0	6
C ₆	0	0	0	0	0	18	0	0	18
C ₃	0	0	0	5	0	0	0	0	5
C ₄	1	0	0	0	40	0	0	0	41
C ₂	21	2	0	0	0	0	0	6	29
gesamt	23	54	5	5	40	18	42	4	221

grau hinterlegt: stärkste Zellbesetzungen

- Hinzufügen irrelevanter Variablen: Die Ergebnisse sollten sich nicht ändern, wenn irrelevante Variablen hinzugefügt werden. Dies kann dadurch realisiert werden, dass eine oder mehrere normalverteilte Zufallsvariablen in die Analyse einbezogen werden.
2. Stabilität bezüglich der Objektauswahl:
- Entfernen von Ausreißern: Die Kerncluster sollten bestehen bleiben, wenn Ausreißer eliminiert werden, Restcluster sollten verschwinden.
 - Zufällige Aufteilung auf zwei Stichproben: Der Datensatz wird in zwei Stichproben zerlegt und für jede Stichprobe wird eine getrennte Clusteranalyse durchgeführt (*Split-Half-Methode*). Die Ergebnisse sollten weitgehend übereinstimmen. Eine

Tab. 12.17: Zusammenhang der 8-Clusterlösung mit der 9-Clusterlösung

Cluster (CL ₈)	Cluster (CL ₉)									gesamt
	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	C ₈	C ₉	
C ₁	0	0	1	1	0	21	0	0	0	23
C ₂	27	0	0	0	0	0	0	0	27	54
C ₃	0	0	5	0	0	0	0	0	0	5
C ₄	0	5	0	0	0	0	0	0	0	5
C ₅	1	0	0	39	0	0	0	0	0	40
C ₆	0	0	0	0	0	0	18	0	0	18
C ₇	2	0	0	0	35	0	0	0	5	42
C ₈	0	0	0	0	0	0	0	30	4	34
gesamt	30	5	6	40	35	21	18	30	36	221

grau hinterlegt: stärkste Zellbesetzungen

alternative Strategie könnte darin bestehen, dass in einer Stichprobe die Clusterzentren geschätzt werden und in der anderen Stichprobe untersucht wird, wie gut die gefundene Clusterlösung zur Reproduktion der Cluster geeignet ist.

3. Stabilität bezüglich der Modellspezifikationen:

- Erhöhung bzw. Reduktion der Clusterzahl: Die untersuchte Clusterlösung – insbesondere die Clusterzentren – sollte sich auch wiederfinden, wenn die Clusterzahl minimal erhöht oder minimal reduziert wird (siehe oben).
- Modifikation des Varianzkriteriums und weiterer Modellparameter (siehe dazu die Abschnitte 12.11 und 12.12).
- Modifikation des Verfahrens: Analyse mit einem anderen Clusteranalyseverfahren, wie zum Beispiel das Ward-Verfahren (siehe Abschnitt 9.5).

Wir empfehlen die Durchführung mehrerer Stabilitätstests. Es werden also p Clusterlösungen $\text{CL}_1, \text{CL}_2, \dots, \text{CL}_p$ berechnet und miteinander entweder hinsichtlich der Mittelwerte oder hinsichtlich der Zuordnung verglichen. Der Vergleich der Zuordnungen der Objekte zu den Clustern ist bis auf das dargestellte Split-Half-Verfahren möglich – der Vergleich der Mittelwerte bei allen dargestellten Tests.

Für den Vergleich der Mittelwerte kann man folgende Strategie wählen, sofern kein spezielles Computerprogramm zur Verfügung steht:

- *Die Clustermittelwerte für die untersuchten Lösungen werden als neue Datenmatrix abgespeichert.*
- *Die (Un-)Ähnlichkeiten der Clustermittelwerte werden berechnet*, zum Beispiel mittels eines Programms zur hierarchischen Clusteranalyse.
- *Anhand der (Un-)Ähnlichkeitsmatrix wird die Stabilität beurteilt.* Werden zum Beispiel zwei Clusterlösungen CL_1 und CL_2 mit K Clustern verglichen, so sollte es K sehr ähnliche Paare geben. Jedes Paar wird von einem Cluster einer Lösung gebildet. Wird eine Lösung mit $K - 1$ Clustern mit K Clustern verglichen, so sollten $K - 2$ Cluster vorliegen, die jeweils aus einem Clusterpaar gebildet werden, das ein Cluster aus CL_1 und ein Cluster aus CL_2 enthält.
- *Ergänzend kann noch eine hierarchische Clusteranalyse durchgeführt werden.* Auch hier sollten zunächst die Cluster unterschiedlicher Clusterlösungen paarweise verschmolzen werden.

Das geschilderte Vorgehen soll für den Vergleich der 7- und 8-Clusterlösung verdeutlicht werden: Die Clustermittelwerte der beiden Clusterlösungen werden abgespeichert und als neue Datenmatrix in eine hierarchische Clusteranalyse einbezogen. Als Distanzmaß werden quadrierte euklidische Distanzen spezifiziert und als Verfahren das gewichtete Mittelwertverfahren (Weighted-Average-Linkage) ausgewählt. In Tabelle 12.18 sind nur die Distanzen zwischen den beiden Lösungen wiedergegeben. Die Matrix erfüllt die

Tab. 12.18: Quadrierte euklidische Distanzen der 7- und 8-Clusterlösungen

Cluster (CL ₇)	Cluster (CL ₈)							
	C ₁ (NO)	C ₂ (GPMAT)	C ₃ (R ₁)	C ₄ (R ₂)	C ₅ (PMAT)	C ₆ (EPMAT)	C ₇ (MAT)	C ₈ (KON)
C ₁	0,8918	0,0031	4,5221	5,9727	0,3807	1,6769	0,2607	0,4400
C ₂	0,9806	0,3276	4,7684	8,5407	1,2701	3,3928	0,0395	0,0862
C ₃	1,2048	3,8237	0,0150	4,7735	3,1756	5,0071	5,0113	3,2819
C ₄	2,6714	1,5444	5,5309	1,9653	0,5449	0,0009	3,1266	3,4455
C ₅	5,2195	5,7591	5,0811	0,0000	3,3401	1,8881	8,6774	8,0959
C ₆	0,7926	0,3356	3,3874	3,3151	0,0042	0,6397	1,2666	1,2150
C ₇	0,0389	0,6539	1,8851	5,9853	0,9636	2,9529	0,9534	0,3119

grau hinterlegt: niedrigste Distanzen

Erwartungen: Es gibt $K - 2 = 8 - 2 = 6$ ähnliche Clusterpaare. C₁ der 8-Clusterlösung entspricht C₇ der 7-Clusterlösung, C₂ der 8-Clusterlösung C₁ der 7-Clusterlösung, C₃ der 8-Clusterlösung entspricht C₃ der 7-Clusterlösung usw. Die Cluster C₇ und C₈ der 8-Clusterlösung weisen zu C₂ der 7-Clusterlösung eine starke Ähnlichkeit auf. Auch das Dendrogramm (Abbildung 12.5) bildet diese Ähnlichkeitsbeziehungen ab und bringt die hohe Stabilität der Lösung zum Ausdruck. Zu einem sehr geringen Distanzniveau werden verschmolzen: Cluster C₃ der 7-Clusterlösung ($7C_3$) mit dem Restcluster der 8-Clusterlösung (8_{R₁}), Cluster C₄ der 7-Clusterlösung ($7C_4$) mit den extremen Postmaterialisten (8_{EPMAT}) usw.

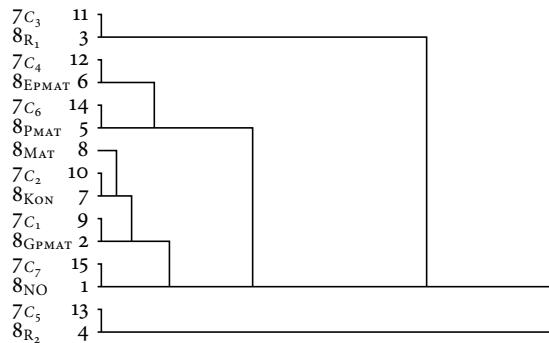


Abb. 12.5: Dendrogramm für die 7- und 8-Clusterlösungen

12.8 Inhaltliche Validitätsprüfung

Die *inhaltliche Validitätsprüfung* besteht darin, dass Hypothesen über die gefundenen Cluster aufgestellt werden. Da nach Inglehart (1979) Wertorientierungen auf Sozialisationsprozesse zurückzuführen sind, in denen Bedürfnisse zu stabilen Wertemustern ausgeformt werden, wird man in einer inhaltlichen Validitätsprüfung Einflüsse von Sozialisationsinstanzen annehmen. In der Untersuchung von Denz (1989) wurden zwei Sozialisationsinstanzen erfasst, nämlich Eltern und Schule. Bezuglich des Einflusses dieser Sozialisationsinstanzen lassen sich folgende Hypothesen für eine Validitätsprüfung formulieren:

- H₁: Postmaterialisten (Cluster 2, 5 und 6) kommen aus höheren sozialen Schichten als die anderen Wertetypen.
- H₂: Postmaterialisten (Cluster 2, 5 und 6) besuchen häufiger die AHS als die anderen Wertetypen, da hier wegen der humanistischen Tradition und der Funktion der Allgemeinbildung postmaterialistische Werte in einem stärkeren Ausmaß vermittelt werden.

Zusätzlich wurde folgender Zusammenhang zwischen der Wahrnehmung von Bedrohungen und den Wertorientierungen angenommen:

- H₃: Postmaterialisten (Cluster 2, 5 und 6) nehmen häufiger eine gesellschaftliche Bedrohung war.

Zur Validitätsprüfung kann die Clusterzugehörigkeit als neue Variable im Datensatz abgespeichert werden. Die Hypothesen können dann mit den gebräuchlichen bivariaten und multivariaten Verfahren geprüft werden. In unserem Beispiel lassen sich beispielsweise die in Tabelle 12.19 dargestellten Verfahren einsetzen.

Eine bivariate Prüfung von Hypothesen kann in ALMO auch innerhalb der Clusteranalyse vorgenommen werden, indem die Kriterienvariablen als *inaktive Deskriptionsvariablen* einbezogen werden. Die Ergebnisse fasst Tabelle 12.20 auf Seite 334 zusammen. Wiedergegeben sind die Clustermittelwerte bzw. Clusteranteilswerte und die entsprechenden z-Werte. Die Berechnung erfolgt analog zur Berechnung der zur Interpretation berechneten Clusterkennwerte (siehe Tabelle 12.8 auf Seite 316). Die Ausprägungen der nominalen Variablen ergeben je Spalte (Cluster) die Summe 1. Das Cluster C₁ (Nichtorientierte) setzt sich also zu 48 Prozent aus Schülerinnen und Schülern einer BHS zusammen, zu 17 Prozent aus Schülerinnen und Schülern einer AHS und zu 35 Prozent aus Schülerinnen und Schülern einer Berufsschule. Männlichen Geschlechts sind 36 Prozent, 64 Prozent weiblichen Geschlechts. Der Durchschnitt fühlt sich etwas von gesellschaftlichen Entwicklungen bedroht und kommt aus der Mittelschicht (mittleres Berufsprestige). Nur ein z-Wert hat einen Absolutbetrag größer 2.

Tab. 12.19: Übersicht über Validitätsprüfungsmöglichkeit für Clusterzugehörigkeit

	bivariat	multivariat
H ₁	Das Berufsprestige der Eltern wird als soziale Herkunftsvariable verwendet. Das Berufsprestige ist eine quantitative Variable. Zur bivariate Prüfung werden die Mittelwerte des Berufsprestige nach Cluster berechnet. Die Unterschiede können mittels Varianzanalyse oder mittels paarweiser t-Tests auf Signifikanz geprüft werden.	Diskriminanzanalyse bzw. äquivalent allgemeines lineares Modell oder multinominale logistische Regression. Abhängige Variable: Clusterzugehörigkeit; unabhängige Variable: Berufsprestige; Kontrollvariablen: Geschlecht, Schultyp
H ₂	Tabellenanalyse mit Schultyp und Clusterzugehörigkeit, da besuchter Schultyp nominalskaliert ist. Signifikanzprüfung mittels Tabellenanalyse möglich.	Diskriminanzanalyse bzw. äquivalent allgemeines lineares Modell oder multinominale logistische Regression. Abhängige Variable: Clusterzugehörigkeit; unabhängige Variable: Schultyp; Kontrollvariablen: Geschlecht, Schultyp
H ₃	Mittelwerte der Gefährdung nach Clusterzugehörigkeit, da Wahrnehmung von Gefährdung quantitatives Messniveau besitzt. Signifikanzprüfung mittels Varianzanalyse oder paarweiser t-Test möglich.	Wie H ₁ . Abhängige Variable: Wahrnehmung der Gefährdung; unabhängige Variable: Clusterzugehörigkeit; weitere unabhängige Variablen: Geschlecht, soziale Herkunft, Schultyp

Hypothese 1, dass Postmaterialisten aus höheren sozialen Schichten kommen, wird durch die Mittelwerte bestätigt. Das Berufsprestige der Eltern liegt über dem Gesamtdurchschnitt. Alle z-Werte für die postmaterialistischen Cluster (C_2 , C_5 und C_6) sind mit 0,17, 0,86 und 0,19 größer 0, während sie für alle anderen Cluster kleiner 0 sind. Ob die Unterschiede zu den anderen Clustern signifikant sind, kann zum Beispiel mittels paarweiser t-Tests oder mittels Kontrasten bei der Varianzanalyse geprüft werden. Die paarweisen t-Tests erbringen das Ergebnis, dass die Hypothese nicht haltbar ist. Von den neun möglichen Unterschieden (C_2 versus C_1 , C_2 versus C_7 , C_2 versus C_8 , C_3 versus C_1 usw.) ist nur ein Unterschied statistisch signifikant (siehe Tabelle 12.21 auf Seite 335).

Hypothese 2, dass Postmaterialisten häufiger aus einer AHS kommen, wird nur partiell bestätigt. Für die Cluster C_5 und C_6 werden positive z-Werte für die AHS berechnet. Schülerinnen und Schüler dieser Cluster kommen also überzufällig häufig aus einer AHS. Eine paarweise Signifikanzprüfung weist nur ein signifikantes Ergebnis aus. Hypothese 2 ist daher nicht haltbar. Allerdings haben Postmaterialisten eines gemeinsam: Sie kommen nicht aus einer Berufsschule, besuchen also eine AHS oder BHS. Alle z-Werte haben sehr hohe negative Zahlen (-2,14, -4,56, -3,35) und weichen signifikant vom Gesamtdurchschnitt ab. Auch beim paarweisen Test werden sechs signifikante Differenzen berechnet.

Tab. 12.20: Clustermittel- bzw. -anteilswerte und entsprechende z-Werte

	C_1 NO	C_2 GPMAT	C_3 R ₁	C_4 R ₂	C_5 PMAT	C_6 EPMAT	C_7 KON	C_8 MAT
<i>Schultyp</i>								
BHS	0,48	0,59	0,20	0,00	0,42	0,39	0,17	0,18
AHS	0,17	0,17	0,60	1,00	0,45	0,50	0,17	0,12
BS	0,35	0,24	0,20	0,00	0,13	0,11	0,67	0,71
<i>Geschlecht</i>								
männlich	0,36	0,60	0,80	0,80	0,53	0,72	0,37	0,38
weiblich	0,64	0,40	0,20	0,20	0,47	0,28	0,63	0,62
<i>Bedrohung</i>								
	2,13	2,58	2,47	2,73	2,76	3,20	2,44	2,22
<i>Berufsprestige Eltern</i>								
	3,57	3,86	3,80	5,00	4,06	3,89	3,79	3,59

(a) Clustermittel- bzw. -anteilswerte

	C_1 NO	C_2 GPMAT	C_3 R ₁	C_4 R ₂	C_5 PMAT	C_6 EPMAT	C_7 KON	C_8 MAT
<i>Schultyp</i>								
BHS	1,05	3,35	—	—	0,74	0,19	-3,43	-2,86
AHS	-1,15	-1,96	—	—	2,30	1,92	-1,72	-2,66
BS	-0,18	-2,14	—	—	-4,56	-3,35	4,08	4,28
<i>Geschlecht</i>								
männlich	-1,32	1,37	—	—	0,28	2,02	-1,79	-1,42
weiblich	1,32	-1,37	—	—	-0,28	-2,02	1,79	1,42
<i>Bedrohung</i>								
	-4,76	0,58	—	—	2,40	5,25	-0,99	-3,23
<i>Berufsprestige Eltern</i>								
	-0,90	0,17	—	—	0,86	0,19	-0,24	-1,25

grau hinterlegt: z-Werte größer bzw. kleiner $\pm 1,96$

(b) z-Werte

Die modifizierte Hypothese 2, dass Postmaterialisten seltener aus einer Berufsschule kommen, ist somit haltbar.

Hypothese 3 wird ebenfalls tendenziell bestätigt. Die z-Werte in der Variablen »Bedrohung« für alle postmaterialistischen Cluster sind größer 0, während sie für alle anderen Cluster kleiner 0 sind. Die paarweise Signifikanzprüfung ermittelt sechs signifikante Unterschiede. Die Hypothese kann daher aufrechterhalten werden.

Tab. 12.21: Paarweise t-Tests für die inhaltliche Validitätsprüfung

Hypothese		Ergebnis
H ₁ : Postmaterialistische Cluster rekrutieren sich häufiger aus höheren sozialen Schichten.	$C_2, C_5, C_6 > C_1, C_7, C_9$	1 von 9 statistisch nicht haltbar
H ₂ : Postmaterialistische Cluster rekrutieren sich häufiger aus der AHS.	$C_2, C_5, C_6 > C_1, C_7, C_9$	1 von 9 statistisch nicht haltbar
H ₂ modifiziert: Postmaterialistische Cluster rekrutieren sich seltener aus der BS.	$C_2, C_5, C_6 < C_1, C_7, C_9$	6 von 9 statistisch haltbar
H ₃ : In postmaterialistischen Clustern wird eine stärkere Bedrohung erlebt.	$C_2, C_5, C_6 > C_1, C_7, C_9$	6 von 9 statistisch haltbar

Anmerkung: Die Restcluster C_3 und C_4 mit jeweils fünf Fällen wurden nicht in die Signifikanzprüfung einbezogen. Für die Testung wurde eine Bonferroni-Korrektur vorgenommen (Bortz 2005, S. 129).

Zusammenfassend ergeben sich somit Hinweise, dass die Hypothesen H₂ (in modifizierter Form) und H₃ haltbar sind – Hypothese H₁ ist dagegen nicht haltbar. Die Schlussfolgerung, die aus diesem Befund gezogen wird, hängt davon ab, welches Gewicht den Hypothesen beigemessen wird. Wird H₁ als zentral betrachtet, wird man den Schluss ziehen, dass die inhaltliche Validität nicht gegeben ist. Wird ihr ein geringeres Gewicht zugesprochen, wird man die Validitätsprüfung als erfolgreich interpretieren.

12.9 Alternative Startwertverfahren

Wir haben bisher als Startwertverfahren die zufällige Zuordnung aller Objekte zu einem der K Cluster verwendet. Weitere Möglichkeiten, Startwerte für die Clusterzentren zu erhalten, sind:

- *Zufällige Auswahl mit systematischem Auswahlabstand*: Es wird für das erste Objekt eine Zufallszahl K_1 zwischen 1 und K erzeugt. Das erste Objekt wird dem Cluster K_1 zugeordnet, das zweite dem Cluster K_{1+1} usw. Überschreitet K_1 den Wert K , wird K_1 gleich 1 gesetzt.
- *Es werden Startwerte mit einem anderen Clusteranalyseprogramm berechnet*, zum Beispiel mit Hilfe des Ward-Verfahrens.
- *Es wird ein spezielles Startwerte-Programm verwendet*: Dies ist das Vorgehen von IBM-SPSS beim K-Means-Verfahren. In älteren Versionen war nur dieses Startwertverfahren implementiert.

- Der Benutzer gibt bestimmte inhaltlich begründete Startwerte ein: Das heißt, der Anwender definiert a priori die Clusterstruktur, die er aufgrund inhaltlicher Überlegungen erwarten würde.

Das von IBM-SPSS verwendete Startwertverfahren wird als *Quick-Clustering-Verfahren* bezeichnet. Es sucht bei K Clustern nach jenen K Objekten, die am weitesten voneinander entfernt sind (SPSS Inc. 2008, S. 659–660). Der Algorithmus ist relativ einfach, die Ergebnisse hängen allerdings von der Reihenfolge der Objekte ab. Es ist daher zu empfehlen, das Verfahren mehrmals mit einer unterschiedlichen Anordnung der Objekte durchzurechnen.

In unserem Beispiel ergibt sich bei zwei Versuchen (natürliche Reihenfolge und zufällige Anordnung) der in Tabelle 12.22 wiedergegebene Zusammenhang. Unmittelbar ersichtlich ist, dass die beiden Lösungen nicht perfekt übereinstimmen. Das Cluster C_1 der natürlichen Anordnung (Zeilenvariable) spaltet sich auf und verteilt sich zu gleichen Teilen auf das Cluster C_2 und C_6 . Cluster C_2 und C_3 bestehen aus einem Objekt. Die beiden Objekte werden in der zweiten Lösung (Spaltenvariable) einem Cluster zugeordnet. Cluster C_4 teilt sich wiederum auf zwei Cluster auf. Die Übereinstimmung (adjustierter RAND-Index) ist gleich 0,568 (siehe Formel 9.4 auf Seite 274). Erwarten würde man einen Wert nahe bei 1,00, da sich die beiden Analysen nur durch die Anordnung der Fälle unterscheiden und dies keinen Einfluss auf die Ergebnisse haben sollte. Der Wert liegt aber deutlich über dem Schwellenwert von 0,224 (siehe Abschnitt 9.5). Wird mit IBM-SPSS gerechnet, sollte man mehrere Lösungen berechnen und jene mit der höchsten erklärten Varianz bzw. der geringsten Fehlerstreuung auswählen.⁷

12.10 Gemischtes Messniveau

Das K-Means-Verfahren setzt strenggenommen quantitative Variablen voraus, da Mittelwerte und Varianzen berechnet werden. Durch entsprechende Datentransformationen und Gewichtungen kann es aber auch für Variablen mit *gemischemtem Messniveau* eingesetzt werden. Ein mögliches Vorgehen wurde bereits in Abschnitt 7.5 beschrieben und soll hier auf das K-Means-Verfahren übertragen werden.⁸ Das Vorgehen umfasst folgende Schritte:

⁷ Diese Fehlerstreuung kann in IBM-SPSS relativ einfach berechnet werden: Es wird die Distanz der Objekte zum Clusterzentrum als neue Variable abgespeichert. Die Variable wird quadriert und ausgezählt. Die Summe dieser Variablen ist gleich der Fehlerstreuung.

⁸ Ein spezielles K-Means-Verfahren für gemischtes Messniveau wird in Abschnitt 12.14 behandelt.

Tab. 12.22: Übereinstimmung zwischen zwei mit SPSS berechneten Clusterlösungen mit unterschiedlicher Anordnung der Objekte

natürliche Anordnung	zufällige Anordnung								gesamt
	1	2	3	4	5	6	7	8	
1	0	18	0	0	18	0	0	0	36
2	0	0	0	1	0	0	0	0	1
3	0	0	0	1	0	0	0	0	1
4	0	0	10	0	0	0	0	24	34
5	50	0	0	0	18	0	0	5	73
6	0	0	0	0	0	0	4	0	4
7	0	0	0	0	0	47	0	12	59
8	0	0	9	0	3	0	1	0	13
gesamt	50	18	19	2	39	47	5	41	221

Anmerkung: adjustierter RAND-Index = 0,568

grau hinterlegt: höchste Zellbesetzungen

- Die nominalen Variablen werden in ihre Dummies aufgelöst, alle Dummies werden benötigt.
- Ordinale Variablen werden, wie in der Forschungspraxis üblich, wie quantitative Variablen behandelt. Alternativ können sie auch als nominale Variablen in die Analyse eingehen.
- Standardisierung oder Gewichtung: Alle Variablen (Dummies, quantitative bzw. ordinale Variablen) werden standardisiert oder mit $1/s$ gewichtet. Alternativ kann die quadrierte Distanz in einer Variablen zum Clustermittelpunkt mit $1/s^2$ gewichtet werden. Alle drei Operationen führen zum gleichen Ergebnis.
- Die Dummies werden zusätzlich mit der Wurzel aus 0,5 gewichtet, um eine »Doppelzählung« der Unterschiede zu vermeiden. Alternativ können die Distanzen in den Dummies mit 0,5 gewichtet werden: Gewichtung der Dummies mit $\sqrt{0,5}$ = Gewichtung der Distanzen in den Dummies mit 0,5.

ALMO führt automatisch eine Auflösung in Dummies und eine Gewichtung der Dummies mit $\sqrt{0,5} = 0,707$ durch. Eine Gewichtung der Distanzen mit $1/s^2$ kann durch Verwendung der Mahalanobisdistanz erreicht werden (siehe Abschnitt 12.12). Wird mit IBM-SPSS gerechnet, empfehlen wir folgendes Vorgehen: Die nominalen Variablen werden in Dummies aufgelöst. Alle Variablen (Dummies, ordinale und quantitative Variablen) werden mittels »DESCRIPTIVES« standardisiert. Die Dummies werden anschließend mit 0,707 (= $\sqrt{0,5}$) multipliziert.

12.11 Modifikation des Algorithmus

Eine erste *Modifikation* des hier dargestellten *Grundalgorithmus der K-Means-Verfahren* geht auf MacQueen (1967) zurück. Die Modifikation besteht darin, dass nach jeder Neuzuordnung eines Objekts unmittelbar die Clusterzentren neu berechnet werden. Dadurch verkürzt sich der erforderliche Rechenaufwand, die Ergebnisse hängen aber von der Reihenfolge der Objekte ab. Diese Modifikation ist daher historisch zu verstehen, mit den heutigen leistungsfähigen Rechnern ist sie überholt.

Eine andere Modifikation des Algorithmus besteht darin, dass eine Neuzuordnung eines Objekts nur dann durchgeführt wird, wenn sich dadurch der Wert der Minimierungsfunktion ändert. Dieses Vorgehen wird als »*Hill-Climbing*«-Methode bezeichnet. Es eignet sich insbesondere dann, wenn anstelle des Varianzkriteriums andere Funktionen minimiert werden sollen. Friedman und Rubin (1967) haben ausgehend von der Zerlegung der Streuungsquadratmatrix \mathbf{SQ}_{ges} ⁹ in eine Streuungsquadratmatrix in den Clustern $\mathbf{SQ}_{\text{in}}(K)$ und zwischen den Clustern $\mathbf{SQ}_{\text{zw}}(K)$ eine Reihe von Kriterien vorgeschlagen:

1. *Minimierung der Spur von $\mathbf{SQ}_{\text{in}}(K)$:* Dieses Kriterium ist gleich dem Varianzkriterium, da in der Diagonalen die Streuungsquadratsummen in den einzelnen Variablen stehen.
2. *Minimierung der Determinante von $\mathbf{SQ}_{\text{in}}(K)$:* Die Determinante ist eine Möglichkeit der Varianzmessung bei m Variablen. Die verallgemeinerte Varianz – gemessen durch die Determinante – soll minimiert werden.
3. *Maximierung der Spur von $[\mathbf{SQ}_{\text{zw}}(K)]^{-1} \cdot \mathbf{SQ}_{\text{in}}(K)$:* Das Verhältnis der Streuung zwischen den Clustern zu der Streuung in den Clustern soll maximiert werden.

Die beiden zuletzt genannten Kriterien setzen voraus, dass die Matrix der Streuungsquadratsumme innerhalb der Cluster bzw. jene zwischen den Clustern vollen Rang besitzt. Werden alle Dummies einer nominalen Variablen einbezogen, ist diese Annahme verletzt. Die Verfahren können daher für nominale oder gemischte Variablen nicht eingesetzt werden. Ziel der dargestellten Modifikationen ist, bestimmte Nachteile des K-Means-Verfahrens zu beseitigen. Das *Determinantenkriterium* (Minimierung der Determinante von $\mathbf{SQ}_{\text{in}}(K)$) beispielsweise beseitigt unterschiedliche Varianzen und Korrelationen zwischen den Variablen. Dies kann auch durch die nachfolgend beschriebene Methode (Verwendung der Mahalanobis-Distanz) realisiert werden. Auch die Mahalanobis-Distanz kann für nominale Variablen nicht berechnet werden.

⁹ Der Fettdruck bedeutet, dass Matrizen untersucht werden. \mathbf{SQ}_{ges} ist also die Matrix der Streuungsquadratsummen der untersuchten Variablen.

12.12 Verwendung der Mahalanobis-Distanz

Anstelle der Definition einer anderen Minimierungsfunktion kann auch die bei der Zuordnung der Objekte verwendete Distanzfunktion geändert werden. Varianzen und Korrelationen zwischen den Variablen werden durch die Verwendung der sogenannten *Mahalanobis-Distanz* (Mahalanobis 1936) beseitigt. Diese ist für zwei Profile \mathbf{x}_g ¹⁰ und \mathbf{x}_{g^*} wie folgt definiert:

$$\text{MAHA}(\mathbf{x}_g, \mathbf{x}_{g^*}) = (\mathbf{x}_g - \mathbf{x}_{g^*}) \cdot \mathbf{W}^{-1} \cdot (\mathbf{x}_g - \mathbf{x}_{g^*})^T,$$

wobei \mathbf{x}_g der $(m \times 1)$ -Merkmalsvektor (Profil) des Objekts g mit den Ausprägungen von g in den untersuchten Variablen ist. Analog stellt \mathbf{x}_{g^*} den Merkmalsvektor des Objekts g^* dar, der Vektor \mathbf{x}_{g^*} kann auch das Mittelwertprofil eines Clusters sein. \mathbf{W} ist die Varianz-Kovarianzmatrix bzw. allgemein die Gewichtung der untersuchten Variablen. Für \mathbf{W} können unterschiedliche Spezifikationen vorgenommen werden:

1. \mathbf{W} ist eine Diagonalmatrix. In der Diagonalen stehen die Varianzen der Variablen:

$$\mathbf{W} = \begin{pmatrix} s_1^2 & 0 & \dots & 0 \\ 0 & s_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & s_p^2 \end{pmatrix}.$$

Diese Methode ist identisch mit einer Standardisierung der Variablen bzw. mit einer Gewichtung der Distanzen mit $1/s_j^2$ zum Erreichen der Vergleichbarkeit (siehe Kapitel 7).

2. \mathbf{W} ist gleich der Varianz-Kovarianzmatrix. In der Diagonalen stehen die Varianzen der Variablen, außerhalb der Diagonalen die Kovarianzen:

$$\mathbf{W} = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & & \vdots \\ \vdots & & \ddots & s_{(p-1)p} \\ s_{p1} & \dots & s_{p(p-1)} & s_p^2 \end{pmatrix}.$$

Diese Methode ist identisch mit einer automatischen Orthogonalisierung. Zusammenhänge und unterschiedliche Varianzen werden beseitigt. Dieses Vorgehen ist nicht unproblematisch (siehe Abschnitt 7.7). Sie sollte nur gewählt werden, wenn hohe Varianzen ein Hinweis auf Messfehler sind und hohe Korrelationen ein Hinweis auf zugrunde liegende gemeinsame Faktoren.

¹⁰ Der Fettdruck von Kleinbuchstaben zeigt Vektoren an.

3. Diese Methode ist identisch mit Spezifikation 1, anstelle der Varianzen werden aber die gepoolten Varianzen mit $w(K)_j^2 = \sum_k n_k \cdot s^2(j|k) / (n - K)$ verwendet. Dadurch soll gewährleistet werden, dass in die Standardisierung nur Unterschiede in den Clustern einfließen und beseitigt werden. Unterschiede zwischen den Clustern sollen erhalten bleiben. Die Matrix \mathbf{W} ist definiert als:

$$\mathbf{W} = \begin{pmatrix} w(K)_1^2 & 0 & \dots & 0 \\ 0 & w(K)_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & w(K)_p^2 \end{pmatrix}.$$

4. Diese Methode ist identisch mit Spezifikation 2, gerechnet wird aber mit gepoolten Varianzen *und* Kovarianzen:

$$\mathbf{W} = \begin{pmatrix} w(K)_1^2 & w(K)_{12} & \dots & w(K)_{1p} \\ w(K)_{21} & w(K)_2^2 & & \vdots \\ \vdots & & \ddots & w(K)_{(p-1)p} \\ w(K)_{p1} & \dots & w(K)_{p(p-1)} & w(K)_p^2 \end{pmatrix}.$$

Die Motivation ist analog zur dritten Methode. Zusätzlich soll durch Berücksichtigung der Kovarianzen eine Orthogonalisierung vorgenommen werden.

Der Modellansatz der Verwendung der gepoolten Varianz-Kovarianzmatrix (Methode 4) oder der gepoolten Varianzen (Methode 3) ist theoretisch attraktiv. Bei der Umsetzung ist es schwierig, gute Startwerte für die gepoolten Varianzen und Kovarianzen zu finden. Praktisch kann so vorgegangen werden, dass im ersten Iterationsschritt die Methoden 1 und 2 eingesetzt werden.

In unserem Beispiel ergeben sich dabei hinsichtlich der Clusterzentren nur geringfügige Änderungen. Als Beispiel wurde die 6-Clusterlösung ausgewählt (siehe Tabelle 12.23). Deutlichere Unterschiede treten bei den Clusteranteilswerten auf. Während beim gewöhnlichen K-Means-Verfahren 28,5 Prozent der Jugendlichen den gemäßigten Postmaterialisten (C_2) zugeordnet werden, sind dies bei den anderen beiden Verfahren nur 19,9 bzw. 17,2 Prozent. Umgekehrt erhöht sich der Anteil des Konsenstypus (C_5). Dies ist darauf zurückzuführen, dass – wie wir aus der Stabilitätsprüfung wissen – diese beiden Cluster nicht besonders gut getrennt sind. Mitunter wird man sich daher für die Anwendung eines probabilistischen Verfahrens entscheiden, um möglichen Überlappungen Rechnung zu tragen.

Zur Beurteilung der Modellanpassung kann eine *modifizierte »erklärte Streuung«* $\tilde{\eta}^2$ berechnet werden:

$$\tilde{\eta}_K^2 = 1 - \frac{\sum_k \sum_{g \in k} \text{MAHA}(g,k)}{\sum_g \text{MAHA}(g,o)},$$

Tab. 12.23: Ergebnisse des K-Means-Verfahrens bei Gewichtung der Distanzen

	Cluster					
	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆
<i>K-Means-Verfahren ohne Gewichtung der Distanzen</i>						
Anteile in Prozent	7,2	28,5	21,3	7,7	16,3	19,0
gMat (\bar{x})	2,25	1,92	2,60	3,64	1,32	1,63
gPmat (\bar{x})	2,76	1,38	1,48	1,37	1,36	1,92
<i>K-Means-Verfahren mit gepoolten Varianzen</i>						
Anteile in Prozent	8,1	19,9	23,1	10,0	21,7	17,2
gMat (\bar{x})	2,31	1,94	2,42	3,45	1,39	1,69
gPmat (\bar{x})	2,72	1,25	1,55	1,29	1,42	1,93
<i>K-Means-Verfahren mit gepoolter Varianz-Kovarianzmatrix</i>						
Anteile in Prozent	5,8	17,2	23,1	13,1	21,7	19,0
gMat (\bar{x})	2,42	1,88	2,36	3,32	1,39	1,69
gPmat (\bar{x})	2,78	1,27	1,53	1,35	1,42	2,00

Abkürzungen: C₁: Nicht-Orientierte, C₂: gemäßigte Postmaterialisten, C₃: Postmaterialisten, C₄: extreme Postmaterialisten, C₅: Konsenstypus, C₆: gemäßigte Materialisten; gMat: Gesamtpunktwerte Materialismus, gPmat: Gesamtpunktwerte Postmaterialismus

wobei MAHA(g, k) die Mahalanobis-Distanz des Objekts g zu seinem Clusterzentrum k ist. MAHA(g, o) ist die Mahalanobis-Distanz des Objekts g zu den Mittelwerten der 1-Clusterlösung. Für unser Beispiel ergibt sich bei einer Gewichtung mit den gepoolten Varianzen eine Modellanpassung von 0,754 und bei einer Gewichtung mit der gepoolten Kovarianzmatrix ein Wert von 0,818. Im Vergleich zum K-Means-Verfahren ($\eta^2 = 0,794$) liegt bei Verwendung der gepoolten Varianz-Kovarianzmatrix eine etwas bessere Modellanpassung vor.

12.13 Konfirmatorisches K-Means-Verfahren

Bei der *konfirmatorischen Clusteranalyse* können analog zum Vorgehen bei den konfirmatorischen Strukturgleichungsmodellen (Bentler 1985; Bollen 1989; Hayduk 1987; Jöreskog und Sörbom 1984; Reinecke 2005) bestimmte Clusterzentren gleich bestimmten Werten gesetzt, also fixiert werden. Diese Werte werden in der Iteration nicht geändert. Daneben können lineare Restriktionen hinsichtlich der Mittelwerte spezifiziert werden, dass beispielsweise der Mittelwert der Variablen j im Cluster k gleich dem Mittelwert der Variablen j^* im Cluster k^* sein soll.

Tab. 12.24: Clusterzentren von fünf angenommenen Wertetypen

	Cluster				
	C ₁	C ₂	C ₃	C ₄	C ₅
Materialismus (gMat)	1,50	1,50	2,50	3,00	3,00
Postmaterialismus (gPmat)	1,50	2,50	1,50	1,00	3,00

Abkürzungen: C₁: Konsensotypus, C₂: gemäßigte Materialisten, C₃: gemäßigte Postmaterialisten, C₄: Postmaterialisten, C₅: Nicht-Orientierte
 grau hinterlegt: kleinste Clustermittelwerte

Das Vorgehen soll für die Wertedaten von Denz (1989) verdeutlicht werden. Die angenommenen Clusterzentren sind in der Tabelle 12.24 dargestellt. Es sollen drei konfirmatorische Analysen durchgeführt werden:

- Alle Clusterzentren werden fixiert.
- Die kleinsten Mittelwerte in jedem Cluster werden fixiert.
- Es werden lineare Restriktionen zwischen den Clustern spezifiziert.

In der ersten konfirmatorischen Analyse sind alle vorgegebenen Clusterzentren fixiert. Das heißt, es wird nur geprüft, wie gut die vorgegebene Clusterstruktur den Daten angepasst ist. Eine Neuberechnung von Clusterzentren findet nicht statt. Führt man die Analyse durch, ergibt sich eine erklärte Varianz von 66,0 Prozent (siehe Tabelle 12.25). Die vorgegebene Clusterstruktur erklärt somit 66,0 Prozent der Gesamtvarianz der Daten. Diese Erklärungskraft ist überzufällig. Bei homogenen Daten (normalverteilte Zufallsdaten) würde sich im Durchschnitt eine erklärte Varianz von 0,583 bzw. von 58,3 Prozent ergeben. Im Vergleich zu einer Analyse, bei der alle Clusterzentren frei variieren können, reduziert sich die erklärte Streuung von 75,2 Prozent auf 66,0 Prozent.

In der zweiten konfirmatorischen Analyse wird untersucht, ob durch eine Abschwächung der Modellannahmen eine deutlich höhere erklärte Streuung erzielt werden kann. Dazu wird der kleinste Mittelwert bzw. werden die kleinsten Mittelwerte der Tabelle 12.24 in jedem Cluster fixiert. Die anderen Mittelwerte werden zur Schätzung freigegeben. Im Cluster 1 werden beide Mittelwerte fixiert, da sie mit 1,50 gleich sind. Im Cluster 2 wird der Mittelwert der materialistischen Wertorientierung fixiert, da er mit einem Wert von 1,50 kleiner als jener der postmaterialistischen Wertorientierung mit 2,50 ist. In den Clustern 3 und 4 wurde der Mittelwert für die postmaterialistische Wertorientierung fixiert und im Cluster 5 wiederum beide Mittelwerte. Unter diesen Modellannahmen ergibt sich eine erklärte Streuung von 71,0 Prozent (siehe Tabelle 12.25). Gegenüber der vorausgehenden Analyse bedeutet dies eine prozentuelle Verbesserung um 7,6 Prozent (= 100 · (0,710 – 0,660)/0,660).

Tab. 12.25: Ergebnisse von konfirmatorischen Analysen mit unterschiedlichen Modellannahmen

Kurzbeschreibung der Modellannahme	erklärte Streuung η_5^2	Erwartungswert der erkl. Streuung für homog. Nullmodell $E(\eta_5^2)$	Standardabw. der erkl. Streuung für homog. Nullmodell $\sigma(\eta_5^2)$
alle Clusterzentren fixiert	0,660	0,583	0,021
kleinste(r) Mittelwert(e) in jedem Cluster fixiert	0,710	0,585	0,020
lineare Restriktionen zwischen den Mittelwerten	0,708	0,598	0,021
alle Mittelwerte frei variierbar	0,752	0,603	0,016

Tabelle 12.25 enthält drei lineare Restriktionen:

- a) Die Mittelwerte in der materialistischen und postmaterialistischen Wertorientierung im Cluster C_1 sollen untereinander gleich und gleich dem Mittelwert in der materialistischen Wertorientierung des Clusters C_2 sowie gleich dem Mittelwert in der postmaterialistischen Wertorientierung des Clusters C_3 sein. Inhaltlich ausgedrückt besagt diese Annahme, dass für den Konsenstypus beide Wertorientierungen gleich wichtig sein sollen. Die materialistischen Werte von C_1 sollen ferner gleich wichtig wie im Cluster C_2 der gemäßigten Materialisten sein, die postmaterialistischen Werte von C_1 sollen gleich wichtig wie im Cluster C_3 der gemäßigten Postmaterialisten sein. Wenn wir die Variablen mit 1 (Materialismus) und 2 (Postmaterialismus) nummerieren, lässt sich diese lineare Restriktion formulieren mit: $\bar{x}_{11} = \bar{x}_{12} = \bar{x}_{21} = \bar{x}_{32}$, wobei der erste Index für das Cluster steht.
- b) Die Ablehnung der postmaterialistischen Werte im Cluster C_2 der gemäßigten Materialisten soll gleich der Ablehnung der materialistischen Werte im Cluster C_3 der gemäßigten Postmaterialisten sein. Formal ausgedrückt soll gelten: $\bar{x}_{22} = \bar{x}_{31}$.
- c) Die Ablehnung der materialistischen Werte im Cluster C_4 der Postmaterialisten soll gleich der Ablehnung der materialistischen und postmaterialistischen Werte im Cluster C_5 der Nicht-Orientierten sein. Formal ausgedrückt: $\bar{x}_{41} = \bar{x}_{51} = \bar{x}_{52}$.

Stellt man diese drei linearen Restriktionen an die zu berechnende Clusterlösung, wird eine erklärte Streuung von 70,8 Prozent erzielt. Die berechneten Clusterzentren sind in der Tabelle 12.26 auf der nächsten Seite wiedergegeben.

Eine konfirmatorische Analyse führt immer zu einer schlechteren Modellanpassung als eine Analyse mit frei variierbaren Parametern. Sie hat aber umgekehrt den Vorteil der leichteren Interpretierbarkeit. Es gilt also abzuwägen, ob der Vorteil der leichteren

Tab. 12.26: Ergebnisse der konfirmatorischen Analyse bei linearen Restriktionen

	Cluster				
	C ₁	C ₂	C ₃	C ₄	C ₅
<i>Startwerte für Clusterzentren</i>					
Materialismus (gMat)	1,50 ^{a)}	1,50 ^{a)}	2,50 ^{a)}	3,00 ^{c)}	3,00 ^{c)}
Postmaterialismus (gPmat)	1,50 ^{a)}	2,50 ^{b)}	1,50 ^{a)}	1,00	3,00 ^{c)}
<i>Clusterzentren nach Iteration</i>					
Materialismus (gMat)	1,52	1,52	2,28	3,16	3,16
Postmaterialismus (gPmat)	1,52	2,28	1,52	1,27	3,16

a), b), c): lineare Restriktionen (siehe Text)

Abkürzungen: C₁: Konsenstypus, C₂: gemäßigte Materialisten, C₃: gemäßigte Postmaterialisten, C₄: Postmaterialisten, C₅: Nicht-Orientierte

Interpretierbarkeit die Abnahme in der erklärten Streuung kompensiert. In unserem Fall ist dies für die erste untersuchte Modellkonstellation, bei der alle Werte fixiert sind, sicherlich nicht der Fall. Die Abnahme ist hier zu groß. Die beiden anderen Modellkonstellationen liegen in der Mitte zwischen dem ersten Modell und dem Modell mit frei variierbaren Mittelwerten. Man wird sich daher – abhängig von inhaltlichen Überlegungen und der leichteren Interpretierbarkeit – für eines der beiden konfirmatorischen Modelle entscheiden. Vorstellbar ist, dass analog zu den Strukturgleichungsmodellen weitere Modellprüfgrößen entwickelt werden.

Technisch ist beim konfirmatorischen K-Means-Verfahren eine Änderung des Algorithmus bei der Neuberechnung der Clusterzentren erforderlich (siehe Abschnitt 12.1). Die Clusterzentren werden wie folgt neu berechnet:

- a) Es werden zunächst alle Clusterzentren entsprechend der Gleichung 12.1 auf Seite 300 berechnet mit

$$\bar{x}_{kj} = \frac{\sum_{g \in k} x_{gj}}{n_{kj}} .$$

- b) Die fixierten Mittelwerte werden gleich den vorgegeben Startwerten gesetzt:

$$\bar{x}_{kj}^{\text{fix}} = \bar{x}_{kj}^{(0)} ,$$

wobei der hochgestellte Index in Klammern auf der rechten Seite für den Startwert steht.

- c) Für jede Gruppe der Mittelwerte, die eine lineare Restriktion erfüllen, wird eine Kleinstes-Quadrat-Schätzung durchgeführt. Bezeichnen wir die Elemente einer Gruppe mit k^* und j^* , so ergibt sich folgender Schätzwert:

$$\bar{x}_{k^*j^*} = \frac{\sum_{k,j \in \{k^*, j^*\}} n_{kj} \cdot \tilde{x}_{kj}}{\sum_{k,j \in \{k^*, j^*\}} n_{kj}}.$$

Das Clusterzentrum für die Gruppe der Elemente, zwischen denen eine lineare Restriktion definiert ist, ist gleich dem gewichteten Gesamtmittelwert. Da ein Mittelwert bekanntlich die Fehlerquadratsumme $\sum (x_i - \bar{x})^2$ minimiert, ist der gewichtete Mittelwert eine Kleinstes-Quadrat-Schätzung.

In ALMO können auch die Clustergrößen fixiert oder durch lineare Restriktionen verknüpft werden. Auf die diesbezüglichen Berechnungsformeln soll hier nicht eingegangen werden. Nicht enthalten sind in ALMO nichtlineare Restriktionen und andere Restriktionen, wie sie zum Beispiel in den Arbeiten von DeSarbo und Mahajan (1984) sowie Ferligoj und Batagelj (1982, 1983) beschrieben werden.

12.14 K-Median- und K-Modus-Verfahren

Analog zum K-Means-Verfahren wurden in den letzten Jahren das *K-Median*- und das *K-Modus-Verfahren* entwickelt. Dadurch soll eine bessere Interpretation der Clusterzentren erreicht werden. Beim K-Means, aber auch den hierarchischen Verfahren zur Konstruktion eines Mittelwertes, werden oft Dezimalzahlen als Mittelwerte berechnet, obwohl nur eine ganzzahlige Skala (beispielsweise von 1 »leicht« bis 5 »schwierig«) zur Verfügung steht. Daher ist oft schwer zu entscheiden, wie ein Wert von 1,27, 1,47 oder 1,52 zu interpretieren ist.

Beim *K-Median-Verfahren* (»k-Median Clustering«, Chen 2006) wird an Stelle des Mittelwerts der Median verwendet. Das Verfahren eignet sich also insbesondere für ordinale Variablen. Anstelle der quadrierten euklidischen Distanz wird die City-Block-Metrik verwendet. Die Minimierungsfunktion lautet:

$$D_{in}(K) = \sum_k \sum_{g \in k} d_{g,k} \rightarrow \min$$

mit $d_{g,k} = \sum_j |x_{gj} - \tilde{x}_{kj}|$. Der Median des Clusters k in der Variablen j wird als \tilde{x}_{kj} bezeichnet.

Beim *K-Modus-Verfahren* (»k-Modes Clustering«; Chaturvedi, Green u. a. 2001) wird als Lagemaß der Modus verwendet. Das Verfahren eignet sich somit für nominale Variablen. Die Minimierungsfunktion ist analog zu oben definiert, unterscheidet sich aber in der Berechnung der Distanzen. Diese sind wie folgt definiert:

$$d_{g,k} = \sum_j d_{gkj}$$

mit

$$d_{gkj} = \begin{cases} 0 & \text{wenn } x_{gj} = x_{kj}^{\text{modus}} \\ 1 & \text{wenn } x_{gj} \neq x_{kj}^{\text{modus}} \end{cases}.$$

Der Modus des Clusters k in der Variablen j wird mit x_{kj}^{modus} bezeichnet. Durch die Verwendung des Modus wird gewährleistet, dass jedes Cluster durch eine konkrete Ausprägung in einer nominalen Variablen gekennzeichnet ist.

Unter Berücksichtigung des unterschiedlichen Messniveaus lässt sich eine allgemeine Distanzfunktion $D_{\text{in}}(K)$ für ein *allgemeines K-Cluster-Verfahren* definieren als:

$$D_{\text{in}}(K) = \sum_k \sum_{g \in k} d_{g,k} = \sum_k \sum_{g \in k} \sum_j w_j d_{gkj} \rightarrow \min ,$$

mit

$$d_{gkj} = |x_{gj} - \bar{x}_{kj}| \quad \text{als Distanzfunktion für quantitative Variablen,}$$

$$d_{gkj} = |x_{gj} - \tilde{x}_{kj}| \quad \text{als Distanzfunktion für ordinale Variablen und}$$

$$d_{gkj} = \begin{cases} 0 & \text{wenn } x_{gj} = x_{kj}^{\text{modus}} \\ 1 & \text{wenn } x_{gj} \neq x_{kj}^{\text{modus}} \end{cases} \quad \text{als Distanzfunktion für nominale Variablen.}$$

Um Vergleichbarkeit zu erreichen, müssen die Distanzen in jeder Variablen gewichtet werden. Andernfalls würden die Variablen mit dem größten Skalenbereich (zum Beispiel Einkommen in Euro) die Ergebnisse sehr stark beeinflussen. Als Gewichtung kann die Spannweite eingesetzt werden:

$$w_j = \frac{1}{\max_{g,g^*}(d_{gg^*j})} .$$

Die maximale Distanz in der Variablen j ist $\max_{g,g^*}(d_{gg^*j})$. Bei nominalen Variablen ist sie gleich 1. Bei ordinalen und quantitativen Variablen entspricht die maximale Distanz der Spannweite $x_j^{\text{max}} - x_j^{\text{min}}$, wobei für das Minimum und Maximum theoretische oder

empirische Werte eingesetzt werden können. Bei Nutzung der empirischen Werte ergibt sich das Distanzmaß von Gower (1971) für gemischte Merkmale (Gordon 1981, S. 23):

$$d_{g,k} = \sum_j d_{gkj}^{\text{nom}} + \underbrace{\sum_j \frac{1}{x_j^{\max} - x_j^{\min}} \cdot |x_{gj} - \bar{x}_{kj}|}_{\text{ordinale Variable}} + \underbrace{\sum_j \frac{1}{x_j^{\max} - x_j^{\min}} \cdot |x_{gj} - \bar{x}_{kj}|}_{\text{quantitative Variable}},$$

mit

$$d_{gkj}^{\text{nom}} = \begin{cases} 0 & \text{wenn } x_{gj} = x_{kj}^{\text{modus}} \\ 1 & \text{wenn } x_{gj} \neq x_{kj}^{\text{modus}} \end{cases}.$$

Das vorgeschlagene K-Cluster-Verfahren minimiert das Distanzmaß von Gower. Bei diesem kommt den nominalen Variablen ein relativ starkes Gewicht zu, da der Unterschied von zwei Objekten in einer nominalen Variablen gleich dem maximalen Unterschied in einer ordinalen oder quantitativen Variablen ist. Der Unterschied zwischen dem größten und kleinsten Einkommen beispielsweise ist gleich dem Unterschied zwischen einem männlichen und einer weiblichen Befragten. Ist dies nicht erwünscht, müssen andere Gewichte verwendet werden.

Auf eine begriffliche Verwirrung sei hingewiesen: Das K-Median-Verfahren wird oft auch als *K-Medoids-Verfahren* bezeichnet und umgekehrt das K-Medoids-Verfahren als K-Median-Verfahren (Kaufman und Rousseeuw 1990, S. 72). Das von Kaufman und Rousseeuw (1990) entwickelten K-Medoids-Verfahren ist ein Repräsentantenverfahren. Jedes Cluster wird durch ein »typisches« Objekt gekennzeichnet. Beim hier behandelten K-Median-Verfahren wird der Median als Lagemaß über alle Objekte eines Clusters berechnet. Das K-Median-Verfahren enthält beispielsweise CLUSTER 3.0 (Eisen und Hoon 2002), eine Open-Source-Software. Ein Fortran-Programm für das K-Modus-Verfahren wird von Chaturvedi, Green u. a. (2001, S. 53) angeboten.

12.15 Anwendungsempfehlungen

1. Wir empfehlen für eine objektorientierte Clusteranalyse die Anwendung des K-Means-Verfahrens bei ausreichender Fallzahl. Ein genauer Schwellenwert für »ausreichend« lässt sich nicht benennen. Die erforderliche Fallzahl hängt von der Zahl der Variablen, deren Messgenauigkeit und der zugrunde liegenden Clusterstruktur ab. Bei kleinen Datensätzen kann für eine objektorientierte Clusteranalyse das Ward-Verfahren angewendet werden (siehe Abschnitt 11.1), wenn sich das K-Means-Verfahren beispielsweise als instabil erweist.

2. Werden Überlappungen vermutet, sollte bei großer Fallzahl ein probabilistisches Verfahren gewählt werden (siehe Teil III), bei kleinerer Fallzahl kann mit dem Complete-Linkage für überlappende Cluster oder mit einem Repräsentantenverfahren gerechnet werden.
3. Zur Vermeidung von lokalen Minima sollte die Methode der multiplen zufälligen Startwerte eingesetzt werden (siehe Abschnitt 12.2).
4. Zur Bestimmung der Clusterzahl hat sich der Abfall des PRE-Koeffizienten bewährt. Ergänzend sollten die weiteren dargestellten Maßzahlen eingesetzt und mehrere formal geeignete Clusterlösungen weiter untersucht werden (siehe Abschnitt 12.2).
5. Bei mehreren formal zulässigen Clusterlösungen sollte eine formale Gültigkeitsprüfung durchgeführt werden. Dadurch ist es eventuell möglich, eine formal beste Lösung auszuwählen (siehe Kapitel 20).
6. Entscheidend ist aber die inhaltliche Interpretierbarkeit und die inhaltliche Gültigkeit. Inhaltliche Interpretierbarkeit liegt vor, wenn den Clustern ein sinnvoller Name gegeben werden kann. Inhaltliche Gültigkeit ist gegeben, wenn Hypothesen über das gemeinsame Auftreten der Cluster mit nicht in die Analyse einbezogenen Kriterienvariablen vorläufig bestätigt werden können.
7. Bei der inhaltlichen Interpretation der Clusterzentren ist auf die Eigenschaften der Variablen zu achten. Werden beispielsweise standardisierte Variablen verwendet, sind absolute Vergleichsurteile zwischen Variablen nicht mehr möglich (siehe Abschnitt 12.4).
8. Eine inhaltliche Gültigkeitsprüfung sollte in jedem Fall durchgeführt werden.

Teil III

Probabilistische Clusteranalyseverfahren

13 Einleitende Übersicht

Die *probabilistischen Clusteranalyseverfahren* unterscheiden sich von den im vorausgehenden Kapitel behandelten deterministischen Verfahren dadurch, dass ein Objekt g jedem Cluster k mit einer bestimmten Wahrscheinlichkeit $\pi(k|g)$ angehört. Wir werden diese Wahrscheinlichkeit im Folgenden als *Zuordnungswahrscheinlichkeit* bezeichnen. In diesem Kapitel werden folgende probabilistische Verfahren und Ansätze behandelt:

- Analyse latenter Klassen für quantitative Variablen bzw. latente Profilanalyse (Kapitel 14)
- Analyse latenter Klassen für nominale, ordinale und gemischtskalierte Variablen (Kapitel 15)
- Latent-GOLD-Ansatz (Kapitel 16)
- Weitere Verfahren (Kapitel 17)

Die ersten zwei Abschnitte dienen der Verfahrensverdeutlichung. Für die *Analyse latenter Klassen* (»Latent Class Analysis«, LCA) und damit für die probabilistische Clusteranalyse ist der Einsatz von *Latent GOLD* zu empfehlen. Es ist ein umfassendes Programm Paket, das neben der Analyse latenter Klassen auch weitere Verfahren enthält. Dargestellt werden des Weiteren die SPSS-Prozedur *TwoStep-Cluster* und der Ansatz des Programms *AutoClass*.

Bis auf die Ausnahme TwoStep-Cluster lassen sich alle Verfahren als Verallgemeinerung des *K-Means-Verfahrens* (siehe Kapitel 12) entwickeln. Sie eignen sich daher nur für eine objektorientierte Clusteranalyse. Technisch bestehen die Modifikationen im Folgenden:

1. Der Schritt 3 des Algorithmus der K-Means-Verfahren (siehe Seite 300), in dem jedes Objekt g dem Cluster zugeordnet wird, zu dem die quadrierte euklidische Distanz minimal ist, wird dahingehend geändert, dass die *Zuordnungswahrscheinlichkeiten* $\pi(k|g)$ berechnet werden. Dazu sind in Abhängigkeit vom Messniveau bestimmte Verteilungsannahmen erforderlich. Ferner geht in die Berechnung die Annahme der – unten näher erläuterten – *lokalen Unabhängigkeit* ein.

2. Die Klassenzentren \hat{x}_{gj} und Klassenanteilswerte $\pi(k)$ (Schritt 2 des Algorithmus des K-Means-Verfahrens) werden als *Maximum-Likelihood-* oder *Bayes-Schätzer* berechnet.

Mit Ausnahme dieser beiden Modifikationen erfolgt die Berechnung nach dem Algorithmus der K-Means-Verfahren (siehe Abschnitt 12.1). Der Algorithmus für probabilistische Clusteranalyseverfahren wird in der Literatur als *EM-Algorithmus* (»Expectation Maximization«) bezeichnet (Dempster, Laird u. a. 1977) und geht auf Goodman (1974) zurück. Der EM-Algorithmus hat sich seit seiner Einführung in zahlreichen Anwendungssituationen bewährt (siehe zum Beispiel Bock und Aitkin 1981; De Soete und DeSarbo 1991; Heckman und Singer 1984; Langeheine und Van de Pol 1990; Rigdon und Tsutakawa 1983; Van de Pol und de Leeuw 1986). Das Konvergenzverhalten zur Lösung einer Schätzaufgabe wurde von Dempster, Laird u. a. (1977) sowie C. F. J. Wu (1983) untersucht: Konvergenz zumindest gegen ein lokales Minimum ist im Regelfall gegeben.

Das Konzept der *lokalen Unabhängigkeit* ist für die in diesem Kapitel behandelten Verfahren zentral. Es ist aus der *Analyse latenter Strukturen* mit der Analyse latenter Klassen als Submodell von Lazarsfeld und Henry (1968) bekannt und geht von folgender Modellvorstellung aus:

1. Den Daten liegen K unbekannte (nicht beobachtete) Klassen zugrunde. Diese werden als *latente Klassen* bezeichnet und erklären
2. die Zusammenhänge zwischen den untersuchten beobachteten (manifesten) Variablen. Werden die (latenten) Klassen als Kontrollvariablen in die Analyse eingeführt, verschwinden die empirischen Zusammenhänge. Die manifesten Variablen sind innerhalb jeder Klasse unabhängig.

Wegen der Beziehung zur Analyse latenter Klassen von Lazarsfeld und Henry (1968) wird für die in diesem Kapitel behandelten Verfahren die Bezeichnung *Analyse latenter Klassen* gewählt. Anstelle von »Clustern« wird von »latenten Klassen« oder kurz von »Klassen« gesprochen, obwohl man sich selbstverständlich unter den Klassen auch Cluster vorstellen kann. Die behandelten Verfahren lassen sich vorstellen als:

1. *Verallgemeinerung des K-Means-Verfahrens*: Die Annahme einer deterministischen Zuordnung der Objekte zu den Klassen wird fallengelassen.
2. *Verallgemeinerung der klassischen Analyse latenter Klassen* von Lazarsfeld und Henry (1968): Neben dichotomen Variablen können nominalskalierte Variablen mit beliebig vielen Ausprägungen, ordinalskalierte und/oder quantitative Variablen untersucht werden.
3. *Submodelle von Mischverteilungsverfahren*: Mischverteilungsverfahren (Kaufmann und Pape 1984, S. 420–445; Wolfe 1970; u. a.) gehen von der Vorstellung aus, dass die

empirische Verteilung der Objekte in den untersuchten Variablen eine Mischung von Wahrscheinlichkeitsverteilungen, zum Beispiel von K m -dimensionalen Normalverteilungen mit den Mittelwertsvektoren μ_k und den Kovarianzmatrizen K_k mit den Elementen $\sigma(k)_{jj^*}^2$ (Kovarianz für $j \neq j^*$ bzw. Varianz für $j = j^*$ zwischen den Variablen j und j^* in der k -ten Normalverteilung) ist. Die Aufgabe von Mischverteilungsverfahren ist die Schätzung des Mischungsverhältnisses und der Parameter der einzelnen Wahrscheinlichkeitsverteilungen. Werden bestimmte zusätzliche Annahmen getroffen, können die Mischverteilungen durch die hier behandelten Verfahren geschätzt werden. So zum Beispiel geht die latente Profilanalyse von der Modellvorstellung aus, dass K m -dimensionale Normalverteilungen vorliegen, wobei die Variablen innerhalb der (latenten) Klassen unabhängig sind.¹

4. Schließlich können die probabilistischen Clusteranalyseverfahren als *Clusteranalyseverfahren interpretiert werden, die eine Modellierung zufälliger Messfehler erlauben* (Espeland und Handelman 1989; Van de Pol und de Leeuw 1986; u. a.).

Die genannten Punkte haben sich in der Praxis im Vergleich mit anderen Klassifizierungsansätzen als vorteilhaft erwiesen. Die Vorteile der probabilistischen Verfahren sind:

- Das durch unterschiedliche Skaleneinheiten und Messniveaus bedingte *Problem der Nichtvergleichbarkeit* tritt nicht auf, da zur Berechnung der Zuordnungswahrscheinlichkeiten $\pi(k|g)$ mit im Intervall $[0,1]$ normierten Wahrscheinlichkeiten $\pi(x_{gj}|k)$ gerechnet wird. Bei den deterministischen Verfahren muss dagegen eine Gewichtung der Variablen oder der Distanzen vorgenommen werden (siehe Abschnitt 7.1).
- Es können *Messfehler* in den Variablen modelliert werden. Messfehler sind clusteranalytisch betrachtet irrelevante Variablen, da sie keine Bedeutung für die Trennung von Clustern haben. Bei deterministischen Verfahren, insbesondere bei hierarchischen, können sie die Clusterstruktur zerstören (siehe Abschnitt 6.6).
- Die probabilistischen Verfahren sind *weniger anfällig für Verzerrungen* durch irrelevante Variablen (Bacher und Vermunt 2010).
- Es werden *erwartungstreue Schätzer für die Clusterzentren ermittelt*, während die deterministischen Verfahren eine »optimale« Partition ermitteln, die bei überlappenden Clusterstrukturen in verzerrten Schätzern der Clusterzentren resultiert (siehe Abschnitt 6.6).
- Zur Bestimmung der Clusterzahl stehen *formal besser begründete Maßzahlen zur Verfügung* (siehe Abschnitt 6.6). Allerdings liefern auch diese häufig keine eindeutige Entscheidungsgrundlage.
- Es können *unterschiedliche Variablentypen modelliert werden* (siehe Abschnitt 16).

¹ Bei den Mischverteilungsmodellen ist die Annahme der lokalen Unabhängigkeit in den latenten Klassen nicht erforderlich.

Umgekehrt haben natürlich die probabilistischen Clusteranalyseverfahren auch bestimmte Nachteile:

- Für eine konvergente und stabile Lösung – insbesondere für die latente Profilanalyse – werden häufig größere Stichproben benötigt als für die K-Means-Verfahren. Das Konvergenzverhalten bei der latenten Profilanalyse hängt von dem *Überlappungsanteil* ab (Kaufmann und Pape 1984, S. 433).² Je geringer der Überlappungsanteil ist, desto besser konvergiert das Verfahren asymptotisch.
- Es werden bestimmte Annahmen getroffen. Sind diese nicht erfüllt, kann dies zu verzerrten Schätzungen führen. So zum Beispiel kann eine Verletzung der Annahme der lokalen Unabhängigkeit eine Überschätzung der Klassenzahl bedingen.

Ein weiterer »Nachteil« der Verfahren kann darin gesehen werden, dass strenggenommen vor der Analyse die *Identifikation des zu schätzenden Modells* untersucht werden muss. Ein zu schätzendes Modell M wird dann als identifiziert bezeichnet, wenn die zu schätzenden Modellparameter P eindeutig bestimmt werden können. Das heißt, es darf kein anderes Modell M^* mit anderen Modellparametern P^* geben, das dieselben Modelldaten produziert. Eine notwendige (aber nicht hinreichende) Bedingung für die Eindeutigkeit (Identifikation) eines Modells ist, dass die Zahl der empirischen Informationen größer oder gleich der Zahl der zu schätzenden Modellparameter ist.³ Bei Anwendung der latenten Profilanalyse kann immer davon ausgegangen werden, dass das untersuchte Modell identifiziert ist, da die notwendige Bedingung (mehr empirische Informationen als Modellparameter) in der Regel erfüllt ist und Mischungen von m -dimensionalen Verteilungen identifizierbar sind (Kaufmann und Pape 1984, S. 422–423). Bei den anderen Verfahren muss dies nicht unbedingt der Fall sein. Die Identifikation eines Modells kann dadurch geprüft werden, dass die gesuchten Modellparameter als Funktion der empirischen Daten ausgedrückt werden (Bacher 1990, S. 92–95; Lazarsfeld und Henry 1968, S. 59–68).⁴

² Überlappungen können zwei Ursachen haben: Messfehler in den Variablen und nicht deutlich getrennte Cluster. Ein weiterer Einflussfaktor ist die Stichprobengröße (siehe Abschnitt 17.3).

³ Bei einer gleichen Zahl zu schätzender Parameter und empirischer Information spricht man von einem »saturierten« Modell, bei einer größeren Zahl empirischer Informationen von einem »überidentifizierten« Modell. Statistische Tests sind nur in überidentifizierten Modellen möglich.

⁴ Mitunter erfordert dieses Vorgehen aber komplexe mathematische Operationen. Deshalb wurden Verfahren zur computerunterstützten Identifikationsprüfung entwickelt (Van de Pol, Langeheine u. a. 1989, S. 29). Dabei wird geprüft, ob lineare Abhängigkeiten zwischen den geschätzten Parametern bestehen. Ist dies der Fall, ist das Modell nicht identifiziert. Umgekehrt kann allerdings aus dem Fehlen von linearen Abhängigkeiten nicht abgeleitet werden, dass ein Modell identifiziert ist.

14 Latente Profilanalyse

14.1 Modellansatz und Algorithmus

Das Modell der *latenten Profilanalyse* bzw. der *Analyse latenter Klassen für quantitative Variablen* wurde im Rahmen der Analyse latenter Strukturen (Gibson 1959; Lazarsfeld 1966; Lazarsfeld und Henry 1968, S. 228–239) entwickelt. Die Modellannahmen sind:

1. Es liegen K latente Klassen vor.
2. Diese besitzen Anteilswerte von $\pi(k)$ in der Grundgesamtheit.
3. Jede Klasse k besitzt in jeder Klassifikationsvariablen j eine Normalverteilung¹ mit dem Mittelwert (Klassenzentrum) μ_{kj} und der Varianz σ_{kj}^2 .
4. In jeder Klasse k sind die Variablen j und j^* paarweise unabhängig.

Die Normalverteilung in den Variablen kommt wie folgt zustande: Der empirisch beobachtete Wert x_{gj} eines Objekts g aus k in der Variablen X_j setzt sich aus einem zufälligen Fehlerterm ε_{gj} und dem Klassenmittelwert μ_{kj} zusammen: $x_{gj} = \mu_{kj} + \varepsilon_{gj}$. Der zufällige Fehlerterm ε_{gj} ist die Realisierung einer normalverteilten Zufallsvariablen ξ_{kj} mit Erwartungswert 0 und Varianz σ_{kj}^2 . Er kann durch zufällige Messfehler und/oder zufällige individuelle Unterschiede in den einzelnen Variablen entstehen. Da die Abweichungen $\varepsilon_{gj} = x_{gj} - \mu_{kj}$ zufällig auftreten, sind sie in zwei Variablen j und j^* unabhängig. Wäre dies nicht der Fall, wäre es nicht sinnvoll, von Zufälligkeit zu sprechen. Formal ausgedrückt sind die Annahmen:

1. $x_{gj} = \mu_{kj} + \varepsilon_{gj} \quad \forall g \in k$.
2. ε_{gj} ist die Realisierung einer normalverteilten Zufallsvariablen ξ_{kj} .
3. ξ_{kj} besitzt einen Erwartungswert von 0 und eine Varianz von σ_{kj}^2 .
4. Die Zufallsvariablen ξ_{kj} sind unabhängig. Es gilt also $\text{cov}(\xi_{kj}, \xi_{kj^*}) = 0$.

Hierbei ist x_{gj} der empirische Wert des Objekts g aus der Klasse k in der Variablen j , μ_{kj} der Klassenmittelwert in der Variablen j und ε_{gj} der zufällige Fehlerterm.

¹ In dem klassischen Ansatz der latenten Profilanalyse ist eine Verteilungsannahme nicht erforderlich, da die Modellparameter nicht über die Maximum-Likelihood-Methode geschätzt werden.

Aus diesen Modellannahmen lassen sich folgende Aussagen über die Mittelwerte, Varianzen und Kovarianzen der Gesamtpopulation ableiten (Lazarsfeld und Henry 1968, S. 229–231):

$$\mu_j = \sum_k \pi(k) \cdot \mu_{kj} , \quad (14.1)$$

$$\sigma_{jj^*} = \sum_k \pi(k) \cdot (\mu_{kj} - \mu_j) \cdot (\mu_{kj^*} - \mu_{j^*}) , \quad (14.2)$$

$$\sigma_j^2 = \sigma_{jj} = \sum_k \pi(k) \cdot \sigma_{kj}^2 + \sum_k \pi(k) \cdot (\mu_{kj} - \mu_j)^2 . \quad (14.3)$$

Der Gesamtmittelwert einer Variablen ist gleich dem gewichteten Mittelwert der Klassenzentren. Die Kovarianz zwischen zwei Variablen j und j^* hängt nur von den Abweichungen der Klassenzentren von den Gesamtmittelwerten ab, die Varianz einer Variablen dagegen auch noch von den Fehlerstreuungen (siehe dazu auch Abschnitt 7.7). Lazarsfeld und Henry (1968, S. 231–233) sowie Lazarsfeld (1966) entwickelten auf der Grundlage der Modellgleichungen 14.1 bis 14.3 sogenannte »accounting equations«, die eine Schätzung der Modellparameter mit Hilfe einer Eigenwertzerlegung ermöglichen. Diese hat mehrere Nachteile (Van de Pol und de Leeuw 1986). Zum Beispiel können negative Zuordnungswahrscheinlichkeiten entstehen. Eine andere Schätzmethode besteht in der Verwendung des EM-Algorithmus. Die Modellparameter werden so geschätzt, dass die *Likelihood-Funktion* (L)

$$L = \prod_g \sum_k \pi(k) \cdot \pi(g|k)$$

bzw. ihr natürlicher Logarithmus, die so genannte *Log-Likelihood-Funktion* (LL)²

$$LL = \log L = \sum_g \log \sum_k \pi(k) \cdot \pi(g|k) \quad (14.4)$$

ein Maximum annimmt. Dabei ist $\pi(k)$ der »wahre« Anteil der Klasse k und $\pi(g|k)$ die (bedingte) Wahrscheinlichkeit des Auftretens des Merkmalsvektors der Person g in der Klasse k . Die bedingte Wahrscheinlichkeit $\pi(g|k)$ ist wegen der lokalen Unabhängigkeit (siehe Kapitel 13) gleich dem Produkt der (bedingten) Auftrittswahrscheinlichkeiten $\pi(x_{gi}|k)$ des Wertes von g in den Variablen j für die Klasse k :

$$\pi(g|k) = \prod_j \pi(x_{gj}|k) . \quad (14.5)$$

² In der angelsächsischen Literatur ist die Abkürzung »log« für den natürlichen Logarithmus üblich, in der deutschen die Abkürzung »ln«. Wir verwenden hier, nicht zuletzt aufgrund der üblichen Bezeichnung »Log-Likelihood«, die angelsächsische Schreibweise.

Die Auftrittswahrscheinlichkeit $\pi(x_{gj}|k)$ ist bei der latenten Profilanalyse als Wert der Dichtefunktion φ der Normalverteilung an der Stelle x_{gj} bei gegebenem Mittelwert μ_{kj} und gegebener Standardabweichung σ_{kj}^2 definiert:

$$\pi(x_{gj}|k) = \varphi(x_{gj}|\mu_{kj}, \sigma_{kj}^2) = \frac{1}{\sigma_{kj}\sqrt{2\pi}} \cdot \exp\left(-\frac{(x_{gj} - \mu_{kj})^2}{2\sigma_{kj}^2}\right). \quad (14.6)$$

Betrachten wir dazu ein Beispiel: Die Klasse k soll in der Variablen j einen Mittelwert μ_{kj} von -1 und eine Varianz σ_{kj}^2 von 2 haben. Das Objekt g besitzt in der Variablen j einen Wert von -2 . Die Auftrittswahrscheinlichkeit des Wertes von -2 in der Klasse k ist entsprechend Gleichung 14.6 gleich dem Wert der Dichtefunktion der Normalverteilung mit den Modellparametern der Klasse k :

$$\begin{aligned} \pi(x_{gj} = -2|k) &= \varphi(x_{gj} = -2|\mu_{kj} = -1, \sigma_{kj}^2 = 2) \\ &= \frac{1}{1,41 \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(-2 - (-1))^2}{2 \cdot 2}\right) \\ &= 0,2197. \end{aligned}$$

Der Wert -2 tritt also mit einer Wahrscheinlichkeit von $0,2197$ in der Klasse k auf. Dieser Wert ergibt sich auch, wenn mit $(-2 - (-1))/\sqrt{2} = 0,707$ der z-Wert berechnet und die Dichtefunktion der Standardnormalverteilung dividiert durch die Standardabweichung verwendet wird. Die gesuchten Parameter $\pi(k)$, μ_{kj} und σ_{kj}^2 werden nun so bestimmt, dass die Log-Likelihood-Funktion ihr Maximum annimmt. Das heißt, es werden Schätzungen gesucht, die (theoretische) Modelldaten mit einer Verteilung erzeugen, die der empirischen Verteilung der Objekte bestmöglich angepasst ist. In die Schätzung gehen die zwei Nebenbedingungen $\sum \pi(k) = 1$ und $\pi(k) > 0$ für alle k ein. Die erste Nebenbedingung besagt, dass die Summe der Klassenanteilswerte gleich 1 sein soll. Die zweite Nebenbedingung bedeutet, dass keine Klasse leer sein soll. Unter Berücksichtigung dieser Nebenbedingung sind bei K Klassen und m Variablen insgesamt $K(1 + 2m) - 1$ Parameter (vgl. Tabelle 14.1) zu schätzen. Wird beispielsweise eine 6-Klassenlösung bei drei Variablen ($m = 3$) gesucht, sind $6 \cdot (1 + 2 \cdot 3) - 1 = 41$ Parameter zu schätzen. Die notwendige Bedingung für ein identifiziertes Modell ist, dass mindestens 41 unterschiedliche Datensätze vorliegen. Die notwendige Bedingung wird in der Regel bei der latenten Profilanalyse immer erfüllt sein. Wie bereits erwähnt, ist beim Vorliegen der notwendigen Identifikationsbedingung das Modell der latenten Profilanalyse dann auch identifiziert.

Beim *EM-Algorithmus* wird nun angenommen, dass die Zuordnungswahrscheinlichkeiten $\pi(k|g)$, mit der die Klasse k bei den Objekten g auftritt, bekannt sind. Dadurch verein-

Tab. 14.1: Berechnungsschema für die Anzahl zu schätzender Parameter bei der latenten Profilanalyse

$K - 1$	Anzahl der Klassenanteilswerte $\pi(k)$: ein Klassenanteilswert ist wegen der Nebenbedingung $\sum \pi(k) = 1$ fixiert
Km	Anzahl der Klassenzentren oder -mittelwerte μ_{kj} : in jeder Klasse für jede Variable ein Klassenzentrum
Km	Anzahl der Klassenstreuungen oder -varianzen σ_{kj}^2 : in jeder Klasse für jede Variable eine Klassenvarianz
$K(1 + 2m) - 1$	Gesamtzahl zu schätzender Parameter bei K Klassen und m Variablen

facht sich die Log-Likelihood-Funktion für die Schätzung (Van de Pol und de Leeuw 1986):

$$\begin{aligned} \text{LL} &= \sum_g \sum_k \pi(k|g) \cdot (\log \pi(k) + \log \pi(g|k)) \\ &= \sum_g \sum_k \pi(k|g) \cdot \log \pi(k) + \sum_g \sum_k \sum_j \pi(k|g) \pi(x_{gj}|k). \end{aligned} \quad (14.7)$$

Die Schätzaufgabe zerfällt also zum einen in die Schätzung der Anteilswerte $\pi(k)$ und zum anderen in die Schätzung der Parameter der einzelnen Normalverteilungen. Die Schätzwerte sind:³

$$p(k) = \frac{\sum_g \pi(k|g)}{n}, \quad (14.8)$$

$$\bar{x}_{kj} = \frac{\sum_g \pi(k|g)x_{gj}}{\sum_g \pi(k|g)}, \quad (14.9)$$

$$s_{kj}^2 = \frac{\sum_g \pi(k|g)(x_{gj} - \bar{x}_{kj})^2}{\sum_g \pi(k|g)}, \quad (14.10)$$

wobei der Schätzwert von $\pi(k)$ mit $p(k)$ bezeichnet wird und n die Anzahl der Objekte ist. Der Schätzwert von μ_{kj} wird mit \bar{x}_{kj} und jener von σ_{kj}^2 mit s_{kj}^2 bezeichnet. Die Schätzwerte kann man sich wie folgt vorstellen: Zur Berechnung der Parameter der Klasse k werden die Datensätze mit $\pi(k|g)$ gewichtet und ausgezählt. Dadurch erhält man die Mittelwerte und Varianzen der Klasse k in den Variablen. Die gewichtete Fallzahl dividiert durch die ungewichtete Fallzahl ergibt den Anteil der Klasse k .

Bei den bisherigen Ausführungen wurde davon ausgegangen, dass die Zuordnungswahrscheinlichkeiten $\pi(k|g)$ bekannt sind. Dies ist natürlich nicht der Fall. Sie können aber

³ Bock (1974, S. 258) und Kaufmann und Pape (1984, S. 432) leiten die Schätzfunktion direkt aus der Gleichung 14.4 ab.

ihrerseits aus den geschätzten Modellparametern mit Hilfe des Satzes von Bayes (Fisz 1980, S. 40–41) berechnet werden mit

$$p(k|g) = \frac{p(k)p(g|k)}{\sum_k p(k) \cdot p(g|k)}, \quad (14.11)$$

wobei $p(k|g)$ der Schätzwert von $\pi(k|g)$, $p(k)$ der Schätzwert von $\pi(k)$ und $p(g|k)$ der Schätzwert von $\pi(g|k)$ ist.

Damit haben wir das Grundprinzip des EM-Algorithmus skizziert. Es besteht aus zwei Schritten:

E-Schritt: Die Zuordnungswahrscheinlichkeiten $\pi(k|g)$ werden aufgrund der geschätzten Modellparameter (Erwartungswerte) berechnet. Die im vorangegangenen M-Schritt geschätzten Modellparameter werden dabei als gegeben angenommen.

M-Schritt: Die Modellparameter $\pi(k)$, μ_{kj} und σ_{kj}^2 werden aufgrund der Zuordnungswahrscheinlichkeiten nach der *Maximum-Likelihood-Methode* geschätzt. Die in dem vorangegangenen E-Schritt ermittelten Zuordnungswahrscheinlichkeiten werden dabei als gegeben angenommen.

Diese Schritte werden solange wiederholt, bis eine konvergente Lösung gefunden ist. Es ergibt sich folgender Schätzalgorithmus:

Schritt 1: Berechnung oder Eingabe von Startwerten für die Modellparameter oder für die Zuordnungswahrscheinlichkeiten. Bei Startwerten für die Zuordnungswahrscheinlichkeiten gehe zu Schritt 3.

Schritt 2: Berechnung der Zuordnungswahrscheinlichkeiten. Entsprechend der Gleichung 14.11 werden die Zuordnungswahrscheinlichkeiten berechnet mit:

$$p(k|g)^{(i)} = \frac{p(k)^{(i-1)} \cdot p(g|k)^{(i-1)}}{\sum_k (p(k)^{(i-1)} \cdot p(g|k)^{(i-1)})},$$

wobei aufgrund der Gleichungen 14.5 auf Seite 356 und 14.6 auf Seite 357 gilt:

$$p(g|k)^{(i-1)} = \prod_j p(x_{gj}|k)^{(i-1)} = \prod_j \varphi\left(x_{gj}|\bar{x}_{kj}^{(i-1)}, s_{kj}^{2(i-1)}\right).$$

Der hochgestellte Index in Klammern ist der Iterationszähler. Beim ersten Durchlauf ist $i = 1$.

Schritt 3: Neuberechnung der Modellparameter. Die Modellparameter werden entsprechend den Gleichungen 14.8 bis 14.10 neu berechnet mit:

$$\begin{aligned} p(k)^{(i)} &= \frac{\sum_g p(k|g)^{(i)}}{n}, \\ \bar{x}_{kj}^{(i)} &= \frac{\sum_g p(k|g)^{(i)} \cdot x_{gj}}{p(k|g)^{(i)}}, \\ s_{kj}^{2(i)} &= \frac{\sum_g p(k|g)^{(i)} \cdot (x_{gj} - \bar{x}_{kj}^{(i)})^2}{\sum_g p(k|g)^{(i)}}. \end{aligned}$$

Schritt 4: Prüfung der Konvergenz. Der Algorithmus wird dann abgebrochen, wenn die Verbesserung der Log-Likelihood-Funktion kleiner einem vorgegebenen Schwellenwert (zum Beispiel 10^{-7}) und/oder die maximale Abweichung der aufeinander folgenden Schätzwerte kleiner einem zweiten Schwellenwert (zum Beispiel 10^{-4}) ist.

Der Algorithmus ist mit dem K-Means-Algorithmus strukturgleich. Sein asymptotisches Konvergenzverhalten für $n \rightarrow \infty$ wurde bereits im einleitenden Kapitel 13 angeprochen und das Verhalten des K-Means-Algorithmus in Abschnitt 12.1 ausführlich untersucht. Zur Vermeidung von lokalen Minima ist das Durchrechnen mit mehreren Startwerten sinnvoll, zum Beispiel, indem unterschiedliche zufällige Zuordnungswahrscheinlichkeiten als Startwerte erzeugt werden (zu multiplen zufälligen Startwerten siehe Abschnitt 12.2).

Zur Veranschaulichung des Algorithmus soll ein Iterationsschritt für die (fiktiven) Daten der Tabelle 11.1 auf Seite 287 durchgerechnet werden (siehe Tabelle 14.2 auf Seite 362). Als Startwerte wurden die in der Tabelle 14.2a auf Seite 362 wiedergegebenen Werte verwendet. Für die erste Klasse wurde ein Anteilswert von 0,333 angenommen. Für die Klassenzentren der ersten Klasse wurden Startwerte von 1,67 in X_1 und von 1,00 in X_2 gewählt, für die Klassenstreuungen (Varianzen) Werte von 1,51 und 0,92. Für das erste Objekt A ergibt sich aufgrund dieser Startwerte eine Zordnungswahrscheinlichkeit $p(1|g = A)$ von 0,017 zur ersten Klasse. Diese berechnet sich wie folgt: Die bedingte Auftrittswahrscheinlichkeit $p(x_{1A} = -2|1)$, dass beim Objekt A in X_1 der Wert -2 auftritt, wenn die Klasse 1 vorliegt, ist gleich dem Wert der Dichtefunktion der Normalverteilung an der Stelle $2,987 = (-2 - 1,67)/\sqrt{1,51}$. Dieser Wert ist ungefähr gleich 0,0038. Analog berechnet sich die bedingte Auftrittswahrscheinlichkeit $p(x_{2A} = 1|2)$. Für sie ergibt sich ein Wert von 0,4159. Die bedingte Wahrscheinlichkeit $p(A|1)$ des Auftretens der Ausprägungen des Objekts A in der Klasse 1 ist daher gleich $p(A|1) = p(x_{1A}|1) \cdot p(x_{2A}|2) = 0,0038 \cdot 0,4159 = 0,0016$. Dieser Wert ist in der mit » $p(g|1)$ « überschriebenen Spalte wiedergegeben. Berechnen wir analog $p(A|2)$ und

$p(A|3)$, ergeben sich Werte von 0,0914 und 0,0001. Für die Zuordnungswahrscheinlichkeit $p(1|A)$ ergibt sich nun entsprechend Gleichung 14.11 auf Seite 359

$$p(1|A) = \frac{p(1)p(A|1)}{p(1)p(A|1) + p(2)p(A|2) + p(3)p(A|3)}$$

ein Wert von

$$p(1|A) = \frac{0,333 \cdot 0,0017}{0,333 \cdot 0,0017 + 0,333 \cdot 0,0914 + 0,333 \cdot 0,0001} = 0,017.$$

Nach derselben Methode können die Zuordnungswahrscheinlichkeiten $p(2|A)$ und $p(3|A)$ berechnet werden. Es ergeben sich die Werte 0,982 und 0,001. Den Nenner bei der Berechnung der Zuordnungswahrscheinlichkeiten benötigen wir zur Berechnung des Wertes der Log-Likelihood-Funktion. Für das erste Objekt ergibt sich ein Wert von $0,333 \cdot 0,0017 + 0,333 \cdot 0,0914 + 0,333 \cdot 0,0001 = 0,031$. Dieser Wert ist in der mit p_{ges} überschriebenen Spalte eingetragen. Der genaue Wert für das Objekt A ist gleich 0,03095. Der natürliche Logarithmus dieses Wertes ist $-3,4753$ und in der mit $\log p_{\text{ges}}$ überschriebenen Spalte eingetragen. Aufsummiert ergeben die p_{ges} -Werte die Likelihood L und die $(\log p_{\text{ges}})$ -Werte die Log-Likelihood LL. Insgesamt ergibt sich ein Log-Likelihood-Wert von $-39,4657$.

Nach der Berechnung der Zuordnungswahrscheinlichkeiten können die Modellparameter berechnet werden. Für die Klasse 1 ergibt sich ein Anteilswert von

$$p(1) = \frac{0,017 + 0,125 + 0,033 + \dots + 0,230}{9} = 0,222,$$

also von 22,2 Prozent. Der Mittelwert in der Variablen X_1 ist gleich

$$\bar{x}_{11} = \frac{0,017(-2) + 0,125(-1) + \dots + 0,2304}{0,017 + 0,125 + \dots + 0,230} = 1,92.$$

Für die Varianz in X_1 in der ersten Klasse ergibt sich ein Wert von

$$s_{11}^2 = \frac{0,017(-2 - 1,92)^2 + 0,125(-1 - 1,92)^2 + \dots + 0,230(4 - 1,92)^2}{0,017 + 0,125 + \dots + 0,230} = 2,36.$$

Die anderen Modellparameter können analog berechnet werden. Sie sind in der Tabelle 14.2a auf der nächsten Seite unter der Spaltenbeschriftung »nach 1. Iteration« wiedergegeben. Da sie sich von den Startwerten unterscheiden, wird man eine erneute Iteration durchführen. Insgesamt werden neun Iterationen benötigt, bis sich der Wert der Log-Likelihood-Funktion an der siebten Kommastelle nicht mehr ändert. Es ergibt sich ein Wert von $-26,094$. Aus diesem Wert lassen sich – wie aus den Streuungsquadratsummen in den Clustern beim K-Means-Verfahren – Modellprüfgrößen berechnen.

Tab. 14.2: Veranschaulichung des EM-Algorithmus für die latente Profilanalyse

	Startwerte			nach 1. Iteration		
	C_1	C_2	C_3	C_1	C_2	C_3
<i>Anteilswert</i>						
$p(k)$	0,33	0,33	0,33	0,222	0,444	0,333
<i>Mittelwert</i>						
X_1	1,67	-1,00	2,67	1,92	-0,90	3,13
X_2	1,00	0,67	1,00	1,09	-0,06	1,98
<i>Varianz</i>						
X_1	1,51	1,51	1,51	2,36	0,79	1,24
X_2	0,92	0,92	0,92	2,36	2,43	1,13

(a) Modellparameter

g	X_1	X_2	$p(g 1)$	$p(g 2)$	$p(g 3)$	$p(1 g)$	$p(2 g)$	$p(3 g)$	Σ	p_{ges}	$\log p_{\text{ges}}$
A	-2	1	0,002	0,091	0,000	0,017	0,982	0,001	1,0	0,03095	-3,4753
B	-1	2	0,007	0,051	0,001	0,125	0,859	0,015	1,0	0,01992	-3,9160
C	-1	-2	0,000	0,003	0,000	0,033	0,963	0,004	1,0	0,00099	-6,9178
D	0	-1	0,006	0,021	0,001	0,211	0,738	0,051	1,0	0,00969	-4,6367
E	1	-1	0,013	0,008	0,006	0,485	0,291	0,224	1,0	0,00914	-4,6951
F	2	2	0,075	0,003	0,068	0,518	0,018	0,464	1,0	0,04858	-3,0245
G	3	2	0,044	0,000	0,075	0,365	0,002	0,633	1,0	0,03976	-3,2249
H	4	2	0,013	0,000	0,044	0,230	0,000	0,770	1,0	0,01884	-3,9718
I	4	3	0,003	0,000	0,009	0,230	0,000	0,770	1,0	0,00370	-5,5994
Σ										-39,4657	

(b) Berechnung der Zuordnungswahrscheinlichkeiten

14.2 Modellprüfgrößen

14.2.1 Bestimmung der Klassenzahl

Zur *Bestimmung der Klassenzahl* wird das Verfahren mit einer unterschiedlichen Anzahl von Klassen durchgerechnet. Wichtig ist wiederum, die 1-Clusterlösung miteinzubeziehen, um prüfen zu können, ob überhaupt eine Cluster- bzw. Klassenstruktur vorliegt. Für die Wertedaten von Denz (1989) ergeben sich die in der Tabelle 14.3 dargestellten Werte für die Log-Likelihood-Funktion, wenn die Gesamtpunktwerte für die postmaterialistische und materialistische Wertorientierung als Klassifikationsvariablen in die Analyse einbezogen werden. Mittels der Methode der multiplen zufälligen Startwerte (siehe Abschnitt 12.2) wurden für die Berechnung 50 Startwerte für jede Clusterzahl erzeugt. Aus

Tab. 14.3: Modellprüfgrößen der latenten Profilanalyse für die Wertedaten von Denz (multiple zufällige Startwerte, $n = 221$)

Cluster zahl K	Wert der Log- Likelihood- funktion LL_K	prozentuelle Verbesserung gegenüber Nullmodell PVO_K	prozentuelle Verbesserung gegenüber vorausgehender Lösung PV_K	AIC_K	BIC_K	$CAIC_K$
1	-374,100	—	—	756,201	769,793	773,793
2	-330,417	11,677	11,677	678,833	709,417	718,417
3	-316,099	15,504	4,333	660,198	707,773	721,773
4	-303,603	18,844	3,953	645,205	709,770	728,770
5	-292,559	21,797	3,638	633,118	714,674	738,674
6	-286,292	23,472	2,142	630,585	729,131	758,131
7	-283,439	24,234	0,997	634,878	750,415	784,415
8	-280,111	25,124	1,174	638,223	770,751	809,751
9	-278,203	25,634	0,681	644,406	793,926	837,926
10	-277,608	25,793	0,214	653,216	819,726	868,726
11	-275,014	26,487	0,934	658,029	841,529	895,529
12	-273,961	26,768	0,383	665,922	866,414	925,414

grau hinterlegt: besonders zu berücksichtigende Größen

den Werten der Log-Likelihood-Funktion⁴ LL_K lassen sich folgende Modellprüfgrößen bestimmen:

Prozentuelle Verbesserung gegenüber dem Nullmodell der 1-Klassenlösung (PVO_K): Diese wird analog zu η^2 (siehe Abschnitt 12.2) berechnet mit

$$PVO_K = 1 - \frac{|LL_K|}{|LL_1|} \quad \text{bzw.} \quad PVO_K = 100 \cdot PVO_K \% ,$$

wobei $|LL_1|$ der Absolutbetrag der Log-Likelihood-Funktion für die 1-Klassenlösung und $|LL_K|$ der Absolutbetrag der Log-Likelihood-Funktion der K -Klassenlösung ist. Für die 5-Klassenlösung beispielsweise beträgt die prozentuelle Verbesserung 21,797 Prozent, da der Absolutbetrag $|LL_1| = 374,101$ und $|LL_5| = 292,559$ ist. PVO_5 ist daher gleich $1 - 292,559/374,101 = 0,218$ (= 21,8 Prozent).

⁴ Allgemein bedeutet ein kleinerer negativer Wert eine bessere Modellanpassung.

Prozentuelle Verbesserung gegenüber der vorausgehenden Klassenlösung (PV_K): Diese Maßzahl ist analog dem PRE-Koeffizienten beim K-Means-Verfahren (siehe Abschnitt 12.2) definiert mit

$$PV_K = 1 - \frac{|LL_K|}{|LL_{K-1}|} \quad \text{bzw.} \quad PV_K = 100 \cdot PV_K \% .$$

Für die 5-Klassenlösung ergibt sich mit $|LL_{5-1}| = |LL_4| = 303,603$ und $|LL_5| = 292,559$ ein Wert von $1 - 292,559/303,603 = 0,03638 (= 3,6\text{ Prozent})$.

Das Informationsmaß von Akaike AIC: Dieses Kriterium (Akaike 1973, 1974; Kaufmann und Pape 1984, S. 443) ist definiert mit:

$$AIC_K = -2(LL_K - m_K) = -2LL_K + 2m_K , \quad (14.12)$$

wobei m_K die Zahl der zu schätzenden Parameter ist. Für die latente Profilanalyse ohne Restriktionen müssen nach Tabelle 14.1 auf Seite 358 $m_K = K + 2mK - 1$ Parameter geschätzt werden. Für fünf Klassen und zwei Variablen sind das 24 ($= 5 + 2 \cdot 2 \cdot 5 - 1$) zu schätzende Parameter. Das Informationsmaß für die 5-Klassenlösung ist daher $2(-292,559 - 24) = 633,118$. Das Informationsmaß von Akaike berücksichtigt wie die F_{max}-Statistik (siehe Abschnitt 12.2) die Tatsache, dass bei einer größeren Klassenzahl in der Tendenz »automatisch« eine bessere Modellanpassung erzielt wird.

Neben dem Informationsmaß von Akaike stehen noch weitere Informationsmaße zur Verfügung (siehe Abschnitt 16.4). In der Tabelle 14.3 auf der vorherigen Seite sind noch BIC (*Bayesian Information Criterion*, Schwarz 1978) und CAIC (*Consistent Akaike Information Criterion*, Akaike 1987; Bozdogan 1987) eingetragen. Die beiden Größen sind definiert als:

$$BIC_K = -2LL + m_K \log n$$

und

$$CAIC_K = -2LL_K + m_K(\log n + 1) .$$

χ^2 -Test für die Likelihood-Quotienten-Teststatistik⁵: In älteren Lehrbüchern wird die Verwendung der Teststatistik $-2 \cdot (LL_{K-1} - LL_K)$ empfohlen ($-2LL$ -Differenzen-Statistik)⁶

⁵ Die englische Bezeichnung lautet »likelihood-ratio-test« und wird meist mit LR abgekürzt.

⁶ Aufgrund der Rechenregeln für Logarithmen gilt:

$$LL_{K-1} - LL_K = \log L_{K-1} - \log L_K = \log \left(\frac{L_{K-1}}{L_K} \right) .$$

Es handelt sich also um eine Log-Likelihood-Differenz, was gleichbedeutend ist mit einem (logarithmierten) Likelihood-Quotienten.

Tab. 14.4: Beziehung zwischen den Modellprüfgrößen der latenten Profilanalyse und jenen des K-Means-Verfahrens

Modellprüfgrößen der latenten Profilanalyse	Modellprüfgrößen des K-Means-Verfahrens	Anwendung zur Bestimmung der Klassenzahl
prozentuelle Verbesserung gegenüber 1-Klassenlösung PV_{OK}	Erklärte Streuung η_K^2	Es werden nur jene Lösungen ausgewählt, für die PV_{OK} einen bestimmten Wert überschreitet.
prozentuelle Verbesserung gegenüber vorausgehender Lösung PV_K	PRE-Koeffizient PRE_K	Es wird (werden) jene Lösung(en) ausgewählt, bei der (denen) nachfolgend ein deutlicher Abfall auftritt.
Informationsmaße (AIC_K , BIC_K usw.)	Maximaler F_{\max} -Wert $F_{\max}(K)$	Es wird jene Lösung mit dem kleinsten Informationsmaß ausgewählt.
Likelihood-Quotienten-Statistiken (gewöhnliche LQ-Statistik und LQ_K (Wolfe))	Bealsche F-Werte	Es wird jene Lösung ausgewählt, die a) im Vergleich zu allen vorausgehenden Lösungen signifikant und b) im Vergleich zu allen nachfolgenden Lösungen nicht signifikant ist. Zur Signifikanzprüfung sollten Bootstrap-Techniken eingesetzt werden.

oder der *modifizierte Likelihood-Quotienten-Teststatistik nach Wolfe* (Kaufmann und Pape 1984, S. 443):

$$LQ_K(\text{Wolfe}) = -\frac{2}{n} \left(n - 1 - m - \frac{K}{2} \right) \cdot (LL_{K-1} - LL_K) .$$

Von diesen Testgrößen wurde angenommen, dass sie approximativ χ^2 -verteilt seien mit $df = m_K - m_{K-1}$ bzw. $df = 2m$ Freiheitsgraden. Heute wird von der Verwendung der χ^2 -Approximation für die $-2LL$ -Differenzen- bzw. LQ-Teststatistik abgeraten, da für sie die approximativen Eigenschaften *nicht erfüllt sind*, das heißt, die Statistiken besitzt asymptotisch *keine* χ^2 -Verteilung (McLachlan und Peel 2000, S. 185–193). Kaufmann und Pape (1996, S. 508–509) ziehen in ihrer Neuauflage auch das Verfahren nach Wolfe in Zweifel und verweisen auf eine Diskussion in McLachlan und Basford (1988). Zur Signifikanzprüfung werden daher heute im Allgemeinen Bootstrap-Verfahren⁷ empfohlen (Efron 1979), die für die dargestellten Differenzen bzw. Quotienten p-Werte liefern, anhand derer Signifikanzaussagen getroffen werden können (siehe dazu Abschnitt 16.4.4).

⁷ Bei diesem Verfahren wird die unbekannte Verteilung einer Statistik durch die synthetische Erzeugung weiterer Datensätze und die erneute Berechnung der Statistiken näherungsweise bestimmt. Aus den empirischen Originaldaten werden dabei mehrfach Stichproben »mit Zurücklegen« gezogen und analysiert (Davier 1997, S. 51; Rost 2004, S. 336).

Die bereits erwähnte Analogie der Modellprüfgrößen zu den Modellprüfgrößen des K-Means-Verfahrens gilt auch für das Vorgehen bei der Bestimmung der Klassenzahl (siehe Tabelle 14.4 auf der vorherigen Seite). Wendet man die einzelnen Strategien an, würden wir uns für folgende Lösungen entscheiden:

- *Prozentuelle Verbesserung PV_K gegenüber vorausgehender Lösung:* Für die 2- und 6-Klassenlösung, da hier anschließend ein Abfall eintritt. Absolut betrachtet sind die prozentuellen Verbesserungen aber gering (< 10 Prozent).
- *Informationsmaße:* Für die 6-Klassenlösung, da hier der Wert des Informationsmaßes AIC_K mit 630,585 am kleinsten ist. Allerdings ist der Wert der 5- und 7-Klassenlösung mit 633,118 bzw. 634,585 nur geringfügig größer. Für BIC_K ergibt sich ein Minimum bei drei Clustern, für $CAIC_K$ bei zwei Clustern.
- *Likelihood-Quotienten-Test:* Dieser steht in ALMO nicht zur Verfügung, die Bootstrap-Testung wird in Abschnitt 16.4 behandelt.
- *Prozentuelle Verbesserung gegenüber Nullmodell der 1-Klassenlösung:* Die Entscheidung hängt ab von einem Schwellenwert, der vorab spezifiziert werden muss. Wird ein Wert von 20 Prozent vorgegeben, würde die 5-Klassenlösung ausgewählt werden.

Für eine bestimmte Klassenlösung können zur Beschreibung die oben genannten entsprechenden *Modellprüfgrößen* verwendet werden. Darüber hinaus können – wie beim K-Means-Verfahren – varianzanalytische Maßzahlen verwendet werden, insbesondere die erklärte Streuung (η^2 , siehe Gleichung 12.3 auf Seite 306). In unserem Beispiel ergibt sich eine erklärte Streuung von 31,9 Prozent für die 3-Klassenlösung. Die 5-Klassenlösung weist eine erklärte Streuung von 52,6 Prozent auf und die 6-Klassenlösung von 61,5 Prozent. Die erklärte Streuung bei der latenten Profilanalyse ist in der Regel kleiner als beim K-Means-Verfahren, da die Fehlerstreuung nicht minimiert wird. Beim K-Means-Verfahren ergibt sich eine erklärte Streuung von 75,2 Prozent für die 5-Clusterlösung.

14.2.2 Zufallstestung einer Klassenlösung

Wie beim K-Means-Verfahren kann auch bei der latenten Profilanalyse mit Hilfe des Nullmodells einer homogenen, normalverteilten Population geprüft werden, ob eine bestimmte Klassenlösung überzufällig ist. Dazu werden Zufallsdatenmatrizen für das homogene Nullmodell erzeugt. Für diese wird geprüft, wie gut sie durch die berechnete Klassenlösung reproduziert werden können. Es wird dazu das Modell mit vorgegebener Klassenstruktur verwendet. Ergibt sich eine annähernd gleich gute Reproduktion – gemessen durch den Wert der Log-Likelihood-Funktion –, wird man die Lösung als Zufallsprodukt betrachten.

Tab. 14.5: Simulationswerte für die Log-Likelihood-Funktion für die 5-Klassenlösung aus einer Zufallstestung bei 20 Simulationen

-452,339	-417,273	-447,587	-475,383	-432,936	-433,723	-504,192
-461,704	-430,683	-433,043	-461,709	-407,436	-441,347	-492,683
-503,631	-504,743	-537,952	-446,600	-481,406	-494,642	

Führt man 20 Simulationen durch, ergeben sich die in der Tabelle 14.5 dargestellten Werte für die 5-Klassenlösung: Alle berechneten Log-Likelihood-Werte sind unter der Annahme einer homogenen Population deutlich kleiner als der empirische Wert von -292,559. Der Mittelwert der Simulationswerte ist gleich -463,051, die Standardabweichung hat einen Wert von 33,953. Konstruieren wir eine z-Teststatistik mit $z = (t - E(t))/\sigma(t)$, wobei t der Wert der empirischen Log-Likelihood-Funktion (LL_K), $E(t)$ der Mittelwert der Log-Likelihood-Werte der simulierten Daten und $\sigma(t)$ deren Standardabweichung ist, ergibt sich ein Wert von 5,02. Dieser ist größer als der kritische Schwellenwert von 2 ($p < 0,05$). Wir können daher die 5-Klassenlösung als überzufällig betrachten. Für die anderen Klassen kann nach demselben Prinzip verfahren werden.

14.3 Beschreibung und Interpretation einer Klassenlösung

Bei der *Beschreibung und Interpretation einer Klassenlösung* wird analog wie beim K-Means-Verfahren vorgegangen. Das bedeutet unter anderem:

1. Für jede Variable kann geprüft werden, ob sie signifikant zur Trennung der Klassen beiträgt. Dazu wird die durch eine Variable erklärte Streuung und ein entsprechender F-Wert berechnet. Da im Unterschied zum K-Means-Verfahren nicht die Streuungsquadratsumme in den Klassen minimiert wird, ist bei der latenten Profilanalyse die Durchführung eines Signifikanztests für den F-Wert angemessener.
2. Es können die paarweisen Unterschiede zwischen den Klassen in den Variablen berechnet werden.
3. Die Variablen innerhalb einer Klasse können zu Variablengruppen zusammengefasst werden.
4. Es können z-Werte zur Beantwortung der Frage, ob signifikante Abweichungen von den Gesamtmittelwerten vorliegen, berechnet werden.
5. Zur Beschreibung und inhaltlichen Validitätsprüfung können Deskriptionsvariablen in die Analyse einbezogen werden. Es ist aber auch möglich, die Klassenzugehörig-

keit abzuspeichern und die üblichen bivariaten und multivariaten Verfahren zur Validitätsprüfung einzusetzen (siehe Abschnitt 12.8 und 18.5).

Wir wollen hier nicht die einzelnen Schritte durchgehen, sondern die 6-Klassenlösung der latenten Profilanalyse mit der 6-Klassenlösung des K-Means-Verfahrens vergleichen. Die bei beiden Verfahren berechneten Klassenzentren und Klassengrößen enthält die Tabelle 14.6. Die Klassenlösungen stimmen hinsichtlich der Klassenzentren sehr gut überein:

Klasse C₁: Es handelt sich bei beiden Verfahren um ein Restcluster, das nur von wenigen Fällen gebildet wird. Es ist durch eine starke Ablehnung des Postmaterialismus gekennzeichnet (Anti-Postmaterialisten).

Klasse C₂: Auch dieses Cluster ist ein Restcluster, das nur aus wenigen Fällen besteht. Materialismus wird deutlich abgelehnt (Anti-Materialisten).

Klasse C₃: Sie könnte als Cluster der Postmaterialisten bezeichnet werden.

Klasse C₄: Diese Gruppierung lässt sich als Konsensotypus bezeichnen. Sie besitzt den größten Anteilswert (latente Profilanalyse: 38,7 Prozent, K-Means: 33,5 Prozent).

Klasse C₅: Die fünfte Klasse lässt sich als Klasse der extremen Postmaterialisten bezeichnen.

Klasse C₆: Sie entspricht dem Typus der Nichtorientierten. Keine Wertepräferenz ist sehr wichtig.

Bezüglich der Klassenanteilswerte treten etwas größere Unterschiede auf. So zum Beispiel besitzt die Klasse C₅ bei der latenten Profilanalyse einen Anteil von 15,8 Prozent, beim K-Means-Verfahren dagegen von 10,9 Prozent. Dies ist auf den unterschiedlichen Modellansatz zurückzuführen. Beim K-Means-Verfahren werden die Objekte deterministisch den Klassen zugeordnet. Besitzt beispielsweise ein Objekt die Zuordnungswahrscheinlichkeit 0,24 für die erste Klasse und die Zuordnungswahrscheinlichkeiten von 0,19 für die anderen 5 Klassen, wird es deterministisch der ersten Klasse zugeordnet (Kaufmann und Pape 1984, S. 449). Die beim K-Means-Verfahren berechneten Anteilswerte der Klassen vermitteln daher nur eine sehr grobe Vorstellung über die Größe der Klassen, wenn Überlappungen vorliegen. Bei beiden Verfahren sind zwei Klassen bzw. zwei Cluster schwach besetzt, nämlich die als Anti-Postmaterialisten und Anti-Materialisten bezeichneten Cluster.

14.4 Überlappungsindizes

Es wurde bereits darauf hingewiesen, dass der *Überlappungsanteil* entscheidend die Konvergenz und Stabilität der Ergebnisse der latenten Profilanalyse beeinflusst. Daher

Tab. 14.6: Ergebnisse der 6-Klassenlösung der latenten Profilanalyse und des K-Means-Verfahrens für die Wertedaten von Denz (50 multiple Zufallsstartwerte je Cluster)

	C_1	C_2	C_3	C_4	C_5	C_6
<i>Latente Profilanalyse</i>						
Anteilswert $p(k)$ in %	1,8	1,8	25,7	38,4	15,8	16,2
gMat (\bar{x})	2,54	4,50	2,34	1,60	2,42	2,10
gPmat (\bar{x})	3,49	1,92	1,48	1,55	1,09	2,17
gMat (s)	0,35	0,20	0,40	0,28	0,78	0,49
gPmat (s)	0,70	0,25	0,15	0,27	0,09	0,37
<i>K-Means-Verfahren</i>						
Anteilswert $p(k)$ in %	4,1	2,3	30,3	33,5	10,9	19,0
gMat (\bar{x})	2,48	4,30	2,29	1,52	3,12	1,72
gPmat (\bar{x})	3,00	2,00	1,48	1,39	1,21	2,05
gMat (s)	0,28	0,44	0,25	0,24	0,29	0,28
gPmat (s)	0,63	0,28	0,24	0,21	0,22	0,26

Abkürzungen: gMat: Gesamtpunktwert für Materialismus, gPmat: Gesamtpunktwert für Postmaterialismus

ist es bei der Modellprüfung sinnvoll, die Überlappungen zu überprüfen. Eine Grobabschätzung des Überlappungsanteils kann dadurch vorgenommen werden, dass die Zuordnungswahrscheinlichkeiten dichotomisiert und alle Ausprägungskombinationen berechnet werden. Als Dichotomisierungsschwelle kann man dabei $1/K$ wählen, also jenen Wert, der sich ergibt, wenn ein Objekt jeder Klasse mit der gleichen Wahrscheinlichkeit angehört.

Für die 5-Klassenlösung beispielsweise ergibt sich folgendes Bild (siehe Tabelle 14.7 auf der nächsten Seite): Insgesamt liegen mit 36 von 221 Befragten 16,3 Prozent in einem Überlappungsbereich. Sie bilden drei Mischtypen: $C_4 + C_5$, $C_2 + C_4$ und $C_2 + C_4 + C_5$. Der Überlappungsanteil ist somit nicht besonders hoch.

Eine genauere Abschätzung des Überlappungsanteils kann durch folgende Maßzahlen gewonnen werden, die im Rahmen von Verfahren zum *Fuzzy-Clustering* (Jain und Dubes 1988, S. 132–133, Kaufman und Rousseeuw 1990, S. 171) entwickelt wurden:

$$\text{DUNN}_K = \left(\frac{1}{n} \sum_g \sum_k p^2(k|g) - \frac{1}{K} \right) / \left(1 - \frac{1}{K} \right),$$

$$\text{BACKER}_K = 1 - \frac{1}{n} \cdot \frac{2}{K-1} \cdot \sum_g \sum_k \sum_{k^* \neq k} \min(p(k|g), p(k^*|g)).$$

Diese beiden Maßzahlen wurden von Dunn (1976, zitiert in Kaufman und Rousseeuw 1990, S. 171) bzw. Backer (1978, zitiert in Jain und Dubes 1988, S. 132–133) entwickelt und

Tab. 14.7: *Dichotome Zuordnung der Objekte zu den Clustern*

C_1	C_2	C_3	C_4	C_5	n
o	o	o	o	1	19
o	o	o	1	o	142
o	o	o	1	1	15
o	o	1	o	o	5
o	o	1	o	1	0
o	o	1	1	o	0
o	o	1	1	1	0
o	1	o	o	o	17
o	1	o	o	1	0
o	1	o	1	o	20
o	1	o	1	1	1
o	1	1	o	o	0
o	1	1	o	1	0
o	1	1	1	o	0
o	1	1	1	1	0
1	o	o	o	o	2
1	o	o	o	1	0
1	o	o	1	o	0
1	o	o	1	1	0
1	o	1	o	o	0
1	o	1	o	1	0
1	o	1	1	o	0
1	o	1	1	1	0
1	1	o	o	o	0
1	1	o	o	1	0
1	1	o	1	o	0
1	1	1	o	o	0
1	1	1	o	1	0
1	1	1	1	o	0
1	1	1	1	1	0
gesamt					221

Abkürzung: o: Objekt gehört nicht dem Cluster an, 1: Objekt gehört dem Cluster an.

grau hinterlegt: Überlappung

sind zwischen 0 und 1 normiert. Sie werden auch als *Partitions-Indexwerte* bezeichnet. Sie sind gleich 1, wenn keine Überlappungen vorliegen. Das heißt, jedes Objekt gehört dann mit einer Wahrscheinlichkeit von 1 nur einer Klasse an. Der Wert 0 tritt auf, wenn jedes Objekt mit derselben Wahrscheinlichkeit jedem Cluster angehört, bei K Klassen also mit einer Wahrscheinlichkeit von $1/K$.

Tabelle 14.8 auf der nächsten Seite verdeutlicht die Eigenschaften der beiden Maßzahlen im folgenden (fiktiven) Beispiel: In der Lösung Klassifikation I gehört jedes Objekt mit einer Wahrscheinlichkeit von 1 einer der drei Klassen an. Die Indexwerte sind daher gleich 1. In der Klassifikation II gehört jedes Objekt mit einer Wahrscheinlichkeit von $1/3$ einer Klasse an. Auf der empirischen Ebene liegt also keine Klassenstruktur vor. Die Indexwerte sind gleich 0. In der Klassifikation III schließlich sind die Objekte A und B eindeutig einer Klasse zuzuordnen, die Objekte C und D dagegen nicht. Sie gehören wiederum mit einer Wahrscheinlichkeit von $1/3$ allen drei Klassen an. Die beiden Indexwerte sind 0,5.

Für das Beispiel von Denz ergeben sich folgende Indexwerte: $DUNN_2 = 0,628$, $DUNN_5 = 0,659$ und $DUNN_6 = 0,674$ sowie $BACKER_2 = 0,760$, $BACKER_5 = 0,887$ und $BACKER_6 = 0,912$. Schwellenwerte zur Interpretation fehlen leider. Eine Studie von Trauwaert (1988) erbringt das Ergebnis, dass der DUNN-Index nur bedingt zur Auswahl der Cluster geeignet ist. Charrad, Lechevallier u. a. (2010) zeigen dagegen, dass für die *Block-Clustering-Methode*⁸ der Index zur Auffinden der Clusterzahl geeignet ist, wenn die Zahl der Cluster für die Zeilen- und Spaltenvariablen gleich ist. Bei ungleicher Clusterzahl ist seine Performanz schlecht. Da unterschiedliche Befunde vorliegen, empfehlen wir, die beiden Maßzahlen ergänzend zu verwenden.

Allgemein lässt sich festhalten: Je größer der Überlappungsanteil ist, desto stärker weichen die Ergebnisse des K-Means-Verfahrens und der latenten Profilanalyse ab. Ab einem bestimmten Überlappungsanteil – genaue Grenzwerte lassen sich nicht angeben – besteht die Gefahr, dass die Ergebnisse der latenten Profilanalyse instabil sind.⁹ In diesem Fall sind Stabilitätsuntersuchungen durchzuführen. Bei großen Stichproben (etwa $n > 2\,000$) kann man dabei so vorgehen, dass man Stichproben unterschiedlicher Größe (zum Beispiel $n = 500$, $n = 750$, $n = 1\,000$) zieht und das Konvergenzverhalten empirisch untersucht.

⁸ Block-Clustering (Mirkin 1996) bezeichnet ein Verfahren, bei dem die Zeilen und Spalten einer Datenmatrix (meistens eine bivariate Tabelle) simultan zu Klassen zusammengefasst werden.

⁹ Unsere bisherigen Erfahrungen legen einen Schwellenwert von 0,7 für den Partitions-Index von Backer nahe.

Tab. 14.8: Veranschaulichung der Bedeutung der Partitions-Indexwerte von Dunn und Backer

Zuordnungs-wahrscheinlichkeiten			kleinster Wert der Zuordnungswahr. in den Clusterpaaren			quadrierte Zuordnungswahrscheinlichkeiten					
C_1	C_2	C_3	C_1, C_2	C_1, C_3	C_2, C_3	Σ	C_1	C_2	C_3	Σ	
<i>Klassifikation I</i>											
A	1	0	0	0	0	0	1	0	0	1	
B	0	1	0	0	0	0	0	1	0	1	
C	0	0	1	0	0	0	0	0	1	1	
D	0	0	1	0	0	0	0	0	1	1	
						Σ	0			Σ	4
											$\text{BACKER} = 1 - \frac{2}{4 \cdot (3-1)} \cdot 0 = 1$
											$\text{DUNN} = \frac{(1/4) \cdot 4 - 1/3}{1-1/3} = 1$
<i>Klassifikation II</i>											
A	1/3	1/3	1/3	1/3	1/3	1/3	1	1/9	1/9	1/9	1/3
B	1/3	1/3	1/3	1/3	1/3	1/3	1	1/9	1/9	1/9	1/3
C	1/3	1/3	1/3	1/3	1/3	1/3	1	1/9	1/9	1/9	1/3
D	1/3	1/3	1/3	1/3	1/3	1/3	1	1/9	1/9	1/9	1/3
						Σ	4			Σ	4/3
											$\text{BACKER} = 1 - \frac{2}{4 \cdot (3-1)} \cdot 4 = 0$
											$\text{DUNN} = \frac{(1/4)(4/3) - 1/3}{1-1/3} = 0$
<i>Klassifikation III</i>											
A	1	0	0	0	0	0	1	0	0	1	
B	0	1	0	0	0	0	0	1	0	1	
C	1/3	1/3	1/3	1/3	1/3	1/3	1	1/9	1/9	1/9	1/3
D	1/3	1/3	1/3	1/3	1/3	1/3	1	1/9	1/9	1/9	1/3
						Σ	2			Σ	8/3
											$\text{BACKER} = 1 - \frac{2}{4 \cdot (3-1)} \cdot 2 = 0,5$
											$\text{DUNN} = \frac{(1/4)(8/3) - 1/3}{1-1/3} = 0,5$

14.5 Überprüfung der Annahme der lokalen Unabhängigkeit

Neben der Inspektion des Überlappungsanteils sollte anhand der Ergebnisse auch geprüft werden, ob die Ergebnisse die *Annahme der lokalen Unabhängigkeit* erfüllen, ob also die Variablen innerhalb der Klassen unkorreliert sind. Dabei kann wie folgt vorgegangen werden: Die Klassenzuordnungswahrscheinlichkeiten $p(k|g)$ werden abgespeichert. Daran anschließend wird für jede Klasse K eine Varianz-Kovarianzmatrix W_K berechnet. Die Zuordnungswahrscheinlichkeiten $p(k|g)$ werden als Gewichte verwendet. Ist die Annahme der lokalen Unabhängigkeit erfüllt, sollten die Variablen unabhängig voneinander sein.

ander sein und die Varianz-Kovarianzmatrix einer Diagonalmatrix entsprechen. Diese Annahme kann mittels eines LQ-Tests (Fahrmeir und Hamerle 1984, S. 74–75) geprüft werden. Die Teststatistik ist definiert als

$$\text{LQ} = -n_K \cdot \log(\det(\mathbf{W}_K)) ,$$

wobei $\det(\mathbf{W}_K)$ die Determinante der Varianz-Kovarianz-Matrix ist. Die gewichtete Fallzahl, auf der die Berechnung der Varianz-Kovarianzmatrix basiert, wird mit n_K bezeichnet. Der Test wird für jede latente Klasse k durchgeführt. Eine andere Möglichkeit der Prüfung der lokalen Unabhängigkeit ist die Berechnung von Modifikationsindizes (siehe dazu Abschnitt 16.4.5).

14.6 Konfirmatorische latente Profilanalyse

Wie bei den K-Means-Verfahren können bei der Analyse latenter Profile *bestimmte Parameter fixiert oder gleichgesetzt werden*.¹⁰ Das Computerprogramm ALMO (siehe Abschnitt 1.10) ermöglicht folgende Restriktionen:

1. Bestimmte Klassenanteilswerte können fixiert werden.
2. Bestimmte Klassenanteilswerte können gleich gesetzt werden – (spezielle) lineare Restriktionen zwischen den Klassenanteilswerten.
3. Bestimmte Klassenzentren können fixiert werden.
4. Bestimmte Klassenzentren können gleichgesetzt werden – (spezielle) lineare Restriktionen zwischen den Klassenzentren bzw. Klassenmittelwerten.

Eine Fixierung von Klassenstreuungen sowie lineare Restriktionen bezüglich der Klassenstreuungen sind derzeit nicht möglich.

Bei *linearen Restriktionen von Anteilswerten* – hier der Spezialfall der Gleichsetzung von Anteilswerten – wird beispielsweise

$$p(k') = p(k'') = \dots$$

gesetzt. Die Schätzwerte berechnen sich dann folgendermaßen:

$$p(k^*) = \sum_{k \in k^*} \frac{p(k)}{r} ,$$

¹⁰ Fixierte Parameter werden während der Iteration nicht geändert.

Tab. 14.9: Theoretisch angenommene Klassenstrukturen für die latente Profilanalyse

Typus	Modell I		Modell II		Modell III	
	gMat	gPmat	gMat	gPmat	gMat	gPmat
APMAT	2,0	3,2	2,5	2,5	2,5	2,5
AMAT	3,2	2,0	3,2	2,0	1,5	2,5
PMAT	3,0	1,0	3,0	1,0	3,0	1,0
GPMAT	1,5	2,5	1,5	2,5	2,5	1,5
KON	1,5	1,5	1,5	1,5	1,5	1,5

Abkürzungen: APMAT: Anti-Postmaterialisten, AMAT: Anti-Materialisten, PMAT: Postmaterialisten, GPMAT: gemäßigte Postmaterialisten, KON: Konsenstypus, NO: Nicht-Orientierte, GMAT: gemäßigte Materialisten, gPmat: Postmaterialismus, gMat: Materialismus.

wobei über alle Anteilswerte $p(k)$ summiert wird, für die eine lineare Restriktion definiert ist. Die Zahl der Anteilswerte der linearen Restriktionen ist r . Bei einer Fixierung wird der Anteilswert während der Iteration nicht geändert.

Für die Menge der Klassenzentren $\bar{x}_{k^*j^*}$, für die eine lineare Restriktion definiert ist, ergeben sich folgende Schätzwerte:

$$\bar{x}_{k^*j^*} = \sum_{\substack{k \in k^* \\ j \in j^*}} \bar{x}_{kj} p(k) \Bigg/ \sum_{\substack{k \in k^* \\ j \in j^*}} p(k) .$$

Wir wollen prüfen, welches der in der Tabelle 14.9 wiedergegebenen Modelle den Daten besser angepasst ist: Fixiert man die Klassenzentren in allen drei Modellen, ergeben sich folgende Log-Likelihood-Funktionswerte: $LL(\text{Modell I}) = -335,078$, $LL(\text{Modell II}) = -331,205$ und $LL(\text{Modell III}) = -330,096$. Die Werte der Log-Likelihood-Funktion sind in unserem Beispiel unmittelbar vergleichbar, da in jedem Modell dieselbe Anzahl von Parametern (fünf Klassenanteilswerte und zehn Klassenstreuungen) zu schätzen ist. Die Log-Likelihood-Werte unterscheiden sich nicht besonders. Die beste Modellanpassung liegt für das Modell III vor.

Sind die Log-Likelihood-Werte nicht vergleichbar, sollte ein Informationsmaß für einen Vergleich der Modelle verwendet werden. Nichtvergleichbarkeit würde zum Beispiel vorliegen, wenn ein Modell aus sechs Klassen besteht und ein anderes nur aus fünf oder wenn in einem Modell alle Werte fixiert sind und in einem anderen nur bestimmte.

Für eine weitere Analyse wird man mitunter das Modell I eliminieren, da es eine leicht schlechtere Modellanpassung erbringt, und die beiden anderen Modelle weiter untersu-

Tab. 14.10: Spezifikationen für zwei Modelle

Typus	Modell IV			Modell V		
	$p(k)$	gMat	gPmat	$p(k)$	gMat	gPmat
NO	1	6	7	1	6	6
AMAT	2	8	9	2	0	7
PMAT	3	0	0	3	8	0
GPMAT	4	0	0	4	6	9
KON	5	0	0	5	9	9

Abkürzungen: siehe Tabelle 14.9

Anmerkung: zu der Bedeutung der ausgewiesenen Zahlenwerte siehe Text.

chen, indem bestimmte Modellannahmen abgeschwächt werden. Wir wollen hier zwei weitere Modelle untersuchen, deren Restriktionen in Tabelle 14.10 dargestellt sind:

Modell IV: Die drei untersuchten Modelle unterscheiden sich nur hinsichtlich der ersten beiden Klassen. Es ist daher naheliegend, die entsprechenden Klassenzentren für eine Schätzung freizugeben.

Modell V: Das Modell II wird wie folgt modifiziert:

1. Die Wichtigkeit der postmaterialistischen Items in der Klasse der gemäßigten Postmaterialisten soll gleich der Wichtigkeit der materialistischen und postmaterialistischen Werte der Klasse des Konsensstypus sein.
2. Die Ablehnung der materialistischen Items in der Klasse der gemäßigten Postmaterialisten soll gleich der Ablehnung beider Wertorientierungen in der Klasse der Nicht-Orientierten sein.
3. Die Zustimmung der Postmaterialisten zu den postmaterialistischen Items soll gleich 1,0 sein.
4. Die Ablehnung der materialistischen Items in der Gruppe der Anti-Materialisten soll gleich 3,2 sein.

Allgemein bedeutet eine Zahl von 0, dass der Parameter fixiert ist. Ist die Zahl größer 0, soll der entsprechende Parameter geschätzt werden. Im Modell IV sind also die Klassenanteilswerte sowie die Klassenzentren der beiden ersten Klassen zu schätzen. Im Modell V sind alle Parameter – mit Ausnahme des Klassenzentrums der zweiten Klasse (AMAT: Anti-Materialisten) in gMat (Materialismus) und jenes der dritten Klasse (PMAT: Postmaterialisten) in gPmat (Postmaterialismus) – zu schätzen. Ist bei mehreren Modellparametern dieselbe Zahl größer 0 angegeben, so besteht zwischen diesen eine lineare Restriktion. Das Modell IV enthält also keine linearen Restriktionen, das Modell V dagegen zwei (Zahlen 6 und 9). Die lineare Restriktion mit der Zahl 9 im Modell V ist

Tab. 14.11: Modellparameter für Modell v (modifiziertes Modell II)

	C_1	C_2	C_3	C_4	C_5
<i>Latente Profilanalyse</i>					
Anteilswert $p(k)$ in %	12,4	4,1	6,5	31,8	45,2
gMat (\bar{x})	2,29	3,20	3,16	2,29	1,54
gPmat (\bar{x})	2,29	1,94	1,00	1,54	1,54
gMat (s)	0,42	1,12	0,40	0,40	0,28
gPmat (s)	0,73	0,33	0,10	0,23	0,33

wie folgt zu lesen: Der Klassenmittelwert in der postmaterialistischen Wertedimension (gPmat) der Klasse der gemäßigten Postmaterialisten (GPMAT) soll gleich den beiden Klassenmittelwerten des Konsenstypus (KON) sein usw.

Schätzt man beide Modelle, ergeben sich folgende Log-Likelihood-Werte: $LL(\text{Modell IV}) = -324,744$ und $LL(\text{Modell V}) = -303,033$. Das Modell v erbringt somit die bessere Modellanpassung. Der Wert der Log-Likelihood-Funktion liegt nahe jenem des Modells mit frei variierenden Parametern ($LL = -292,559$). Auch die BIC-Werte unterscheiden sich nur gering ($BIC = 714,674$ für das Modell mit frei variierenden Parametern und $BIC = 708,631$ für Modell v). Die geschätzten Werte sind in der Tabelle 14.11 dargestellt. Bei der Interpretation des Clusters C_2 (Anti-Materialisten) ist Vorsicht angebracht, da eine hohe Standardabweichung vorliegt. Mitunter wird man daher die Suche nach einem geeigneteren Modell fortsetzen. So zum Beispiel können bei linearen Restriktionen auch für das Modell III gute Ergebnisse erzielt werden. Letztlich ist es Aufgabe einer inhaltlichen Validitätsprüfung zu entscheiden, welches Modell zur Beschreibung und »Erklärung« der Daten angemessener ist. Mitunter werden dazu weitere Datenerhebungen erforderlich sein, indem man beispielsweise Jugendliche qualitativ befragt, um die mit den einzelnen Typen verbundenen Tiefenstrukturen herauszuarbeiten. Bei der Auswahl dieser Jugendlichen kann man sich an den sozialstrukturellen Charakteristiken der Typen orientieren, die berechnet werden können, indem die erhobenen sozialstrukturellen Variablen als Deskriptionsvariablen in die Analyse einbezogen werden.

Zu Anwendungsempfehlungen für dieses Kapitel verweisen wir auf die Empfehlungen in Abschnitt 16.4.6.

15 Analyse latenter Klassen für nominale, ordinale und gemischtskalierte Variablen

15.1 Modellansatz und Algorithmus für nominale Variablen

Im Unterschied zur latenten Profilanalyse setzt die *Analyse latenter Klassen für nominalskalierte Variablen* – wie ihr Name sagt – nur nominalskalierte Variablen voraus. Die Modellannahmen sind:

1. Es liegen K latente Klassen (Muster) mit den Mittelwerten $\pi(k)$ vor.
2. Die Wahrscheinlichkeit, dass in der latenten Klasse k die nominale Variable j mit der Ausprägung i auftritt, ist gleich $\pi(i(j)|k)$.
3. Wegen der Annahme der lokalen Unabhängigkeit ist die Auftrittswahrscheinlichkeit des Merkmalsvektors eines Objekts g , wenn die Klasse k vorliegt, gleich:

$$\pi(g|k) = \pi(x_{g1}|k) \cdot \pi(x_{g2}|k) \cdot \dots \cdot \pi(x_{gm}|k) = \prod_{j=1}^m \pi(x_{gj}|k), \quad (15.1)$$

wobei x_{gj} der Wert des Objekts g in der nominalen Variablen j und m die Anzahl der nominalen Variablen ist. Die bedingte Wahrscheinlichkeit des Auftretens der Ausprägung des Objekts g in der Variablen j für die Klasse k wird mit $\pi(x_{gj}|k)$ bezeichnet.

Die Modellparameter sind:

1. die Anteilswerte $\pi(k)$ der latenten Klassen $k (k = 1, \dots, K)$ und
2. die bedingten Auftrittswahrscheinlichkeiten $\pi(j(i)|k)$ für das Auftreten der Ausprägungen i in den nominalen Variablen j in den latenten Klassen k .

Die *bedingten Auftrittswahrscheinlichkeiten* entsprechen den Klassenzentren der latenten Profilanalyse. Im Unterschied zur latenten Profilanalyse stellen die Klassenvarianzen keine Modellparameter dar, da sie mit $\pi(j(i)|k) \cdot (1 - \pi(j(i)|k))$ aus den bedingten Auftrittswahrscheinlichkeiten berechnet werden können. Die Schätzung der Modellparameter erfolgt wiederum über den EM-Algorithmus. Dazu werden die nominalen Variablen in ihre Dummies aufgelöst. Wir wollen mit $x_{gj(i)}$ den Wert des Objekts g in der i -ten Dummy-Variablen (Ausprägung) der nominalen Variablen j bezeichnen. Es gilt $x_{gj(i)} = 1$, wenn Objekt g in der nominalen Variablen j die Ausprägung i besitzt, andernfalls gilt $x_{gj(i)} = 0$. Mit $\pi(k|g)$ sollen wiederum die (wahren) Zuordnungswahrscheinlichkeiten der Objekte g zu den Klassen k und mit $p(k)$ und $p(j(i)|k)$ die Schätzwerte der gesuchten Modellparameter bezeichnet werden. Unter Verwendung dieser Notation ergeben sich folgende Schätzgleichungen (Van de Pol und de Leeuw 1986):¹

$$p(k) = \frac{\sum_g \pi(k|g)}{n}, \quad (15.2)$$

$$p(j(i)|k) = \frac{\sum_g \pi(k|g) \cdot x_{gj(i)}}{\sum_g \pi(k|g)}. \quad (15.3)$$

Die beiden Schätzgleichungen 15.2 und 15.3 sind vollkommen strukturgleich mit jenen der latenten Profilanalyse (Gleichungen 14.8 und 14.9 auf Seite 358). Die Auftrittswahrscheinlichkeiten $\pi(j(i)|k)$ sind die Klassenzentren der Dummies der nominalen Variablen. Im Unterschied zur latenten Profilanalyse werden die Wahrscheinlichkeiten $\pi(x_{gj}|k)$ des Auftretens der Werte des Objektes g in den Variablen j , wenn die Klasse k vorliegt, nicht über die Dichtefunktion der Normalverteilung berechnet, sondern mit:

$$\pi(x_{gj}|k) = \pi(j(i)|k) \quad \text{für } x_{gj(i)} = 1. \quad (15.4)$$

Die Wahrscheinlichkeit $\pi(x_{gj}|k)$ ist also gleich der Auftrittswahrscheinlichkeit der Ausprägung i der nominalen Variablen j , die das Objekt besitzt. Hat beispielsweise das Objekt g in der nominalen Variablen j die Ausprägung 1, so ist die Wahrscheinlichkeit $\pi(x_{g1}|k)$ gleich der Auftrittswahrscheinlichkeit der Ausprägung 1 in der Klasse k , also gleich $\pi(j(1)|k)$. Das zugrunde liegende Verteilungsmodell ist das einer Poly- bzw. Multinomialverteilung (Fisz 1980, S. 195–197). Die Zuordnungswahrscheinlichkeiten $\pi(k|g)$ werden wie bei der latenten Profilanalyse über den Satz von Bayes bestimmt.

Der EM-Algorithmus sieht folgendermaßen aus:

¹ Van de Pol und de Leeuw (1986) entwickeln die Schätzgleichungen für das allgemeine latente Markovmodell (siehe dazu auch Langeheine und Van de Pol 1990). Dieses enthält als Submodell die Analyse latenter Klassen für nominalskalierte Variablen. Die Schätzgleichungen für die Analyse latenter Klassen sind beispielsweise auch in Andersen (1991, S. 426–429) wiedergegeben.

Schritt 1: Berechnung oder Eingabe von Startwerten für die Modellparameter oder für die Zuordnungswahrscheinlichkeiten. Bei Startwerten für die Zuordnungswahrscheinlichkeiten gehe zu Schritt 3.

Schritt 2: Berechnung der Zuordnungswahrscheinlichkeiten nach Bayes mit²

$$p(k|g) = \frac{p(k) \cdot p(g|k)}{\sum p(k) \cdot p(g|k)},$$

wobei die Wahrscheinlichkeit $p(g|k)$ des Auftretens des Objekts g in der Klasse k entsprechend der Gleichung 15.1 auf Seite 377 geschätzt wird. Die dafür benötigten Auftrittswahrscheinlichkeiten $\pi(x_{gj}|k)$ werden entsprechend Gleichung 15.4 berechnet.

Schritt 3: Schätzung der Modellparameter entsprechend den Gleichungen 15.2 und 15.3. Für Zuordnungswahrscheinlichkeiten $\pi(k|g)$ werden die Schätzwerte $p(k|g)$ aus dem zweiten Schritt verwendet.

Schritt 4: Prüfung der Konvergenz analog der latenten Profilanalyse.

In die Schätzung gehen folgende Nebenbedingungen ein:

$$\sum_k \pi(k) = 1, \quad (15.5)$$

$$\pi(k) > 0, \quad (15.6)$$

$$\sum_i \pi(j(i)|k) = 1 \quad \forall \text{ Variablen } i \text{ und Klassen } k. \quad (15.7)$$

Die beiden ersten Nebenbedingungen haben wir bereits bei der latenten Profilanalyse eingeführt. Sie besagen, dass die Summe der Klassenanteilswerte gleich 1 und keine Klasse leer sein soll. Die dritte Nebenbedingung bedeutet, dass die Summe der Auftrittswahrscheinlichkeiten der Ausprägungen einer Variablen in einer Klasse k gleich 1 sein soll. In Tabelle 15.1 auf der nächsten Seite ist ein Berechnungsschema für die Anzahl der zu schätzenden Parametern bei K Klassen und m nominalen Variablen unter Berücksichtigung der Nebenbedingungen angegeben.

Wird für drei nominale Variablen mit jeweils drei Ausprägungen eine 2-Klassenlösung gesucht, sind $2 \cdot ((3+3+3)-3+1) - 1 = 13$ Modellparameter zu schätzen. Diesen stehen $3 \cdot 3 \cdot 3 = 27$ unterschiedliche Datenvektoren gegenüber. Die notwendige Bedingung für ein identifiziertes Modell, wonach mehr oder zumindest gleichviele empirische Informationen (unterschiedliche Merkmalsvektoren) als Modellparameter vorliegen, ist somit erfüllt. Die allgemeine Bedingung lautet: Die Zahl der Ausprägungskombinationen der nominalen Variablen muss größer/gleich der Zahl der zu schätzenden Parameter sein.

² Aus Gründen der einfacheren Schreibweise wurde auf den Index für die i -te Iteration verzichtet.

Tab. 15.1: Berechnungsschema für die Anzahl zu schätzender Parameter bei der Analyse latenter Klassen mit nominalskalierten Variablen

$K - 1$	Anzahl der Klassenanteilswerte $\pi(k)$. Wegen der Nebenbedingung $\sum \pi(k) = 1$ ist ein Klassenanteilswert fixiert.
$K \cdot \sum_{j=1}^m (m_j - 1)$	Anzahl der Auftrittswahrscheinlichkeiten $\pi(j(i) k)$. Wegen der Nebenbedingung 15.7 sind in jeder Klasse für jede Variable $m_j - 1$ Auftrittswahrscheinlichkeiten zu schätzen (m_j : Zahl der Ausprägungen der Variablen j), insgesamt also die angegebene Zahl.
$K \cdot \left(\sum_j m_j - m + 1 \right) - 1$	Gesamtzahl zu schätzender Parameter für K Klassen und m Variablen.

Sind nicht alle Ausprägungskombinationen besetzt, kann das Modell empirisch nicht identifiziert sein. In unserem Beispiel wäre dies der Fall, wenn nur 12 der 27 möglichen Kombinationen empirisch besetzt wären. In diesem Fall wäre eine Schätzung sinnlos. Vor einer Analyse sollte daher immer geprüft werden, ob die notwendige Identifikationsbedingung erfüllt ist. Praktisch kann diese Bedingung zum Beispiel mit dem K-Means-Verfahren überprüft werden. Dazu wird die Clusterzahl gleich der Zahl zu schätzender Parameter gesetzt. Wir bezeichnen diese Zahl mit m_K . Findet das K-Means-Verfahren keine m_K Cluster, ist die notwendige Bedingung der Modellidentifikation nicht erfüllt.

Wir wollen den Algorithmus anhand eines Rechenbeispiels veranschaulichen: Gegeben seien zwei nominalskalierte Variablen X_1 und X_2 mit jeweils drei Ausprägungen. Es soll eine 2-Klassenlösung berechnet werden. Vor dem ersten Iterationsschritt sollen Schätzwerte wie in Tabelle 15.2 vorliegen. Die Interpretation der Schätzwerte soll exemplarisch für die latente Klasse 1 (C_1) dargestellt werden. Die latente Klasse 1 ($k = 1$) hat einen Anteilswert von 0,5. Die Ausprägung 1 der nominalen Variablen 1 tritt mit einer Wahrscheinlichkeit von 0,8 ($p(1(1)|1)$ auf, die Ausprägung 2 der nominalen Variablen 1 mit einer Wahrscheinlichkeit von 0,2 und die Ausprägung 3 mit einer Wahrscheinlichkeit

Tab. 15.2: Schätzwerte vor dem ersten Iterationsschritt

	C_1 ($k = 1$)	C_2 ($k = 2$)
$p(k)$	0,5	0,5
$p(1(1) k)$	0,8	{= 1}
$p(1(2) k)$	0,2	0,0
$p(1(3) k)$	0,0	0,2
$p(2(1) k)$	0,6	{= 1}
$p(2(2) k)$	0,2	0,1
$p(2(3) k)$	0,2	0,7

von 0. Da die Auftrittswahrscheinlichkeiten als Mittelwerte der Dummies (Anteilswerte) interpretiert werden können, ist auch folgende Interpretation möglich: In der Klasse 1 tritt mit 80 Prozent die Ausprägung 1 in der Variablen 1 auf und mit 20 Prozent die Ausprägung 2. Für die zweite Variable ist die Auftrittswahrscheinlichkeit der Ausprägung 1 gleich 0,6 ($p(2(1)|1)$) usw. An den Schätzwerten lassen sich auch die dargestellten Nebenbedingungen veranschaulichen: Die Summe der Anteilswerte der beiden Klassen ist gleich 1. Da beide Klassen einen Anteilswert von 0,5 haben, ist auch die zweite Nebenbedingung, nach der keine Klasse leer sein darf, erfüllt. Die dritte Nebenbedingung besagt, dass die Summe der Auftrittswahrscheinlichkeiten jeder Variablen in jeder Klasse gleich 1 ist. Auch diese Nebenbedingung ist erfüllt. Für die nominale Variable 1 in der Klasse 1 gilt – wie man leicht nachrechnen kann – beispielsweise:

$$p(1(1)|1) + p(1(2)|1) + p(1(3)|1) = 0,8 + 0,2 + 0 = 1 .$$

Die Berechnung der Zuordnungswahrscheinlichkeiten $p(k|g)$ zeigt Tabelle 15.3. Sie soll für das erste Objekt A dargestellt werden. Objekt A besitzt in der Variablen X_1 die Ausprägung 1 und in der Variablen X_2 ebenfalls die Ausprägung 1. Die Wahrscheinlichkeit für das Auftreten dieses Antwortmusters in der ersten latenten Klasse ist gleich $0,8 \cdot 0,6 = 0,48$, da $p(1(1)|1) = 0,8$ und $p(2(1)|1) = 0,6$. Die Auftrittswahrscheinlichkeit des Objekts A für die latente Klasse 2 ist gleich $0 \cdot 0,1 = 0$. Der Likelihood-Wert p_{ges} des Objekts A ist – wie bei der latenten Profilanalyse – gleich $0,24 + 0,00 = 0,24$, da $p(1) = 0,5$, $p(2) = 0,5$ und $p(A|1) = 0,48$ und $p(A|2) = 0$. Daraus ergibt sich ein Log-Likelihood-Wert von $-1,4271$. Die Zuordnungswahrscheinlichkeiten für das Objekt A lassen sich über den Satz von Bayes berechnen. Für $p(1|A)$ ergibt sich ein Wert von

$$p(1|A) = \frac{0,5 \cdot 0,48}{0,5 \cdot 0,48 + 0,5 \cdot 0} = 1 ,$$

da $p(A|1) = 0,48$ und $p(A|2) = 0$ ist. Die Zuordnungswahrscheinlichkeit des Objekts A zur Klasse 1 ist gleich 1, jene zur Klasse 2 gleich 0. Das Objekt A wird also mit einer Wahrscheinlichkeit von 1 der Klasse 1 zugeordnet.

Für die anderen Objekte können die Zuordnungswahrscheinlichkeiten analog berechnet werden. Es ergeben sich die in der Tabelle 15.3 auf der nächsten Seite dargestellten Werte. Zur Neuberechnung der Modellparameter werden die nominalen Variablen in ihre Dummies aufgelöst und mit den entsprechenden Zuordnungswahrscheinlichkeiten multipliziert. Da beide Variablen drei Ausprägungen haben, wird jede nominale Variable in drei Dummies aufgelöst. Tabelle 15.4 auf Seite 383 zeigt exemplarisch die Berechnung der Modellparameter für die erste nominale Variable in der ersten Klasse. Da das Objekt A in der nominalen Variablen 1 die Ausprägung 1 besitzt, ist der Wert in der Dummy-Variablen $X_{1(1)}$ gleich 1. Dieser wird mit der Zuordnungswahrscheinlichkeit des Objekts A zur ersten Klasse ($k = 1$) multipliziert. Analog wird für die anderen Dummies und die

Tab. 15.3: Veranschaulichung der Berechnung der Zuordnungswahrscheinlichkeiten für die Analyse latenter Klassen für nominalskalierte Variablen

Objekt	Variablen	Auftritts-wahrscheinlichkeit der Objekte		Likelihood-Wert	Log-Likelihood-Wert	Zuordnungs-wahrscheinlichkeit	
		X_1	X_2			$p(1 g)$	$p(2 g)$
g		$p(g 1)$	$p(g 2)$	p_{ges}	$\log p_{\text{ges}}$		
A	1	0,48	0,00	0,24	-1,4271	1,00	0,00
B	1	0,16	0,02	0,16	-1,8326	1,00	0,00
C	2	0,04	0,04	0,04	-3,2189	0,50	0,50
D	3	0,00	0,56	0,28	-1,2730	0,00	1,00
E	3	0,00	0,16	0,08	-2,5257	0,00	1,00
				Σ	-10,2773		

weiteren Objekte vorgegangen. Berechnet man die Spaltensumme der Zuordnungswahrscheinlichkeiten für die Klasse 1 (Spalte $p(g|1)$), ergibt sich entsprechend Gleichung 15.2 auf Seite 378 die mit den Zuordnungswahrscheinlichkeiten gewichtete Fallzahl der Klasse 1. Der Spaltensummenwert ist gleich 2,5. Division mit der Fallzahl ergibt den Anteilwert der Klasse 1. In unserem Beispiel ist dieser gleich 0,5. Bilden wir die Spaltensummen der mit den Zuordnungswahrscheinlichkeiten multiplizierten (gewichteten) Dummies und dividieren diese mit der gewichteten Fallzahl, erhalten wir entsprechend Gleichung 15.3 auf Seite 378 die Schätzwerte für die Auftrittswahrscheinlichkeiten der Ausprägungen der nominalen Variablen 1 in der Klasse 1. Es ergeben sich Werte von 0,8, 0,2 und 0,0. Sie sind also mit den der Iteration vorausgehenden Schätzwerten identisch. Dies gilt auch für die anderen Modellparameter.

Damit dürfte das Grundprinzip der Schätzung der Modellparameter verdeutlicht worden sein. *Abschließend ist auf eine Besonderheit des Algorithmus hinzuweisen:* Besitzt ein Schätzwert für eine Auftrittswahrscheinlichkeit $p(i(j)|k)$ den Wert 0 oder 1, wird er während der Iteration nicht mehr geändert. Bei der Eingabe von Startwerten ist also darauf zu achten, dass keine Werte von 0 oder 1 eingegeben werden. Neue Computerprogramme, wie Latent GOLD (siehe Kapitel 16) oder AutoClass (siehe Abschnitt 17.1) vermeiden dies, indem mittels Bayes-Technik Bestrafungsterme eingefügt werden.

15.2 Modellprüfung und Interpretation

Es können die für die latente Profilanalyse entwickelten *Modellprüfgrößen* verwendet werden. Ihre Anwendung soll anhand der Analyse der Freizeitaktivitäten von Kindern dargestellt werden (siehe Kapitel 4 und 10 sowie die Abschnitte 5.2 und 5.3).

Tab. 15.4: Neuberechnung der Modellparameter für die erste Variable in der ersten Klasse

g	Variablen		Zuordnungs-wahrscheinlichkeit		Dummy multipliziert mit Zuordnungswahrsch.			
	X ₁	X ₂	p(1 g)	p(2 g)	X ₁₍₁₎	X ₁₍₂₎	X ₁₍₃₎	usw.
A	1	1	1,00	0,00	1 · 1,00	0 · 1,00	0 · 1,00	
B	1	2	1,00	0,00	1 · 1,00	0 · 1,00	0 · 1,00	
C	2	2	0,50	0,50	0 · 0,50	1 · 0,50	0 · 0,50	
D	3	3	0,00	1,00	0 · 0,00	0 · 0,00	1 · 0,00	
E	3	2	0,00	1,00	0 · 0,00	0 · 0,00	1 · 0,00	
Σ			2,50	2,50	2,00	0,50	0,00	

$$p(1) = 2,50/5 = 0,50$$

$$p(1(1)|1) = 2,00/2,5 = 0,8$$

$$p(1(2)|1) = 0,50/2,5 = 0,2$$

$$p(1(3)|1) = 0,00/2,5 = 0,0$$

Für eine erste Analyse wurde angenommen, dass vier bis acht latente Klassen vorhanden sind. Die entsprechenden Testgrößen zeigt die Tabelle 15.5: Betrachten wir zunächst die prozentuellen Verbesserungen PV_K gegenüber der jeweils vorausgehenden Lösung, so zeigt sich, dass die 7- und 8-Klassenlösung kaum mehr den Wert der Log-Likelihood-Funktion der vorausgehenden Lösung verbessern: Die 7-Klassenlösung verbessert die 6-Klassenlösung nur mehr um 9,1 Prozent, die 8-Klassenlösung die 7-Klassenlösung um 7,7 Prozent (Spalte »PV_K«). Nach der 6-Klassenlösung tritt somit ein deutlicher Abfall auf, so dass dieses Kriterium eine 6-Klassenlösung nahelegt. Der kleinste Wert des Informationsmaßes von Akaike liegt allerdings für die 8-Klassenlösung vor. Die

Tab. 15.5: Ergebnisse der Analyse latenter Klassen für die Freizeitaktivitäten von Kindern (n = 2745)

K	LL _K	PV _{OK}	AIC _K	AIC _{3K}	PV _K
1 ^{a)}	-32 024,300	0,000	— ^{b)}	— ^{b)}	— ^{b)}
4	-30 166,865	5,800	60 599,73	60 084,73	— ^{c)}
5	-30 090,713	6,038	60 389,43	59 869,43	25,2
6	-29 984,312	6,370	60 218,62	59 593,62	35,4
7	-29 956,915	6,456	60 205,83	59 475,83	9,1
8	-29 933,790	6,528	60 201,58	59 366,58	7,7

a) ALMO berechnet aus Modelltestgründen immer automatisch die 1-Klassenlösung.

b) nicht definiert

c) Diese Werte werden nicht berechnet, da keine vorausgehende 3-Klassenlösung vorliegt.

Werte für die 6- und 7-Klassenlösungen sind aber nur geringfügig kleiner. Der AIC tendiert zudem zu einer Überschätzung der Modellanpassung und der Klassenzahl. BIC und CAIC führen zu Modellen mit einer geringeren Klassenzahl. In einer neuen Evaluationsstudie berichten Fonseca und Cardoso (2007), dass das Informationsmaß AIC₃ (*Akaike's Information Criterion 3*, Bozdogan 1993) bei nominalen Merkmalen die bekannten Cluster am besten wiederentdeckt. Bei quantitativ-kontinuierlichen Variablen schneidet BIC am besten ab. Bei gemischten Merkmalen erweist sich das Informationsmaß ICL-BIC (*Integrated Completed Likelihood Criterion*, Biernacki, Celeux u. a. 2000) als am besten geeignet. Der AIC₃ ist definiert als

$$\text{AIC}_3K = -2 \cdot \text{LL}_K + 3 \cdot m_K$$

und unterscheidet sich vom AIC dadurch, dass als Bestrafung nicht der Term $2 \cdot m_k$ sondern $3 \cdot m_k$ verwendet wird, mit m_k als Parameteranzahl. Der ICL-BIC ist definiert als

$$\text{ICL-BIC} = \text{BIC} + 2 \cdot \text{EN}(S) . \quad (15.8)$$

Als zusätzlicher Bestrafungsterm wird die Entropie der Klassenzuordnungswahrscheinlichkeiten $pi(k|g)$ der Objekte $\text{EN}(S) = -\sum_{g=1}^n \sum_{k=1}^K \pi(k|g) \log \pi(k|g)$ berücksichtigt.

In unserem Beispiel ergibt sich auch für AIC₃ ein Minimum bei acht Klassen (siehe Tabelle 15.5 auf der vorherigen Seite). Die größere Bestrafung von Modellen mit mehr Parametern durch die Multiplikation der Zahl der Modellparameter m_k mit dem Faktor 3 (bei AIC Faktor 2) – also von Modellen mit größeren Klassenzahl – führt zu keiner Reduktion der Klassenzahl. Allerdings rücken alle Werte näher zueinander, das heißt der Unterschied im Informationsmaß zwischen der 4-Klassenlösung und der 8-Klassenlösung ist für AIC₃ kleiner als für AIC.

Wir wollen nachfolgend die 6-Klassenlösung weiter beschreiben, da nach der 6-Klassenlösung ein deutlicher Abfall im PV_K-Koeffizient auftritt. Die Ergebnisse sind in der Tabelle 15.6 dargestellt. Da sehr viele Variablen untersucht werden, wird man zur Erleichterung der Interpretation untersuchen, ob in den einzelnen Klassen Variablengruppen gebildet werden können. Dabei ergeben sich die in Tabelle 15.7 auf Seite 386 dargestellten Variablengruppen.

Die letzten drei Klassen lassen sich relativ einfach interpretieren:

Klasse 4: Typus der sehr aktiven Kinder

Klasse 5: Typus der inaktiven oder deprivierten Kinder

Klasse 6: Typus der Mittelaktiven, wobei Fernsehen und Radfahren überzufällig häufig auftreten

Tab. 15.6: 6-Klassenlösung bei einer Analyse der Freizeitaktivitäten von Kindern ($n = 2745$)

		C_1	C_2	C_3	C_4	C_5	C_6
Anteilswert $p(k)$		0,283	0,188	0,157	0,063	0,095	0,213
Ausrufen	ja	0,55	0,18	0,36	0,72	0,08	0,32
Ausrufen	nein	0,45	0,82	0,64	0,28	0,92	0,68
Freunde	ja	0,93	0,79	0,89	0,92	0,49	0,66
Freunde	nein	0,07	0,21	0,11	0,08	0,51	0,34
Familie	ja	0,71	0,53	0,74	0,89	0,25	0,39
Familie	nein	0,29	0,47	0,26	0,11	0,75	0,61
Basteln	ja	0,67	0,33	0,35	0,91	0,06	0,21
Basteln	nein	0,33	0,67	0,65	0,09	0,94	0,79
Comics	ja	0,61	0,21	0,56	0,86	0,11	0,47
Comics	nein	0,39	0,79	0,44	0,14	0,89	0,53
Musizieren	ja	0,38	0,36	0,20	0,63	0,06	0,06
Musizieren	nein	0,62	0,64	0,80	0,37	0,94	0,94
Haustiere	ja	0,61	0,61	0,61	0,78	0,19	0,48
Haustiere	nein	0,39	0,39	0,39	0,22	0,81	0,52
Kino	ja	0,04	0,04	0,21	0,51	0,05	0,08
Kino	nein	0,96	0,96	0,79	0,49	0,95	0,92
Konzert	ja	0,18	0,17	0,22	0,56	0,05	0,04
Konzert	nein	0,82	0,83	0,78	0,44	0,95	0,96
Musikhören	ja	0,93	0,63	0,81	0,97	0,23	0,65
Musikhören	nein	0,07	0,37	0,19	0,03	0,77	0,35
Kirche	ja	0,55	0,66	0,38	0,73	0,27	0,29
Kirche	nein	0,45	0,34	0,62	0,27	0,73	0,71
Fernsehen	ja	0,84	0,47	0,93	0,97	0,25	0,84
Fernsehen	nein	0,16	0,53	0,07	0,03	0,75	0,16
Computersp.	ja	0,31	0,09	0,78	0,80	0,14	0,48
Computersp.	nein	0,69	0,91	0,22	0,20	0,86	0,52
Buch	ja	0,92	0,79	0,62	0,94	0,21	0,52
Buch	nein	0,08	0,21	0,38	0,06	0,79	0,48
Vereinsv.	ja	0,10	0,07	0,30	0,56	0,07	0,03
Vereinsv.	nein	0,90	0,93	0,70	0,44	0,93	0,97
Radfahren	ja	0,92	0,88	0,93	1,00	0,42	0,85
Radfahren	nein	0,08	0,12	0,07	0,00	0,58	0,15
Spazieren	ja	0,88	0,56	0,55	0,98	0,18	0,38
Spazieren	nein	0,12	0,44	0,45	0,02	0,82	0,62
alleine Sp.	ja	0,81	0,44	0,75	0,96	0,10	0,57
alleine Sp.	nein	0,19	0,56	0,25	0,04	0,90	0,43
Parties	ja	0,18	0,12	0,30	0,70	0,05	0,09
Parties	nein	0,82	0,88	0,70	0,30	0,95	0,91
Sport	ja	0,69	0,57	0,92	0,91	0,33	0,48
Sport	nein	0,31	0,43	0,08	0,09	0,67	0,52

Tab. 15.7: Variablengruppen in den sechs Klassen

Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5	Klasse 6
Ausruhen, Comics, Haustiere, Kirche (0,58)	Basteln, Fern- sehen (0,41)	Musizie- ren, Kino, Konzert, Ver- einsv., Parties (0,27)	Ausruhen, Musizieren, Haustiere, Kino, Kon- zert, Kirche, Computer sp., Vereinsv., Parties (0,67)	Ausruhen, Basteln, Co- mics, Musi- zieren, Kino, Konzert, Computer sp., alleine Sp., Parties (0,09)	Buch, Sport, Comics, Haustiere, alleine Sp. (0,50)
Musizieren (0,42)	Familie, Haustiere, Musikhö- ren, Kirche, Spaziereng. (0,59)	Familie, Musikhören, Computer sp., alleine Sp. (0,73)	Freunde, Familie, Bas- teln, Comics, Musik, Fern- sehen, Buch, Radfahren, Spazieren, Sport (0,94)	Familie, Haustiere, Musik, Kir- che, Fernse- hen, Buch, Spazieren, Sport (0,28)	Freunde (0,67)
Freunde, Musikhö- ren, Buch, Radfahren, Spazieren (0,91)	Musizieren, Comics, Konzerte (0,19)	Comics, Haustiere, Buch, Spazie- ren (0,59)	Freunde, Radfahren (0,46)	Ausruhen, Kirche, Spa- zieren, Com- puter sp. (0,33)	
Vereinsv. (0,09)	Freunde, Buch (0,81)	Ausruhen, Basteln, Kir- che (0,41)			Musizie- ren, Kino, Konzert, Ver- einsv., Parties (0,05)
Computer sp. (0,30)	Computer sp., Vereinsv., Parties (0,10)	Freunde, Fernsehen, Radfahren, Sport (0,91)			Basteln (0,18)
Fernsehen, alleine Sp. (0,83)	Radfahren (0,90)				Fernsehen, Radfahren (0,82)
Konzert, Parties (0,19)	Kino (0,04)				
Kino (0,04)	alleine Sp. (0,44)				
	Sport (0,57)				

Die Klassen 1 bis 3 haben ein selektives Freizeitmuster:

Klasse 3: Typus von Kindern, für den das Spielen im Vordergrund steht. Diese Klasse ist zum einen stark spielorientiert (alleine Spielen, mit Freunden spielen, Computerspiele), zum anderen wird Musik gehört, mit der Familie etwas unternommen sowie Sport betrieben und Rad gefahren.

Klasse 2: Typus des zurückgezogenen Freizeittypus. Die Klasse ist dadurch gekennzeichnet, dass nur drei Freizeitaktivitäten häufig ausgeübt werden. Der Typus lässt sich wie folgt beschreiben: Die Kinder sind zum einen in der Wohnung sehr zurückgezogen. Sie lesen in der Freizeit ein Buch, im Freien spielen sie mit Freunden oder fahren Rad.

Klasse 1: Unter Umständen familienorientierter Typus (ein Name lässt sich nur schwer finden). Diese Klasse ist im Unterschied zur Klasse 2 durch eine stärkere Familienorientierung gekennzeichnet. Mit der Familie wird etwas gemeinsam unternommen, unter anderem ein Spaziergang, oder es wird gemeinsam gebastelt oder gemeinsam ferngesehen. Daneben wird noch ein Buch gelesen, Rad gefahren, Musik gehört, Sport betrieben und alleine oder gemeinsam mit Freunden gespielt.

Zusammenfassend lässt sich festhalten, dass zwar eine inhaltliche Interpretation möglich ist, dass man diese aber – vor allem in Hinblick auf die schlecht interpretierbare Klasse 1 – in einer konfirmatorischen Analyse noch zu verbessern versuchen wird.

15.3 Konfirmatorische Analyse

Wie bei der latenten Profilanalyse oder dem K-Means-Verfahren können auch in der Analyse latenter Klassen für nominalskalierte Variablen *bestimmte Parameter fixiert oder durch lineare Restriktionen verknüpft werden*. Für den Fall, dass sich die (speziellen) linearen Restriktionen (siehe Abschnitt 14.6) nur auf die Ausprägungen einer Variablen beziehen, ergeben sich die bisherigen Schätzwerte mit

$$p(j(i^*)|k) = \frac{\sum_{i \in i^*} p(j(i)|k)}{r},$$

wobei r die Zahl der Ausprägungen ist, für die eine lineare Restriktion definiert ist. Wegen der Nebenbedingung 15.5 auf Seite 379, nach der die Summe der bedingten Antwortwahrscheinlichkeiten gleich 1 sein soll, gestaltet sich die Behandlung linearer Restriktionen schwieriger, wenn sich diese auf unterschiedliche latente Klassen und/oder unterschiedliche Variablen beziehen. Der Algorithmus zur Schätzung der Modellparameter, für die

Tab. 15.8: Beispiel für proportionale Anpassung der Schätzwerte für die Auftrittswahrscheinlichkeiten

k	ohne Restriktion		mit Restriktion		mit Restriktion und Skalierung	
	1	2	1	2	1	2
$p(k)$	0,40	0,60	0,40	0,60	0,40	0,60
$p(1(1) k)$	0,60	0,80	0,72	0,72	0,72	0,72
$p(1(2) k)$	0,20	0,20	0,20	0,20	$0,20 \cdot 0,70 = 0,14$	$0,20 \cdot 1,40 = 0,28$
$p(1(3) k)$	0,20	0,00	0,20	0,00	$0,20 \cdot 0,70 = 0,14$	$0,00 \cdot 1,40 = 0,00$
$\sum p(1(i) k)$	1,00	1,00	1,12	0,92	1,00	1,00

eine lineare Restriktion definiert ist, sieht folgendermaßen aus, wobei die bedingten Auftrittswahrscheinlichkeiten, für die eine lineare Restriktion definiert ist, mit $p(j^*(i^*)|k^*)$ bezeichnet werden sollen:

Schritt 1: Die bedingten Auftrittswahrscheinlichkeiten werden zunächst ohne lineare Restriktionen geschätzt.

Schritt 2: Für die Modellparameter mit einer linearen Restriktion erfolgt die Schätzung mit

$$p(j^*(i^*)|k^*) = \frac{\sum_{\substack{i \in i^* \\ j \in j^* \\ k \in k^*}} (p(k) \cdot p(j(i)|k))}{\sum_{\substack{i \in i^* \\ j \in j^* \\ k \in k^*}} p(k)} .$$

Schritt 3: Nach der Schätzung der Modellparameter mit linearen Restriktionen findet eine Anpassung an die Nebenbedingung $\sum p(j(i)|k) = 1$ für alle j und k statt. Dabei wird wie folgt vorgegangen: Es wird die Summe $P_j = \sum_{i \in i^*} p(j(i)|k)$ der frei variierenden Auftrittswahrscheinlichkeiten für jede Variable j und jede Klasse k berechnet. Ferner wird die Summe $P_{jk}^* = 1 - \sum_{i \in i^*} p(j(i)|k)$ berechnet, also die verbleibende Summe nach Abzug der Auftrittswahrscheinlichkeiten mit linearen Restriktionen. Die frei variierenden Parameter $p(j(i)|k)$ werden mit Hilfe dieser beiden Größen mit $P_{jk}^* \cdot p(j(i)|k)/P_j$ proportional angepasst.

Wir wollen uns diese proportionale Anpassung anhand eines einfachen Beispiels verdeutlichen, das in Tabelle 15.8 wiedergegeben ist: Für die Variable $j = 1$ wird die lineare Restriktion getroffen, dass die erste Ausprägung in den Klassen 1 und 2 dieselbe Auftrittswahrscheinlichkeit besitzen soll. Die Schätzwerte ohne lineare Restriktionen finden sich links in der Tabelle. Da für $p(1(1)|1)$ und $p(1(1)|2)$ eine lineare Restriktion definiert ist, gilt für die Schätzwerte der Auftrittswahrscheinlichkeiten

$$p(1(1)|1) = p(1(1)|2) = \frac{0,6 \cdot 0,4 + 0,8 \cdot 0,6}{0,4 + 0,6} = 0,72 .$$

Tab. 15.9: Für eine konfirmatorische Analyse spezifizierte Freizeittypen

Freizeittypus	Nr.	Charakteristik
sehr aktive Kinder	I	Alle Freizeitaktivitäten treten mit einer Wahrscheinlichkeit von 0,9 auf.
mittel aktive Kinder	II	Alle Freizeitaktivitäten treten mit einer Wahrscheinlichkeit von 0,5 auf.
inaktive Kinder	III	Alle Freizeitaktivitäten treten mit einer Wahrscheinlichkeit von 0,1 auf.
medialer Freizeittypus	IV	Musikhören, Computerspielen, Buchlektüre und Fernsehen treten mit einer Wahrscheinlichkeit von 1,0 auf.
verwalteter Freizeittypus	V	Vereinsveranstaltungen und der Besuch von Parties treten mit einer Wahrscheinlichkeit von 1,0 auf.
kontrollierter Freizeittypus	VI	Pädagogisch für wenig wertvoll gehaltene Freizeitaktivitäten (Comics, Computerspiele, Fernsehen) treten mit einer Wahrscheinlichkeit von 0,1 auf, das Lesen eines Buches mit 0,9.
zurückgezogener Freizeittypus	VII	Musikhören und Buchlektüre tritt mit einer Wahrscheinlichkeit von 0,9 auf.
kultureller Freizeittypus	VIII	Konzert-, Theater- oder Ausstellungsbesuch sowie ein Kinobesuch, Buchlektüre, Musizieren und Musikhören treten mit einer Wahrscheinlichkeit von 0,9 auf.

Die Nebenbedingung $\sum p(j(i)|k) = 1$ ist somit nicht erfüllt. Wir berechnen nun die beiden Summenwerte P_{jk} und P_{jk}^* für beide latente Klassen. Es gilt: $P_{11} = 0,2 + 0,2 = 0,4$, da die Ausprägungen 2 und 3 frei variieren können, und $P_{11}^* = 1 - 0,72 = 0,28$, da für die erste Ausprägung eine lineare Restriktion (0,72) definiert ist. Der Skalierungsfaktor für die frei variierbaren Parameter ist daher für die erste Klasse $P_{11}^*/P_{11} = 0,28/0,40 = 0,70$. Für die zweite Klasse ergibt sich ein Skalierungsfaktor von 1,4. Die Auftrittswahrscheinlichkeiten sind rechts in der Tabelle 15.8 dargestellt und erfüllen die Nebenbedingung 15.5.

Das hier dargestellte Vorgehen setzt voraus, dass die Summenwerte P_{jk}^* größer 0 sind und dass mindestens eine Auftrittswahrscheinlichkeit frei variieren kann. Mitunter kann P_{jk}^* kleiner 0 werden. Dies stellt einen Entartungsfall dar und bedeutet, dass die getroffenen Modellannahmen nicht zutreffen.

Wir wollen nun die Technik der konfirmatorischen Analyse anwenden, um für das Beispiel der Freizeitaktivitäten eine bessere Interpretation zu erhalten. Dazu werden die Typen aus Tabelle 15.9 angenommen. Die Typen sind nur partiell definiert. Die zur Definition verwendeten Variablen werden in der Analyse fixiert, alle anderen können

frei variieren. Zu Testzwecken wird ferner eine Analyse gerechnet, in der alle Parameter frei variieren können. Das Modell mit fixierten Parametern erbringt eine schlechtere Modellanpassung ($\text{LL} = -29\,923,688$, $\text{AIC} = 60\,181,376$, $\text{AIC}_3 = 60\,348,376$) als das Modell mit frei variierenden Parametern ($\text{LL} = -30\,856,183$, $\text{AIC} = 61\,732,366$, $\text{AIC}_3 = 61\,742,366$). Um Anhaltspunkte für Modellmodifikationen zu gewinnen, wird man die beiden Modelle gegenüberstellen und/oder das Modell mit frei variierenden Parametern interpretieren und mögliche Restriktionen ins Auge fassen. Wir haben mehrere inhaltlich plausible Modelle durchprobiert, aber kein einfach interpretierbares Modell mit einer guten Modellanpassung gefunden.

15.4 Modellansatz und Algorithmus für ordinale und gemischtskalierte Variablen

Die Schätzgleichungen der Modellparameter für die bisher behandelten Verfahren sind – wie wir gesehen haben – vollkommen strukturgleich, wenn bei nominalen Variablen mit Dummies gerechnet wird. Dies legt nahe, ein *allgemeines probabilistisches Messmodell für gemischte Variablen* zu definieren (Bacher 2000), in dem die Zuordnungswahrscheinlichkeiten wie folgt berechnet werden:

$$\pi(g|k) = \frac{\pi(k) \cdot \pi(g|k)^{\text{nom}} \cdot \pi(g|k)^{\text{ord}} \cdot \pi(g|k)^{\text{quant}}}{\sum_{k^*} \pi(k^*) \cdot \pi(g|k^*)^{\text{nom}} \cdot \pi(g|k^*)^{\text{ord}} \cdot \pi(g|k^*)^{\text{quant}}}. \quad (15.9)$$

Dabei ist:

$\pi(g|k)^{\text{nom}}$ Auftrittswahrscheinlichkeiten der Ausprägungen des Objekts g in den nominalen Variablen für die Klasse k . Diese werden entsprechend den Gleichungen 15.1 auf Seite 377 und 15.4 auf Seite 378 berechnet.

$\pi(g|k)^{\text{ord}}$ Auftrittswahrscheinlichkeiten der Ausprägungen des Objekts g in den ordinalen Variablen für die Klasse k . Die Berechnung dieser Wahrscheinlichkeiten wird in Gleichung 15.10 dargestellt.

$\pi(g|k)^{\text{quant}}$ Auftrittswahrscheinlichkeiten der Ausprägungen des Objekts g in den quantitativen Variablen in der Klasse k . Diese werden entsprechend den Gleichungen 14.5 auf Seite 356 und 14.6 auf Seite 357 berechnet.

Ordinale Variablen werden mittels einer Binomialverteilung modelliert. Dazu wird für jede Variable j ein Parameter $\pi(j|k)$ definiert. Dieser kann als Wahrscheinlichkeit interpretiert werden, dass eine Antwortschwelle überschritten wird, zum Beispiel von »sehr wichtig« zu »wichtig«. Die Überschreitungswahrscheinlichkeit soll für alle Antwortschwellen gleich sein. Angenommen wird, dass die Kategorien mit 0 beginnend kodiert

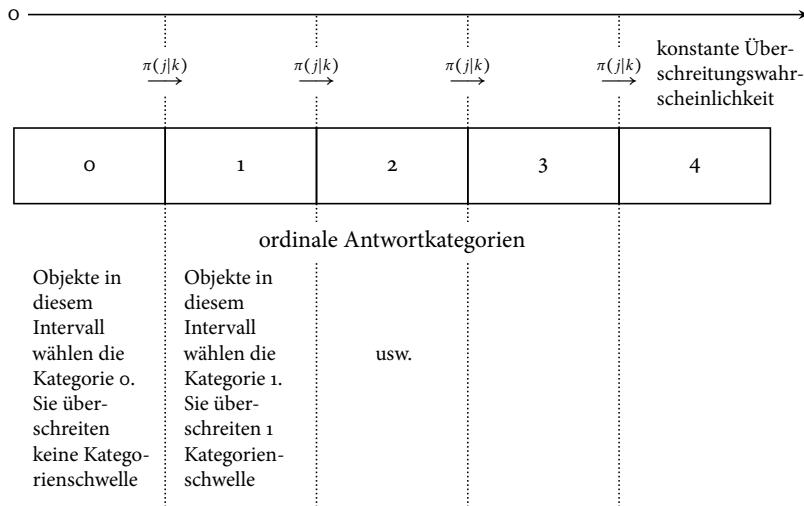


Abb. 15.1: Ordinales Antwortmodell für die Analyse latenter Klassen

sind (siehe Abbildung 15.1). Um die Antwortkategorie i zu erreichen, müssen i Schwellen überschritten werden. Die Zahl der Antwortschwellen ist m_j . Die Wahrscheinlichkeit ist dann:

$$\pi(j(i)|k) = \binom{m_j}{i} \cdot \pi(j|k)^i \cdot (1 - \pi(j|k))^{m_j-i} . \quad (15.10)$$

Betrachten wir ein Beispiel zur Verdeutlichung der unter dieser Annahme entstehenden Antwortwahrscheinlichkeiten: Es soll eine Einstellungsfrage mit fünf Antwortkategorien (0: »lehne stark ab«, 1: »lehne ab«, 2: »unentschieden«, 3: »stimme zu«, 4: »stimme stark zu«) vorliegen. Die Zustimmungswahrscheinlichkeit $\pi(j|1)$ soll 0,2 sein. Für die erste Ausprägung ($i = 0$) berechnet sich die Auftrittswahrscheinlichkeit als

$$\pi(j(0)|1) = \binom{4}{0} \cdot 0,2^0 \cdot (1 - 0,2)^4 = 0,4096 .$$

Für die zweite Ausprägung ($i = 1$, da mit 0 begonnen wird) ist die Auftrittswahrscheinlichkeit gleich

$$\pi(j(1)|1) = \binom{4}{1} \cdot 0,2^1 \cdot (1 - 0,2)^3 = 0,4096 .$$

Für die verbleibenden drei Antwortkategorien ergeben sich Werte von 0,1536, 0,0256 und 0,0016.

Aus den Antwortwahrscheinlichkeiten $\pi(j(i)|k)$ kann die Überschreitungswahrscheinlichkeit berechnet werden:

$$\pi(j|k) = \frac{1}{m_j} \sum_{i=0}^{m_j} i \cdot \pi(j(i)|k) , \quad (15.11)$$

wobei m_j die Zahl der Antwortschwellen der Variablen j ist (Ausprägungszahl minus 1). In unserem Beispiel ergibt sich der vorgegebene Wert von:

$$\pi(j|k) = \frac{1}{4}(0 \cdot 0,4096 + 1 \cdot 0,4096 + 2 \cdot 0,1536 + 3 \cdot 0,0256 + 4 \cdot 0,0016) = \frac{0,8}{4} = 0,2.$$

Die Modellparameter können – abhängig vom Messniveau – entsprechend den bei den einzelnen Modellen angeführten Schätzgleichungen bestimmt werden.

Wir wollen das eben skizzierte Modell als *Analyse latenter Klassen für gemischte Variablen* bzw. kurz als *allgemeine latente Klassenanalyse* bezeichnen. Es besteht aus folgenden Modellparametern:

- Den Klassenanteilswerten $\pi(k)$ der latenten Klassen k ($k = 1, 2, \dots, K$). Diese erfüllen die Nebenbedingungen $\sum \pi(k) = 1$ und $\pi(k) > 0$ (die Summe der Klassenanteilswerte ist gleich 1. Keine Klasse soll leer sein).
- Den (bedingten) Auftrittswahrscheinlichkeiten $\pi(j(i)|k)$ für das Auftreten der Ausprägung i in der nominalen Variablen j in der Klasse k . Diese erfüllen die Nebenbedingung $\sum \pi(j(i)|k) = 1$ für alle j und k (die Summe der Auftrittswahrscheinlichkeiten in einer Klasse soll gleich 1 sein).
- Den (bedingten) Überschreitungswahrscheinlichkeiten $\pi(j|k)$ für die ordinalen Variablen.
- Den Klassenzentren μ_{kj} und Klassenstreuungen σ_{kj}^2 für die quantitativen Variablen.

Ein Berechnungsschema für die Anzahl der Modellparameter eines K -Klassenmodells findet sich in Tabelle 15.10 wieder. In eine Analyse sollen folgende Variablen einbezogen werden:

- nominale Variablen: Geschlecht (2 Ausprägungen), Familienstand (4 Ausprägungen) und berufliche Tätigkeit (7 Ausprägungen),
- ordinale Variablen: Schulbildung (6 Ausprägungen) und
- quantitative Variablen: Einkommen und Kinderzahl.

Tabelle 15.11 zeigt die Berechnung der Anzahl der Modellparameter für eine 6-Klassenlösung aus diesem Beispiel. Insgesamt sind 95 Parameter zu schätzen. Zur Modellprüfung können die bereits bekannten Teststatistiken und Strategien eingesetzt werden. Wie bei allen in diesem Abschnitt behandelten Modellen können die Modellparameter (mit Ausnahme der Klassenvarianzen) fixiert oder durch lineare Restriktionen verbunden werden.

Zu Anwendungsempfehlungen für dieses Kapitel verweisen wir auf die Empfehlungen in Abschnitt 16.4.6.

Tab. 15.10: Berechnungsschema für die Anzahl der Modellparameter eines K-Klassenmodells mit gemischten Variablen

$K - 1$	Anzahl der Anteilswerte $\pi(k)$ der latenten Klassen
$K \cdot \sum_{j=1}^{\text{NOM}} (m_j - 1)$	Anzahl der (bedingten) Auftrittswahrscheinlichkeiten $\pi(j i) k$ bei den nominalen Variablen
$K \cdot \text{ORD}$	Anzahl der Überschreitungswahrscheinlichkeiten $\pi(j k)$ für die ordinalen Variablen
$2 \cdot K \cdot \text{QUANT}$	Anzahl der Klassenzentren und Klassenstreuungen für quantitative Variablen
$K \cdot \left(1 + \sum_{j=1}^{\text{NOM}} (m_j - 1) + \text{ORD} + 2 \cdot \text{QUANT} \right) - 1$	Gesamtzahl zu schätzender Parameter für K Klassen

NOM: Anzahl nominaler Variablen, m_j : Anzahl der Ausprägungen der nominalen Variablen j , ORD: Anzahl ordinaler Variablen, QUANT: Anzahl quantitativer Variablen

Tab. 15.11: Berechnung der Anzahl der Modellparameter eines 6-Klassenmodells für ein Beispiel mit gemischten Variablen

$6 - 1 = 5$	Anteilswerte für die latenten Klassen
$6 \cdot (2 - 1) = 6$	Auftrittswahrscheinlichkeiten für nominale Variable Geschlecht
$6 \cdot (4 - 1) = 18$	Auftrittswahrscheinlichkeiten für nominale Variable Familienstand
$6 \cdot (7 - 1) = 36$	Auftrittswahrscheinlichkeiten für nominale Variable berufliche Fähigkeit
$6 \cdot 1 = 6$	Überschreitungswahrscheinlichkeiten für ordinale Variable Schulbildung
$6 \cdot 2 = 12$	Klassenzentren und Klassenstreuungen für quantitative Variable Einkommen
$6 \cdot 2 = 12$	Klassenzentren und Klassenstreuungen für quantitative Variable Kinderzahl
$\sum = 95$	Gesamtzahl zu schätzender Parameter

16 Latent-GOLD-Ansatz

16.1 Allgemeiner Ansatz und Überblick

Von Vermunt und Magidson (2000, 2002, 2005b) wird ein allgemeiner Ansatz für verschiedene Typen der latenten Klassenmodelle (»LC models«) vorgeschlagen, den die Autoren in einem eigenen Softwarepaket »Latent GOLD« umgesetzt haben. Dieses Programm hat sich im praktischen Vergleich mit anderer Software und anderen Verfahren bei Klassifizierungsaufgaben als sehr geeignet erwiesen (siehe Abschnitt 17.3). Der Latent-GOLD-Ansatz umfasst drei Spezialfälle, die in eigenen Programmodulen umgesetzt werden. Es handelt sich dabei um das *LC-Cluster-Modell*, das *LC-Faktoren-Modell* und das *LC-Regressions-Modell*. Im Folgenden soll der allgemeine Latent-GOLD-Ansatz vorgestellt und auf das Modell LC-Cluster näher eingegangen werden.

Latent GOLD enthält alle bisher in diesem Teil behandelten Verfahren als Submodelle und darüber hinaus folgende Verallgemeinerungen und Erweiterungen:¹

- Neben nominalen, ordinalen und quantitativ-kontinuierlichen Variablen können noch sogenannte *Zählvariablen* (»counts«), wie zum Beispiel die Zahl der Kinder oder die Zahl der im letzten Jahr begangenen Delikte modelliert werden.
- Neben den bisher behandelten Modellprüfgrößen werden weitere Maßzahlen zur Beurteilung der Modellgüte berechnet.
- Die Annahme der lokalen Unabhängigkeit kann abgeschwächt werden.
- Einflüsse von weiteren Variablen auf die Klassenzugehörigkeit und die Klassifikationsmerkmale können spezifiziert werden.
- Daten, die auf komplexen Stichprobenverfahren basieren, können korrekt analysiert werden.

Mit dem LC-Cluster-Ansatz von Latent GOLD können Modelle spezifiziert werden, welche die in der Abbildung 16.1 auf der nächsten Seite dargestellte Struktur besitzen. Es werden drei Variablengruppen unterschieden: *Indikatoren*, *latente Klassen* und *Kovariaten*. Die Indikatoren entsprechen den bisher als Klassifikationsvariablen bzw. -merkmale

¹ Es handelt sich nur um eine beispielhafte Aufzählung.

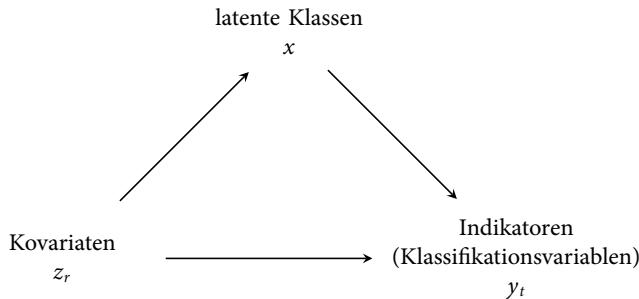


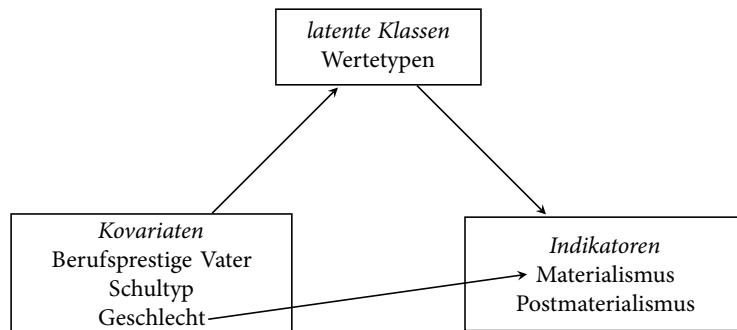
Abb. 16.1: Allgemeine Modellstruktur von Latent GOLD

bezeichneten Variablen. Sie gehen in die Klassenbildung ein. In Latent GOLD werden sie mit y_t abgekürzt. Die latente Klassenvariable wird mit x abgekürzt. Bei der Analyse latenter Klassen ist x eine nominale Variable, bei der in Latent GOLD ebenfalls enthaltenen latenten Faktorenanalyse ist x eine dichotome oder ordinalskalierte Variable. Die bisher behandelten probabilistischen Verfahren enthalten keine Kovariaten. Kovariaten sind Variablen, die einen direkten Einfluss auf x und/oder auf die y_t -Variablen haben.² Sie werden in Latent GOLD mit z_r abgekürzt. Es ist auch möglich, *inaktive Kovariaten* zu spezifizieren. Diese gehen dann nicht in die Schätzung ein. Die für die inaktiven Variablen berechneten Ergebnisse dienen nur der Deskription.

Mit Latent GOLD lässt sich somit folgendes in der Abbildung 16.2 wiedergegebene Modell für die Denz-Daten schätzen. In dem Beispiel wird angenommen, dass die bisher in einem weiteren Analyseschritt zur Validierung verwendeten sozio-demographischen Variablen (Berufsprestige Vater, Geschlecht, besuchter Schultyp) einen Einfluss auf die latente Klassenzugehörigkeit haben. Die inhaltliche Validitätsprüfung, bei der Hypothesen über die latenten Klassen formuliert und empirisch geprüft werden, kann also direkt mit Latent GOLD durchgeführt werden. In der Praxis wird dabei schrittweise vorgegangen. Zunächst wird wie bisher ein Modell ohne Kovariaten geschätzt und interpretiert. In einem weiteren Schritt werden dann Kovariaten hinzugenommen. Wird in einer Validierungshypothese keine kausale, sondern nur eine korrelative Struktur angenommen, zum Beispiel »latente Klassen x korrelieren mit z_h «, dann kann eine inaktive Kovariate definiert werden.

Zusätzlich wird in dem Beispiel ein direkter Einfluss des Geschlechts auf den Materialismus vermutet. Die Skala enthält ein Item bezüglich der Einstellung zur Landesverteidigung. Es wird angenommen, dass wegen der ab dem 18. Lebensjahr drohenden Einberufung viele – auch materialistisch orientierte – junge Männer diesem Item negativ

² Bei der Berechnung der Wirkung einer Kovariaten wird angenommen, dass die Wirkungen aller anderen Kovariaten konstant gehalten werden. Es werden also durchgehend partielle Effekte berechnet.

Abb. 16.2: *Latent-GOLD-Modell für die Denz-Daten*

gegenüberstehen. Durch die Spezifikation eines direkten Effekts des Geschlechts auf den Materialismus wird der Einfluss der persönlichen Betroffenheit statistisch kontrolliert.

Zu den Möglichkeiten und unterschiedlichen Analysemodellen von Latent GOLD siehe ergänzend Tabelle 16.1. Nachfolgend wird nur auf das LC-Cluster-Modell näher eingegangen, da es der »klassischen« LCA entspricht und die Klassifikationsaufgabe im Vordergrund steht. Die in Latent GOLD benutzte Notation fasst Tabelle 16.2 auf der nächsten Seite zusammen.

Tab. 16.1: *Analysemodelle von Latent GOLD*

Modell bzw. Submodelle	Messniveau der manifesten Variablen	latente Variable
LC Cluster (LCA) Submodelle: – klassische LCA – latente Profilanalyse – IRT-Modelle (zum Beispiel Proctor)	beliebige dichotom quantitativ dichotom	latente Klassen: eine nominalskaulierte Variable
LC Factor Submodelle: – Item-Response-Modelle (IRT-Modelle)	beliebig dichotom	latente Faktoren: eine oder mehrere ordinalskalierte Variable
LC Regression Submodelle: – multiple Regression – Wachstumsmodelle	beliebig, aber derselbe Variablentypus quantitativ Messwiederholungen	latente Klassen: eine nominalskaulierte Variable mit unterschiedlichen funktionalen Zusammenhängen

Tab. 16.2: *Latent GOLD-Notation*

Symbol	Beschreibung
f	Wahrscheinlichkeitsdichte
p	Wahrscheinlichkeit
i	Fallindex
n	Anzahl der Fälle
t, T	Indikatorindex
y_{it}	Antwort von Fall i bei Indikator t
m_k	Anzahl der geschätzten Parameter
m, m_t	Kategorie einer nominalen bzw. ordinalen Antwortvariablen
y_m^*, y_m^{t*}	Wert einer Kategorie m einer nominalen bzw. ordinalen Antwortvariablen
h, H	Index für eine Indikatorsubgruppe
T_h^*	Gesamtzahl von Indikatoren in Subgruppe h
y_{ih}	Vektor der Antworten in Subgruppe h
x	nominale latente Variable, eine bestimmte latente Klasse
K	Gesamtzahl latenter Klassen
r, R	Kovariatenindex
z_{ir}^{cov}	Kovariate
μ	Mittelwert bei kontinuierlichen y_{it}
η	linearer Prädiktor
β, γ	Regressionsparameter im Modell für y_{it} bzw. für x
w_i	Fallgewicht
u, U	Index der Kovariatenmuster
i^*, I^*	Index der eindeutigen Datenmuster

16.2 Modellansatz der Latent-Class-Clusteranalyse

Das LC-Cluster-Modell bei Latent GOLD besteht aus einer nominalen latenten Variablen x , T nominalen, ordinalen oder kontinuierlichen Antwort- bzw. Indikatorvariablen y_t und R numerischen oder nominalen Kovariaten z_r^{cov} , die x direkt beeinflussen. Die Kovariaten können auch einen direkten Einfluss auf die Indikatoren haben. Als Indikatoren sind auch sogenannte Zählvariablen (»counts«) zulässig. In Latent GOLD werden die Variablen abweichend von der bisherigen Klassifikation teilweise anders bezeichnet und eingeteilt. Die Terminologie soll daher kurz erklärt werden:

- Nominale Variablen sind Variablen mit nominalem, ordinale Variablen solche mit ordinalem Messniveau.
- Mit kontinuierlichen Variablen sind Variablen mit reellen Zahlenwerten gemeint, also zum Beispiel Einkommens-, Zeit-, Prozent- und Gewichtsangaben. Kontinuierliche Variablen haben ein quantitatives Messniveau (mindestens Intervallskalenniveau).

- Mit Zählvariablen sind Variablen mit ganzzahligen Zahlenwerten größer oder gleich 0 gemeint, also zum Beispiel Zahl der Kinder, Zahl der Wohnräume. Zählvariablen haben quantitatives Messniveau.
- Mit numerischen Variablen sind kontinuierliche Variablen oder Zählvariablen gemeint, also quantitative Variablen.

Auf Seiten der Kovariaten sind in Latent GOLD keine ordinalen Variablen vorgesehen. Sie müssen entweder als nominale oder als numerische Variablen in die Analyse einbezogen werden. Als Indikatoren sind ordinale Variablen hingegen explizit als Variablentyp vorgesehen. Ordinale Variablen können auch als Zählvariablen modelliert werden. Latent GOLD stellt hier den Verteilungstyp der Binomialverteilung zur Verfügung. Diese Modellierung entspricht jener von ALMO für ordinale Variablen (siehe Abschnitt 15.4).

Das allgemeine Modell lautet:

$$f(\mathbf{y}_i | \mathbf{z}_i^{\text{cov}}) = \sum_{x=1}^K p(x | \mathbf{z}_i^{\text{cov}}) \prod_{h=1}^H f(y_{ih} | x, \mathbf{z}_i^{\text{cov}}), \quad (16.1)$$

wobei die exakte Form der jeweiligen klassenspezifischen Wahrscheinlichkeitsverteilung $f(y_{ih} | x, \mathbf{z}_i^{\text{cov}})$ von dem Skalenniveau der Indikatorvariablen abhängt (siehe dazu später).

In der bisherigen Notation würde der LC-Modellansatz geschrieben werden als:

$$\pi(g|z) = \sum \pi(k|z) \prod \pi(x_{gi}|k, z). \quad (16.2)$$

$\pi(g|z)$ ist die Wahrscheinlichkeit, dass die für den Fall g beobachten Variablenwerte in den Indikatorenvariablen bei gegebenen Merkmalsausprägungen in den Kovariaten z auftreten. Diese Größe ist gleich der Summe der Wahrscheinlichkeiten $\pi(k|z)$ des Auftretens der latenten Klasse k bei gegebenen Merkmalsausprägungen in den Kovariaten z multipliziert mit dem Produkt des Auftretens der Variablenwerte in den Indikatorenvariablen bei gegebener Klasse und Merkmalsausprägungen in den Kovariaten. Der möglicherweise auf den ersten Blick schwer nachvollziehbare Modellansatz soll zunächst für das klassische Modell ohne Kovariaten beschrieben werden.

16.2.1 Der klassische Ansatz

Das LCA-Modell in seiner Standardform, wie es in Kapitel 15 erörtert wurde, ist im Latent-GOLD-Ansatz ein *LC-Cluster-Modell mit kategorialen Indikatoren y_{it} und keinen*

Kovariaten (Vermunt und Magidson 2000, S. 21, 22). Ohne Kovariaten vereinfacht sich das Grundmodell aus Gleichung 16.1 auf der vorherigen Seite zu:

$$f(\mathbf{y}_i) = \sum_{x=1}^K p(x) \prod_{t=1}^T f(y_{it}|x) . \quad (16.3)$$

Als Beispiel kann für drei kategoriale Indikatoren ($T = 3$) dann formuliert werden:

$$p(y_{i1} = m_1, y_{i2} = m_2, y_{i3} = m_3) = \sum_{x=1}^K p(x) \prod_{t=1}^3 p(y_{it} = m_t|x) ,$$

wobei die Annahme der lokalen Unabhängigkeit (siehe Abschnitt 14.5) gilt:

$$\prod_{t=1}^3 p(y_{it} = m_t|x) = p(y_{i1} = m_1|x)p(y_{i2} = m_2|x)p(y_{i3} = m_3|x) .$$

Die Indikatoren y_{i1} bis y_{i3} werden innerhalb jeder latenten Klasse als paarweise unabhängig betrachtet. Die konditionalen bzw. bedingten (Antwort-)Wahrscheinlichkeiten $p(y_{it} = m_t|x)$ (Kategorie m der Variable y_{it} unter der Bedingung von x) werden in Latent-GOLD folgendermaßen parametrisiert:

$$p(y_{it} = m|x) = \frac{\exp(\eta_{m|x}^t)}{\sum_{m'=1}^{M_t} \exp(\eta_{m'|x}^t)} ,$$

wobei für den linearen Prädiktor $\eta_{m|x}^t$ der latenten Klasse x für die Ausprägung m der Indikatorvariablen t gilt:

$$\eta_{m|x}^t = \beta_{mo}^t + \beta_{mx}^t .$$

Die Wahrscheinlichkeiten $p(y_{it} = m|x)$ sind die in den vorausgehenden Abschnitten genannten Auftrittswahrscheinlichkeiten $p(x_{gi}|k)$. Der Koeffizient β_{mo}^t lässt sich als Konstanteneffekt auf die Ausprägung m der Indikatorvariablen t interpretieren. Der Koeffizient β_{mx}^t bildet den Effekt der latenten Klasse x auf die Ausprägung m der Indikatorvariablen t ab. Ist β_{mx}^t negativ, tritt die untersuchte Ausprägung m der Indikatorvariablen t in der latenten Klasse x weniger häufig auf, ist der Koeffizientenwert positiv, tritt die betrachtete Ausprägung häufiger auf.

Die in Latent GOLD gewählte Reparametrisierung soll anhand eines Beispiels veranschaulicht werden: In die Analyse werden drei nominalskalierte Variablen y_1 , y_2 und y_3 einbezogen. Es werden drei latente Klassen vorgegeben. Latent GOLD ermittelt die in Tabelle 16.3 dargestellten Ergebnisse.

Tab. 16.3: Ergebnisse der Parameterschätzung im LC-Modell für nominale Indikatorvariablen

	overall	C_1	C_2	C_3	Wald	p-value	R^2
<i>Models for Indicators</i>							
y_1							
1		-0,1370	0,4669	-0,3298	9,3976	0,052	0,2596
2		-1,4190	0,4008	1,0182			
3		1,5560	-0,8677	-0,6883			
y_2							
1		-0,5628	0,4270	0,1358	8,1496	0,017	0,1681
2		0,5628	-0,4270	-0,1358			
y_3							
1		1,2066	-0,1813	-1,0253	6,2960	0,1800	0,1871
2		0,8531	0,6030	-1,4561			
3		-2,0597	-0,4217	2,4814			
<i>Model for Clusters</i>							
Intercept	0,3453	-0,0724	-0,2728	0,9207	0,6300		
<i>Intercepts</i>							
y_1							
1	0,4212				1,5374	0,4600	
2	-0,1503						
3	-0,2709						
y_2							
1	0,0886				0,2715	0,6000	
2	-0,0886						
y_3							
1	1,0203				2,0227	0,3600	
2	0,7893						
3	-1,8096						

Die linearen Prädiktoren $\eta_{m|x=1}^1$ für die Ausprägungen $m = (1, 2, 3)$ der Indikatorvariablen y_1 für die erste latente Klasse ($x = 1$) berechnen sich wie folgt:

$$\eta_{1|x=1}^1 = 0,4212 + (-0,1370) = 0,2842 ,$$

$$\eta_{2|x=1}^1 = -0,1503 + (-1,4190) = -1,5693 ,$$

$$\eta_{3|x=1}^1 = -0,2709 + (1,5560) = 1,2851 .$$

Für die Ausprägungen $m = (1, 2)$ der Indikatorvariablen y_2 ergeben sich folgende lineare Prädiktoren für die erste latente Klasse:

$$\eta_{1|x=1}^2 = 0,0886 + (-0,5628) = -0,4742 ,$$

$$\eta_{2|x=1}^2 = -0,0886 + (+0,5628) = 0,4742 .$$

Für die Ausprägungen $m = (1, 2, 3)$ der Indikatorvariablen y_3 nehmen die linearen Prädiktoren für die erste latente Klasse folgende Werte an:

$$\begin{aligned}\eta_{1|x=1}^3 &= 1,0203 + (-1,2066) = -0,1863 , \\ \eta_{2|x=1}^3 &= 0,7893 + (0,8531) = 1,6424 , \\ \eta_{3|x=1}^3 &= -1,8096 + (-2,0597) = -3,8693 .\end{aligned}$$

Analog können die linearen Prädiktoren für die Ausprägungen der Indikatorvariablen für die anderen beiden Klassen berechnet werden. Die linearen Prädiktoren müssen bestimmte Nebenbedingungen erfüllen, damit sie identifiziert sind. Die Zeilensummen über die latenten Klassen müssen für jede Ausprägung null ergeben (zum Beispiel: $-0,137 + 0,4669 + (-0,3298) = 0$). Die Spaltensummen für jede Variable in jedem Cluster müssen null ergeben (zum Beispiel: $-0,137 + (-1,419) + 1,556 = 0$). Letztere Bedingung gilt auch für den Gesamteffekt (»overall«) jeder Variablen (zum Beispiel: $0,4212 + (-0,1503) + (-0,2709) = 0$). Formal ausgedrückt lauten die Nebenbedingungen:

$$\begin{aligned}\sum_{x=1}^K \beta_{mx}^t &= 0 \quad \forall \text{ Ausprägungen } m \text{ der Indikatoren } t , \\ \sum_{j=1}^{m_t} \beta_{jx}^t &= 0 \quad \forall \text{ Indikatoren } t \text{ und alle latenten Klassen } x , \\ \sum_{j=1}^{m_t} \beta_{jo}^t &= 0 \quad \forall \text{ Indikatoren } t .\end{aligned}$$

Auch andere Formen der Nebenbedingungen sind möglich. So zum Beispiel kann der Effekt der ersten oder letzten Ausprägung gleich null gesetzt werden. Auf der Basis der linearen Prädiktoren lassen sich die Auftrittswahrscheinlichkeiten $p(y_{it} = m|x)$ berechnen. Die Berechnung soll exemplarisch für die Variable y_1 und die erste Klasse ($x = 1$) angeschrieben werden:

$$\begin{aligned}p(y_{i1} = 1|x = 1) &= \frac{\exp(0,2842)}{\exp(0,2842) + \exp(-1,5693) + \exp(1,2851)} = 0,2579 , \\ p(y_{i1} = 2|x = 1) &= \frac{\exp(-1,5693)}{\exp(0,2842) + \exp(-1,5693) + \exp(1,2851)} = 0,0404 , \\ p(y_{i1} = 3|x = 1) &= \frac{\exp(1,2851)}{\exp(0,2842) + \exp(-1,5693) + \exp(1,2851)} = 0,7017 .\end{aligned}$$

Die Auftrittswahrscheinlichkeit für die erste Ausprägung der Variablen y_1 in der ersten Klasse $x = 1$ ist gleich 0,2579 bzw. 25,79 Prozent, jene für die zweite Ausprägung gleich

Tab. 16.4: Profile für das Beispiel des LC-Modell für nominale Indikatorvariablen

	C_1	C_2	C_3
Anteile in %	45,51	29,97	24,53
<i>Indicators</i>			
y_1			
1	0,2579	0,6023	0,2838
2	0,0404	0,3184	0,6170
3	0,7017	0,0794	0,0992
y_2			
1	0,2792	0,7372	0,6104
2	0,7208	0,2628	0,3896
y_3			
1	0,6412	0,3590	0,2871
2	0,3574	0,6243	0,1481
3	0,0014	0,0167	0,5648

0,0404 und jene für die dritte Ausprägung gleich 0,7017. Die Bedingung $\sum_{m'=1}^{M_t} p(y_{it} = m'|x = 1)$ ist selbstverständlich erfüllt. Die Auftrittswahrscheinlichkeiten werden in Latent GOLD als »Profile« ausgegeben. Sie sind für alle drei Klassen und alle Variablen in Tabelle 16.4 wiedergegeben.

Aus den Ergebnissen können auch die Klassenauftrittswahrscheinlichkeiten berechnet werden mit. Dazu werden die in der Ausgabe unter »Model for Clusters« angeführten Koeffizienten (»Intercepts«) verwendet:

$$p(x = k) = \frac{\exp(\gamma_{0k})}{\sum_{x'=1}^K \exp(\gamma_{0x'})} .$$

In dem Beispiel ergeben sich folgende Auftrittswahrscheinlichkeiten für die latenten Klassen:

$$p(x = 1) = \frac{\exp(0,3453)}{\exp(0,3453 + (-0,0724) + (-0,2728))} = 0,4551 ,$$

$$p(x = 2) = \frac{\exp(-0,0724)}{\exp(0,3453 + (-0,0724) + (-0,2728))} = 0,2997 .$$

$$p(x = 3) = \frac{\exp(-0,2729)}{\exp(0,3453 + (-0,0724) + (-0,2728))} = 0,2453 .$$

Die erste Klasse tritt mit einer Wahrscheinlichkeit von 45,51 Prozent auf. Für die zweite latente Klasse ist die Auftrittswahrscheinlichkeit gleich 29,97 Prozent und für die dritte gleich 24,53 Prozent.

16.2.2 Erweiterung mit Kovariaten

Das beschriebene »klassische« LC-Cluster-Modell kann über »aktive« Kovariaten z_{ir} als Prädiktoren für die Klassenzugehörigkeiten erweitert werden (Vermunt und Magidson 2000, S. 22, 23; für den Ansatz siehe Clogg 1981; Dayton und Macready 1988; Hagenaars 1993). Führt man in das oben genannte Beispiel mit drei kategorialen Indikatoren y_{i1} bis y_{i3} zwei Kovariaten z_{i1} und z_{i2} ein, so modelliert man:

$$p(y_{i1} = m_1, y_{i2} = m_2, y_{i3} = m_3 | z_{i1}^{\text{cov}}, z_{i2}^{\text{cov}}) = \sum_{x=1}^K p(x | z_{i1}^{\text{cov}}, z_{i2}^{\text{cov}}) \prod_{t=1}^3 p(y_{it} = m_t | x).$$

In diesem Modell wird angenommen, dass die Kovariaten z_1 und z_2 nur die Klassenzugehörigkeit beeinflussen, nicht aber die Indikatorvariablen bei gegebener latenten Variablen x . Der direkte Einfluss der Kovariaten auf die Klassenzugehörigkeiten $p(x | z_{i1}^{\text{cov}}, z_{i2}^{\text{cov}})$ wird wiederum durch ein multinomiales logistisches Regressionsmodell geschätzt. Höhere Interaktionsterme werden vom Modell ausgeschlossen. Es gilt damit:

$$p(x | z_{i1}^{\text{cov}}, z_{i2}^{\text{cov}}) = \frac{\exp(\eta_x | z_{i1}, z_{i2})}{\sum_{x'=1}^K \exp(\eta_{x'} | z_{i1}, z_{i2})},$$

wobei für den linearen Prädiktor für die latente Klasse x , gegeben die Kovariaten z_1 und z_2 , folgende Funktion angenommen wird:

$$\eta_x | z_{i1}, z_{i2} = \gamma_{x0} + \gamma_{x1} z_{i1} + \gamma_{x2} z_{i2}.$$

Der Index x bringt zum Ausdruck, dass die Zugehörigkeit zur latenten Klasse x untersucht wird, während die Indizierung mit m und t bedeutet, dass die Ausprägung m der Indikatorvariablen t untersucht wird. Der Koeffizient γ_{x0} kann als Konstanteneffekt auf die Klassenzugehörigkeit interpretiert werden, die Koeffizienten γ_{x1} und γ_{x2} als Effekte der Kovariaten z_1 und z_2 auf die Klassenzugehörigkeit. Das Modell kann noch um Effekte der Kovariaten auf die Indikatorvariablen erweitert werden. In diesem Fall wird der lineare Prädiktor $\eta_{m|x}^t$ für die Ausprägung m der Indikatorvariablen t um die Effekte β_{mj}^t der Kovariaten z_j erweitert:

$$\eta_{m|xz_{i1}z_{i2}}^t = \beta_{mo}^t + \beta_{mx}^t + \beta_{m1}^t z_{i1} + \beta_{m2}^t z_{i2}.$$

Die allgemeine Gleichung lautet:

$$\eta_{m|xz_{i1}z_{i2}}^t = \beta_{mo}^t + \beta_{mx}^t + \sum_j \beta_{mj}^t z_{ij}.$$

Zur Verdeutlichung soll das obige Beispiel fortgesetzt werden. Es werden die Kovariaten z_1 und z_2 in das Modell aufgenommen. Angenommen wird, dass beide Kovariaten die

Tab. 16.5: Ergebnisse der Parameterschätzung im LC-Modell für nominale Indikatorvariablen mit Kovariaten

	overall	C_1	C_2	C_3	Wald	p-Wert	R^2
<i>Models for Indicators</i>							
γ_1							
1		-0,0880	0,2646	-0,1766	17,8129	0,0013	0,2045
2		-1,3570	0,7844	0,5726			
3		1,4450	-1,0490	-0,3960			
γ_2							
1		-0,6658	-0,0688	0,7346	10,4138	0,0055	0,1612
2		0,6658	0,0688	-0,7346			
γ_3							
1		1,5993	-1,2048	-0,3945	5,8694	0,2100	0,1664
2		1,1137	0,1706	-1,2843			
3		-2,7130	1,0342	1,6788			
<i>Model for Clusters</i>							
Intercept (γ_{x0})		1,7749	-6,9993	5,2244	12,9635	0,0015	
$z_1(\gamma_{x1})$		-0,3212	1,8454	-1,5242	19,9936	0,0000	
$z_2(\gamma_{x2})$		0,3387	0,2137	-0,5525	0,7516	0,6900	
<i>Intercepts</i>							
γ_1							
1		0,4134			5,2501	0,0720	
2		-0,1165					
3		-0,2969					
γ_2							
1		0,2871			1,2226	0,2700	
2		-0,2871					
γ_3							
1		1,7850			5,1929	0,0750	
2		1,1365					
3		-2,9215					
Indicator γ_3							
1 2 3 Wald p-Wert							
<i>Direct Effects</i>							
z_1		-0,2493	-0,1313	0,3805	3,3721	0,1900	

Klassenzugehörigkeit beeinflussen. Zusätzlich wird ein Effekt von z_1 auf die dritte Indikatorvariable y_3 vermutet. Die von Latent GOLD berechneten Ergebnisse gibt Tabelle 16.5 auf der vorherigen Seite wieder.

In der Tabelle werden zunächst die direkten Effekte β_{mx}^t der latenten Klassen auf die Indikatorvariablen dargestellt. Die direkten Effekte $\beta_{mj=1}^t$ ($m = 3$) der Kovariaten z_1 auf y_3 werden in der Tabelle ganz unten aufgeführt. Der direkte Effekt von z_1 auf die erste Ausprägung von y_3 ist $-0,2493$. Dies bedeutet, dass die Wahrscheinlichkeit der Ausprägung »1« abnimmt, wenn z_1 größere Werte hat. Auch für die zweite Ausprägung von y_3 liegt mit $-0,1313$ ein negativer Effekt vor. Die Wahrscheinlichkeit des Auftretens dieser Ausprägung nimmt ebenfalls mit steigenden Werten von z_1 ab. Für die dritte Ausprägung wird ein positiver Effekt ausgewiesen. Die Auftrittswahrscheinlichkeit erhöht sich, wenn z_1 höhere Werte aufweist. Wiederum gilt die Nebenbedingung, dass die Summe der Effekte einer Kovariaten über alle Ausprägungen hinweg null ergibt: $\sum_{m'} \beta_{m'j}^t = 0$. Wie beim gewöhnlichen Modell existieren daneben auch andere Möglichkeiten der Parametrisierung.

Die direkten Effekte der Kovariaten auf die Klassenzugehörigkeit werden unter »Model for Clusters« ausgegeben. Für die latenten Klassen ergeben sich folgende lineare Prädiktoren:

$$\begin{aligned}\eta_{x=1|z_{i1}, z_{i2}} &= 1,7749 + (-0,3212)z_{i1} + 0,3387z_{i2}, \\ \eta_{x=2|z_{i1}, z_{i2}} &= -6,9993 + 1,8454z_{i1} + 0,2137z_{i2}, \\ \eta_{x=3|z_{i1}, z_{i2}} &= 5,2244 + (-1,5242)z_{i1} + (-0,5525)z_{i2}.\end{aligned}$$

Wiederum gilt die Nebenbedingung, dass die Zeilensummen null sein müssen. Der direkte Effekt von $-0,3212$ von z_1 auf die erste latente Klasse bedeutet, dass die Wahrscheinlichkeit der Zugehörigkeit zur ersten latenten Klassen abnimmt, wenn z_1 höhere Werte aufweist. Die Kovariate z_2 hat auf die erste latenten Klasse mit $0,3387$ einen positiven Effekt. Die Wahrscheinlichkeit der Zugehörigkeit zur ersten Klasse steigt folglich, wenn z_2 höhere Werte hat. Die direkten Effekte für die anderen Klassen sind analog zu interpretieren.

Die Kovariaten können in Latent GOLD numerisch (quantitativ) oder kategorial (nominal) sein. Im obigen Beispiel wurden numerische Kovariaten betrachtet. Bei kategorialen Kovariaten wird für jede Ausprägung der Kovariaten ein direkter Effekt berechnet und ausgewiesen.

16.2.3 Ordinale Indikatorvariablen

Im Fall von *ordinalen Indikatorvariablen* y_{it} wird eine Restriktion in der Art eingeführt, dass gilt:

$$\beta_{mx}^t = \beta_x^t \cdot y_m^{t^*},$$

wobei $y_m^{t^*}$ die Kategorienwerte der ordinalen Variablen t^* sind. Da ordinale Variablen im Allgemeinen ganzzahlig kodiert sind, vereinfacht sich die Formel zu:

$$\beta_{mx}^t = \beta_x^t \cdot m.$$

Für die erste Ausprägung gilt $\beta_{1x}^t = 1\beta_x^t$, für die zweite Ausprägung $\beta_{1x}^t = 2\beta_x^t$, für die dritte $\beta_{1x}^t = 3\beta_x^t$ usw. Für jede Indikatorvariable wird für jede latente Klasse nur ein Parameter berechnet.

In Tabelle 16.6 auf der nächsten Seite werden zu Demonstrationszwecken alle Variablen des bisherigen Beispiels als ordinal definiert. Für y_1 ergibt sich für die erste latente Klasse ein Wert von $\beta_1^1 = -2,5574$. Das bedeutet, dass in der ersten latenten Klasse eher geringere Werte in y_1 auftreten. Für die zweite Klasse ist β_2^1 gleich 3,5087. Der Wert bedeutet, dass höhere Werte auftreten. Für die dritte Klasse ist der Effekt gleich -0,9513. Auch er ist negativ, aber näher bei null als jener für die erste latente Klasse. Es besteht auch hier eine Tendenz zu kleineren Werten, aber deutlich schwächer als für die erste Klasse. Für y_2 sind die Koeffizientenwerte: $\beta_1^2 = -0,2892$, $\beta_2^2 = 0,8366$ und $\beta_3^2 = -0,9513$. Wiederum gelten folgende Nebenbedingungen:

$$\sum_{x=1}^K \beta_x^t = 0,$$

$$\sum_{j=1}^{m_t} \beta_{j0}^t = 0.$$

Beispielhaft sollen die Auftrittswahrscheinlichkeiten für die Ausprägungen von y_1 für die erste latente Klasse berechnet werden. Die linearen Prädiktoren sind:

$$\eta_{1|x=1}^1 = -0,8607 + -2,5574 \cdot 1 = -3,41810,$$

$$\eta_{2|x=1}^1 = 0,8117 + -2,5574 \cdot 2 = -4,3031,$$

$$\eta_{3|x=1}^1 = 0,0490 + -2,5574 \cdot 3 = -7,6232.$$

Tab. 16.6: Ergebnisse der Parameterschätzung im LC-Modell für ordinale Indikatorvariablen

	overall	C_1	C_2	C_3	Wald	p-value	R^2
<i>Models for Indicators</i>							
y_1		-2,5574	3,5087	-0,9513	20,7791	0,00003	0,7341
y_2		-0,2892	0,8366	-0,5474	13,8795	0,00097	0,0796
y_3		-1,2887	-1,7981	3,0867	5,8628	0,05300	0,5007
<i>Model for Clusters</i>							
Intercept		0,3488	0,1488	-0,4976	3,4129	0,18000	
<i>Intercepts</i>							
y_1							
1	-0,8607				8,6597	0,013	
2	0,8117						
3	0,0490						
y_2							
1	0,0267				0,1072	0,7400	
2	-0,0267						
y_3							
1	-0,0508				2,6626	0,2600	
2	1,1151						
3	-1,0644						

Mit diesen Ergebnissen lassen sich die Auftrittswahrscheinlichkeiten der Ausprägungen in y_1 für die erste latente Klasse bestimmen:

$$p(y_{it} = 1|x = 1) = \frac{\exp(-3,41810)}{\exp(-3,41810) + \exp(-4,3031) + \exp(-7,6232)} = 0,7005 ,$$

$$p(y_{it} = 2|x = 1) = \frac{\exp(-4,3031)}{\exp(-3,41810) + \exp(-4,3031) + \exp(-7,6232)} = 0,2891 ,$$

$$p(y_{it} = 3|x = 1) = \frac{\exp(-7,6232)}{\exp(-3,41810) + \exp(-4,3031) + \exp(-7,6232)} = 0,0105 .$$

Eine Besonderheit der gewählten Parametrisierung für ordinale Variablen ist, dass im 1-Klassenmodell die Ordinalität nicht berücksichtigt wird. Das 1-Klassenmodell für ordinale Indikatorenvariablen ist identisch mit dem 1-Klassenmodell für nominale Variablen. Auch im Modell mit ordinalen Indikatorvariablen können Kovariaten berücksichtigt werden (Vermunt und Magidson 2005b), auf eine explizite Darstellung wird hier jedoch verzichtet.

16.2.4 Kontinuierliche Indikatorvariablen

Im Fall von *kontinuierlichen Indikatorvariablen* y gestaltet sich die grundlegende Wahrscheinlichkeitsstruktur wie folgt (Vermunt und Magidson 2005b, S. 25, 26):

$$f(\mathbf{y}_i) = \sum_x^K p(x) f(\mathbf{y}_i|x) .$$

Nimmt man für das am wenigsten restriktive Modell an, dass die Indikatoren klassenspezifischen Multinormalverteilungen folgen, so kann man formulieren (siehe auch Banfield und Raftery 1993; Wolfe 1970):

$$f(\mathbf{y}_i|x) = (2\pi)^{-K_m/2} \det \left(\sum_x \exp \left(-\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_x)' \boldsymbol{\Sigma}_x^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_x) \right) \right)^{-1/2} .$$

Für jede Klasse werden spezifische Mittelwerte $\boldsymbol{\mu}_x$ und Varianz-Kovarianzmatrizen $\boldsymbol{\Sigma}_x$ spezifiziert. Das restriktivste Modell setzt alle Kovarianzen gleich 0, welches der Annahme der lokale Unabhängigkeit entspricht. Dann ergibt sich:

$$f(\mathbf{y}_i) = \sum_{x=1}^K p(x) \prod_{t=1}^T f(y_{it}|x) ,$$

mit

$$f(y_{it}|x) = \frac{1}{\sigma_{tx}\sqrt{2\pi}} \exp \left(-\frac{(y_{it} - \mu_{tx})^2}{2\sigma_{tx}^2} \right) .$$

Auch im Modell mit kontinuierlichen Indikatorvariablen können Kovariaten berücksichtigt werden (Vermunt und Magidson 2005b). Das Vorgehen wird im Anwendungsbeispiel in Abschnitt 16.5 dargestellt. Das Basismodell entspricht der in Kapitel 14 behandelten latenten Profilanalyse.

16.2.5 Zählvariablen

Sollen, wie oben schon erwähnt, *Zählvariablen* zum Einsatz kommen, so können sie mit Hilfe von *Poisson-* oder *Binomialverteilungen* modelliert werden. Die Poissonverteilung, auch als »Verteilung der seltenen Ereignisse« bekannt, liefert Voraussagen über die Anzahl des Eintretens seltener, zufälliger und voneinander unabhängiger Ereignisse innerhalb eines Intervalls. Die Verteilung hängt davon ab, wie viele Ereignisse im Durchschnitt innerhalb dieses Intervalls erwartet werden. Die Poissonverteilung wird häufig in kriminologischen Untersuchungen angewandt. Erfasst wird hier, wie oft ein bestimmtes

Delikt in einem bestimmten Zeitraum (zum Beispiel im letzten Jahr, in den letzten fünf Jahren usw.) ausgeübt wurde. Angenommen wird, dass das Delikt ein seltenes Ereignis ist.

In Latent GOLD wird die Poissonverteilung folgendermaßen spezifiziert (Vermunt und Magidson 2005b, S. 11–12):

$$\begin{aligned} P(y_{it}|x, z_i, e_{it}) &= \frac{1}{y_{it}!} (\theta_{t,x,z_i} e_{it})^{y_{it}} \exp(-\theta_{t,x,z_i} e_{it}) \\ &= \frac{1}{y_{it}!} (\mu_{t,x,i})^{y_{it}} \exp(-\mu_{t,x,i}) . \end{aligned}$$

Hierbei bezeichnet θ_{t,x,z_i} die sogenannte *Poissonrate*, also die oben genannte mittlere Anzahl an Ereignissen im untersuchten Intervall. Der Term e_{it} ist der Beitrag von Fall i am Ereignis t , wird aber im LC-Cluster-Modell generell gleich 1 gesetzt. Es gilt des Weiteren: $\mu_{t,x,i} = \theta_{t,x,z_i} e_{it}$. Für die Poissonrate kann hinsichtlich des linearen Prädiktors η_{x,z_i}^t formuliert werden:

$$\theta_{t,x,z_i} = \exp(\eta_{x,z_i}^t) \quad \text{bzw.} \quad \mu_{t,x,i} = \exp(\eta_{x,z_i}^t) e_{it} .$$

In dem Fall, in dem nur Fälle mit Zählvariablen ungleich 0 in die Analyse einbezogen werden, kann die »abgeschnittene« (»truncated«) Poissonverteilung verwendet werden:

$$P(y_{it}|x, z_i, e_{it}, y_{it} > 0) = \frac{P(y_{it}|x, z_i, e_{it})}{1 - P(0|x, z_i, e_{it})} , \quad (16.4)$$

mit $P(y_{it}|x, z_i, e_{it})$ als Poissonwahrscheinlichkeit des Auftretens von y_{it} Ereignissen und $P(0|x, z_i, e_{it}) = \exp(-\mu_{t,x,i})$ als Poissonwahrscheinlichkeit des Auftretens von 0 Ereignissen. Auf diese Verteilung kann man zum Beispiel zurückgreifen, wenn nur Personen untersucht werden sollen, die mindestens ein Delikt begangen haben. Damit kann das Problem bei der Analyse von seltenen Ereignissen, das darin besteht, dass die Nullausprägungen (kein Ereignis) häufiger als aufgrund der angenommenen Verteilung angenommen, auftreten (»zero inflation«), berücksichtigt werden. Für (latente) Regressionsmodelle wird mit der negativen Binomialverteilung (Hilbe 2007) ein spezielles Verteilungsmodell für diese Problemsituation angeboten, das allgemein eine Modellierung der Überdispersion erlaubt (Vermunt und Magidson 2005a, S. 59).

Sollen Zählvariablen mit Hilfe der Binomialverteilung modelliert werden – beispielsweise um multiple »Versuche« oder »Ziehungen« beschreiben zu können, wird formuliert (Vermunt und Magidson 2005b, S. 12–14):

$$P(y_{it}|x, z_i, e_{it}) = \binom{e_{it}}{y_{it}} (\pi_{t,x,z_i})^{y_{it}} (1 - \pi_{t,x,z_i})^{(e_{it}-y_{it})} .$$

In diesem Fall bezeichnet e_{it} die Anzahl von maximalen Versuchen für ein Individuum i in der Indikatorvariablen t . Bei der LC-Cluster-Analyse wird e_{it} gleich dem höchsten beobachteten Wert in der Klassifikationsvariablen y_t gesetzt. Es wird also angenommen, dass es zumindest eine Person gibt, bei der das Maximum empirisch auftritt. Für die Parametrisierung gilt:

$$\pi_{t,x,z_i} = \frac{\exp(\eta_{x,z_i}^t)}{1 + \exp(\eta_{x,z_i}^t)}.$$

Mit der Binomialverteilung lässt sich ebenfalls die Zahl der Delikte in einem bestimmten Zeitraum untersuchen. Mathematisch lässt sich zeigen, dass die Poissonverteilung die Grenzverteilung der Binomialverteilung ist, wenn die maximale Zahl der Versuche gegen unendlich geht ($e_{it} \rightarrow \infty$). In diesem Fall ist die durchschnittliche Anzahl der Versuche der Poissonverteilung gleich dem Erwartungswert der Binomialverteilung: $\mu_{t,x,i} = e_{it} \cdot \pi_{t,x,z_i}$. Die Beziehung zwischen beiden Variablen zeigt Abbildung 16.3 auf der nächsten Seite. Angenommen wurde eine maximale Zahl von sechs Versuchen und eine Auftrittswahrscheinlichkeit eines einzelnen Ereignisses von 0,1. Die Poissonverteilung fällt in dem Beispiel zum einen zwischen 0 und 1 Ereignissen steiler ab, dann aber flacher als die Binomialverteilung, da die maximale Zahl nicht auf 6 begrenzt ist. Ist bei seltenen Ereignissen ein zunächst steiler Abfall und anschließend ein flacher Verlauf erwünscht, wird man sich für die Poissonverteilung entscheiden. Die Poissonverteilung wird auch dann ausgewählt werden, wenn die Zahl der Versuche sehr groß ist bzw. wenn es empirisch einige wenige Fälle mit sehr großen Werten gibt. Umgekehrt ermöglicht die Binomialverteilung eine flexiblere Modellierung. Es lassen sich links- und rechtschiefe sowie flache oder spitze Verteilungen modellieren. Sie wird daher in ALMO zur Modellierung ordinaler Variablen genutzt (siehe Abschnitt 15).

Ebenso wie bei der Poissonverteilung (Gleichung 16.4) kann eine abgeschnittene Binomialverteilung verwendet werden. Es gilt dann, dass $P(y_{it}|x,z_i,e_{it})$ und $P(0|x,z_i,e_{it}) = (1 - \pi_{t,x,z_i})^{e_{it}}$ Binomialwahrscheinlichkeiten für y_{it} bzw. 0 Ereignisse sind.

16.3 Parameterschätzung

Um die *Parameter der vorgestellten Modelle zu schätzen*, kommen bei Latent GOLD die *Maximum-Likelihood-*(ML) und die *Posterior-Mode-Methode* (PM) zum Einsatz. Der Vektor, der die unterschiedlichen Modellparameter γ und β enthält, für die die Wahrscheinlich-

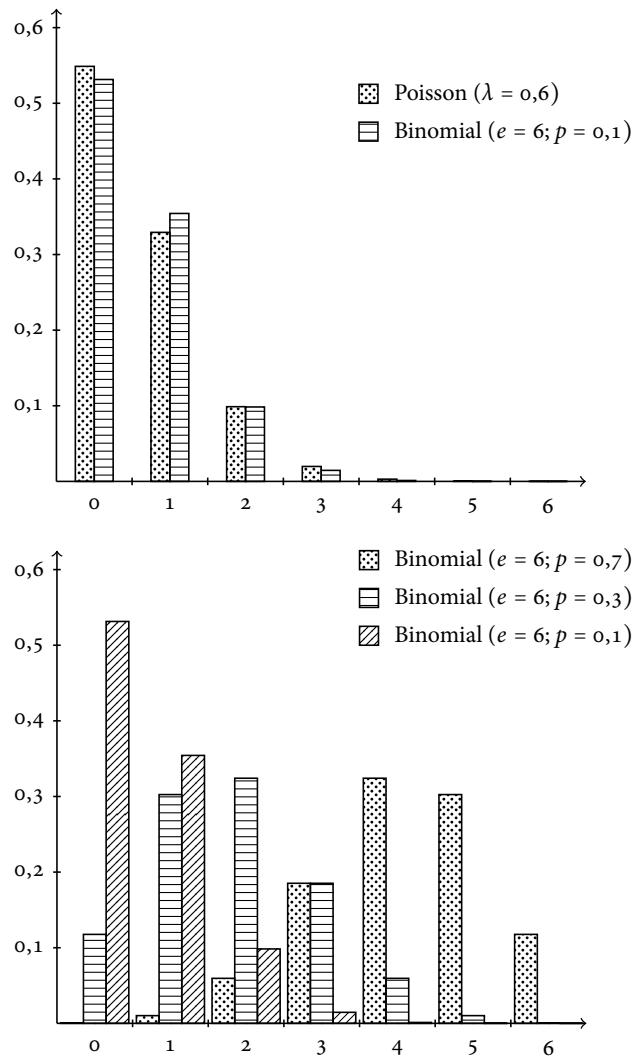


Abb. 16.3: Veranschaulichung der Poission- und Binomialverteilung

keitsfunktion maximiert werden muss, wird als $\boldsymbol{\theta}$ bezeichnet. Die Maximum-Likelihood-Schätzung beruht damit auf der Maximierung der folgenden Log-Likelihoodfunktion:

$$\text{LL} = \sum_{i=1}^n w_i \log f(y_i | z_i, \boldsymbol{\theta}),$$

mit y_i als abhängigen bzw. Indikatorvariablen und z_i als unabhängigen Variablen bzw. Kovariaten. Die Fallanzahl wird mit n , ein einzelner Fall mit i bezeichnet. Der Term

$f(\mathbf{y}_i | \mathbf{z}_i, \boldsymbol{\vartheta})$ beschreibt die Wahrscheinlichkeitsdichte, die mit dem Fall i unter den gegebenen Variablenwerten in den Kovariaten \mathbf{z} und den Parameterwerten $\boldsymbol{\vartheta}$ verbunden ist und w_i stellt einen möglichen Gewichtungsparameter des Falles i dar.

Um das Problem von *Randlösungen* bei der Schätzung zu umgehen, was gleichbedeutend damit ist, dass möglicherweise keine Maximum-Likelihood-Schätzer existieren, wird in Latent GOLD auf Aspekte des Bayes'schen Schätzverfahrens zurückgegriffen (Details siehe Abschnitt 17.1). Dabei wird die Unsicherheit über den Parametervektor mit Hilfe einer Wahrscheinlichkeitsdichte ausgedrückt. Grundlage für Schätzungen nach dem Bayes'schen Verfahren sind folgende Voraussetzungen:

- vor jeglicher Beobachtung ist die Priori-Dichte³ anzugeben (der sogenannte »Prior«),
- nach einer Beobachtung der Daten wird daraus eine Posteriori-Dichte⁴ (der sogenannte »Posterior«).

Bei Latent GOLD kommen spezielle Varianten, »Dirichlet-Priors« und »inverse Wishart-Priors«, zum Einsatz (Clogg, Rubin u. a. 1991; Gelman, Carlin u. a. 1995; Schafer 1997). Die darauf beruhende Schätzmethode wird dann allerdings nicht mehr Maximum-Likelihood- sondern Posterior-Mode-Methode (PM-Methode) genannt. Bezeichnet man die genannten Priors für den Parametervektor $\boldsymbol{\vartheta}$ mit $p(\boldsymbol{\vartheta})$ und den Posterior mit \mathcal{P} , so lautet die Schätzaufgabe für die PM-Schätzung: Finde Parameter für $\boldsymbol{\vartheta}$, so dass die folgende Log-Posterior-Funktion maximiert wird:

$$\begin{aligned}\log \mathcal{P} &= \text{LL} + \log p(\boldsymbol{\vartheta}) \\ &= \sum_{i=1}^n w_i \log f(\mathbf{y}_i | \mathbf{z}_i, \boldsymbol{\vartheta}) + \log p(\boldsymbol{\vartheta}),\end{aligned}$$

was gleichbedeutend ist mit der Suche nach einer Lösung für:

$$\frac{\partial \log \mathcal{P}}{\partial \boldsymbol{\vartheta}} = \frac{\partial \text{LL}}{\partial \boldsymbol{\vartheta}} + \frac{\partial \log p(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = 0.$$

Von der Anwenderin können bestimmte Parameter der Priors $p(\boldsymbol{\vartheta})$ so gesetzt werden, dass $\log p(\boldsymbol{\vartheta}) = 0$ gilt, womit aus der PM-Schätzung wieder eine ML-Schätzung wird. Generell kann die PM-Schätzung als eine Art ML-Schätzung mit einer »Bestrafungsfunktion« $p(\boldsymbol{\vartheta})$ angesehen werden, die Lösungen des Schätzproblems, die zu nahe am Rand liegen, »bestraft« und somit erschweren oder verhindern soll.

Konkret müssen drei Parameter für die Priori-Verteilungen definiert werden, die als α_1 , α_2 und α_3 bezeichnet werden. Der Parameter α_1 bezieht sich auf die Priori-Verteilung der

³ In der Literatur werden die Bezeichnungen »A-priori-Dichte«, »Priori-Dichte« bzw. »A-priori-Verteilung«, »Priori-Verteilung« synonym verwendet.

⁴ Auch in diesem Fall wird »A-posteriori« und »Posteriori« synonym verwendet.

Klassenanteile. Je größer der Parameter gewählt wird, desto stärker ist die Annäherung an eine Gleichverteilung der Klassengrößen und desto schwächer ist der Effekt von Kovariaten. Für $\alpha_1 = 0$ erfolgt eine ML-Schätzung für die Klassenanteile. Die Voreinstellung ist $\alpha_1 = 1$. Der Parameter α_2 steuert die bedingten Auftrittswahrscheinlichkeiten innerhalb der latenten Klassen. Je größer der Wert gewählt wird, desto stärker nähern sich die bedingten Auftrittswahrscheinlichkeiten einer Gleichverteilung an. Die Voreinstellung ist wiederum $\alpha_2 = 1$; die Einstellung $\alpha_2 = 0$ resultiert in einer ML-Schätzung für die Auftrittswahrscheinlichkeiten. Der Parameter α_3 schließlich beeinflusst die Schätzung der Varianzen und Kovarianzen innerhalb der latenten Klassen. Je größer er gesetzt wird, desto größer werden die Varianzen innerhalb der latenten Klassen und desto schwächer werden die Kovarianzen innerhalb der latenten Klassen, sofern sie modelliert werden. Werden alle drei Parameter gleich 0 gesetzt ($\alpha_1 = \alpha_2 = \alpha_3 = 0$), wird eine ML-Schätzung berechnet. Alle nachfolgend berichteten Ergebnisse werden – sofern nichts anderes angeführt ist – mit den Voreinstellungen ($\alpha_1 = \alpha_2 = \alpha_3 = 1$) gerechnet.

Der Einfluss der Priori-Parameter hängt von der Stichprobengröße ab. Je größer diese ist, desto schwächer ist ihr Effekt. Tabelle 16.7 veranschaulicht exemplarisch den Einfluss des Parameters α_1 für ein 2-Klassenmodell in Abhängigkeit von der Stichprobengröße. Wiedergeben ist nur der Klassenanteil der ersten Klasse, jener der zweiten ergibt sich residual. Für $\alpha_1 = 0$ ergibt sich bei einer Stichprobengröße von 221 ein Anteil von 0,7493. Dieser reduziert sich schrittweise, wenn α_1 erhöht wird. Für $\alpha_1 = 10$ beträgt er 0,7028. Der Effekt ist also eher schwach. Halbiert man die Stichprobe auf $n = 110,5$ Fälle, indem jeder Fall mit 0,5 gewichtet wird, resultiert ein stärkerer Effekt. Der Klassenanteils Wert geht von 0,7518 auf 0,6726 zurück. Wird die Stichprobe durch Gewichtung mit 0,1 auf $n = 22,1$ Fälle verkleinert, verstärkt sich der Effekt und für $\alpha_1 = 10$ wird beinahe der bei zwei Kategorien für eine Gleichverteilung erwartete Wert von 0,50 erreicht. Bei einer Vergrößerung der Stichprobe schwächt sich der Einfluss des Priori-Parameters ab. Bei einer Verzehnfachung auf $n = 2210$ ist er so gut wie nicht mehr auffindbar. Bei in den Sozialwissenschaften üblichen größeren Stichproben von $n = 2\,000$ oder mehr Fällen, ist der Einfluss also gering. Bei kleinen Stichproben kann die Wahl der Priori-Parameter einen entscheidenden Einfluss ausüben. Anhaltspunkte für eine »richtige« Wahl fehlen derzeit leider. Eine falsche Wahl von Priori-Verteilungen kann zu falschen Ergebnissen führen (McLachlan und Peel 2000, S. 125–126). Wir empfehlen ein schrittweises Vorgehen und zunächst mit einem kleinen Wert der Parameter zu beginnen. Wird eine Randlösung berechnet, sollten die Parameter in einem nächsten Schritt erhöht werden, wenn diese nicht erwünscht ist.

Die Parameterschätzung wird bei Latent GOLD mit einer Kombination aus zwei iterativen Verfahren durchgeführt: dem *Newton-Raphson-* (NR-Verfahren) und dem *Expectation-Maximization-Verfahren* (EM-Verfahren, siehe Abschnitt 14.1). Eine Eigenschaft des EM-

Tab. 16.7: Einfluss des Priori-Parameters α_1

α_1	$n = 21,1^a$	$n = 110,5^a$	$n = 221$	$n = 442^a$	$n = 2210^a$
0	0,7675	0,7518	0,7493	0,7481	0,7470
1	0,7200	0,7408	0,7437	0,7451	0,7465
2	0,6868	0,7308	0,7384	0,7425	0,7459
3	0,6619	0,7215	0,7333	0,7398	0,7453
5	0,6272	0,7051	0,7237	0,7345	0,7442
10	0,5818	0,6726	0,7028	0,7224	0,7414

a) rechnerische Größe durch Gewichtung der Fälle

Algorithmus ist es, am Anfang des Iterationsprozesses in relativ großen Schritten gegen das Maximum der Likelihoodfunktion, gegen Ende jedoch sehr langsam, mit vielen Iterationen voranzuschreiten. Im Vergleich zum EM-Algorithmus konvergiert der NR-Algorithmus dafür mit weniger, dafür aber rechenintensiveren Iterationen als der EM-Algorithmus. Der größte Nachteil des NR-Algorithmus ist jedoch, dass in vielen Fällen sehr gute *Startwerte* erforderlich sind, um konvergieren zu können, beim EM-Algorithmus sind relativ grobe, oftmals sogar zufällige Startwerte ausreichend (Andréß, Hagenaars u. a. 1997, S. 220–223). Die von Latent GOLD verwendete Kombination aus beiden Algorithmen versucht, sich beide Stärken zu eignen zu machen: Zu Beginn des Iterationsprozesses wird wegen seiner Stabilität der EM-Algorithmus eingesetzt, nahe dem Optimum aber auf den NR-Algorithmus umgeschaltet, um die schnellere Iterationsgeschwindigkeit auszunutzen (Vermunt und Magidson 2000, S. 165, 167). Im einzelnen wird der EM-Algorithmus so lange durchgeführt, bis entweder das voreingestellte Maximum der Rechenschritte (»Iteration Limits EM«) erreicht oder ein *Konvergenzkriterium* erfüllt ist (»EM Tolerance«). Danach wird auf den NR-Algorithmus umgeschaltet und dieser so lange benutzt, bis auch hier die Anzahl der eingestellten Rechenschritte erreicht ist (»Iteration Limits Newton-Raphson«) oder das *Gesamtkonvergenzkriterium* erfüllt ist (»Tolerance«).

Zur Vermeidung von lokalen Minima rechnet Latent GOLD automatisch mit multiplen zufälligen Startwerten. Die Standardeinstellung sind 10 Versuche. Nach 50 Iterationen wird die beste Lösung ausgewählt und mit dieser weitergerechnet. Die Zahl der Startwerte und die Zahl der Iterationen, mit denen alle Lösungen gerechnet werden, kann im Menüpunkt »Technical« geändert werden. Wir empfehlen mitunter die Zahl zur Prüfung der Stabilität zu erhöhen. Bei einer erneuten Durchführung der Berechnungen kann der Fall eintreten, dass die Ergebnisse nicht exakt reproduziert werden. Dies ist ein Hinweis, dass ein lokales Minimum gefunden wurde. Die Ergebnisse unterscheiden sich in der LL-Funktion aber oft nur geringfügig, die Klassenprofile kaum. Ursache hierfür ist, dass die Schätzfunktion stark zerklüftet ist und auch in der Nähe des globalen Minimums mehrere kleine Täler vorliegen, in denen der Algorithmus verharrt (siehe

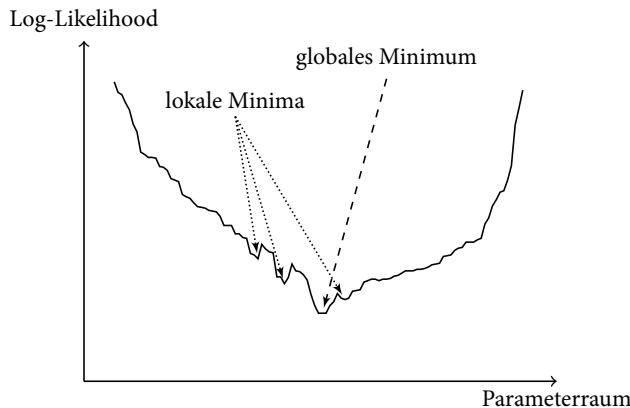


Abb. 16.4: Zerklüftete Log-Likelihood-Funktion mit lokalen Minima nahe dem globalen Minimum (fiktive Werte)

Abbildung 16.4 auf der nächsten Seite). Glatter verlaufende Schätzfunktionen können möglicherweise mit Bayes-Schätzverfahren (siehe Abschnitt 17.1) erzeugt werden. Eine für eine bestimmte Clusterlösung ermittelte Lösung kann dadurch reproduziert werden, dass der entsprechende Startwert des Zufallzahlengenerators, der von Latent GOLD dokumentiert wird, bei der erneuten Berechnung eingegeben und die Zahl der Versuche gleich 1 gesetzt wird.

16.4 Statistiken zur Modellanpassung

Ist die Parameterschätzung für ein bestimmtes Modell bzw. für mehrere Modelle durchgeführt worden, so müssen diese begutachtet werden. Latent GOLD stellt dem Anwender zu diesem Zweck eine Reihe von Statistiken zur Verfügung, anhand derer Aussagen über die Geltung und Güte eines geschätzten Modells gemacht werden können (Vermunt und Magidson 2005b, S. 58–76). Grundlegende Informationen sind hierbei die Fallzahl (n), die Anzahl der geschätzten Parameter (m_k) und Informationen über die (pseudo-)zufälligen Startwerte der Berechnung. Der Output⁵ umfasst für ein berechnetes Modell des Weiteren die geschätzten Klassengrößen $\hat{p}(x)$, die klassenspezifischen Antwortwahrscheinlichkeiten $\hat{\pi}_{m|t,x}$ für nominale und ordinale Variablen, darüber hinaus die klassenspezifischen Mittelwerte $\sum_{m=1}^{M_t} y_m^{t^*} \hat{\pi}_{m|t,x}$ für ordinale Variablen und die klassenspezifischen Mittelwerte $\hat{\mu}_{t,x}$ für kontinuierliche Antwortvariablen sowie die klassenspezifischen Mittelwerte $\hat{E}(z_{ir}|x)$ und Wahrscheinlichkeiten $\hat{p}(z_{ir} = a|x)$ der Kovariaten. Daneben werden die

⁵ Zu weiteren, hier nicht dokumentierten Maßzahlen siehe Vermunt und Magidson (2005b).

geschätzten Werte für β und γ der linearen Prädiktoren η , die Schätzungen der Fehlervarianzen und -kovarianzen σ und deren zugehörige asymptotischen Standardfehler $\widehat{se}(\beta)$, $\widehat{se}(\gamma)$ und $\widehat{se}(\sigma)$ ausgegeben.

Ob bestimmte Parameter signifikant sind, lässt sich mit Hilfe eines *Wald- χ^2 -Tests* (W^2) überprüfen. Außerdem kann für jeden Indikator der Anteil der *erklärten Varianz* $R_{y_i}^2$ bestimmt werden. Für die diskreten Variablen werden auch die geschätzten und die beobachteten Zellhäufigkeiten \hat{m}_{i*} und n_{i*} sowie die *standardisierten Residuen* \hat{r}_{i*} berichtet. Die Wald- χ^2 -Statistik (W^2) für einen Parameter ϑ_h ist definiert als:

$$W^2(\hat{\vartheta}_h) = \left(\frac{\hat{\vartheta}_h}{\widehat{se}(\vartheta_h)} \right)^2,$$

mit $\widehat{se}(\vartheta_h)$ als Standardfehler für den geschätzten Parameter ϑ_h . Latent GOLD bietet unterschiedliche Schätzmethoden für die Standardfehler an. Für ein Set von Parametern $\hat{\vartheta} = (\hat{\vartheta}_1, \hat{\vartheta}_2, \dots, \hat{\vartheta}_H)$ wird die Wald- χ^2 -Teststatistik (W^2) wie folgt berechnet:

$$W^2 = (\mathbf{c}' \hat{\vartheta})' (\mathbf{c}' \widehat{\sum}(\vartheta) \mathbf{c})^{-1} (\mathbf{c}' \hat{\vartheta}),$$

wobei \mathbf{c}' ein Zeilenvektor ist, der nur aus Nullen und Einsen besteht. Eine Eins an einer bestimmten Stelle bedeutet, dass der entsprechende Parameter für die Testung ausgewählt wird. Für das getestete Set der linearen Restriktionen gilt als Nullmodell: $\mathbf{c}' \vartheta = 0$. Die Wald-Statistik ist χ^2 -verteilt mit der Anzahl der Restriktionen (Anzahl der Einsen) als Anzahl der Freiheitsgrade df.

Darüber hinaus können bei dem LC-Cluster-Modell drei Gruppen von Prüfstatistiken verwendet werden, die auch einen Vergleich von Modellen – beispielsweise mit unterschiedlichen Klassenanzahlen – ermöglichen: 1) *χ^2 -Statistiken*, 2) *Log-Likelihood-Statistiken* und 3) *Klassifikations-Statistiken*. Neben diesen Prüfstatistiken soll im Weiteren noch auf das Verfahren des *parametrischen Bootstrap* im Hinblick auf Signifikanztests und die Analyse der *bivariaten Residuen* eingegangen werden.

16.4.1 χ^2 -Statistiken

Bei *kategorialen*, nicht jedoch bei *kontinuierlichen* Indikatorvariablen, können die *Likelihood-Ratio- χ^2 -Statistik* (L^2), die *Pearson- χ^2 -Statistik* (χ^2) und die *Cressie-Reed- χ^2 -Statistik* (CR^2) verwendet werden. Dazu werden zunächst die empirischen Auftrittshäufigkeiten n_{i*} aller I^* eindeutigen Datenmuster i^* berechnet. Ein Datenmuster i^* ist dadurch bestimmt, dass alle ihm angehörigen Fälle $i \in i^*$ dieselben Ausprägungen in den Kovariaten und Indikatoren haben. Werden die Fälle mit w_i gewichtet, gilt: $n_{i^*} = \sum_{i \in i^*} w_i$. Benötigt

werden des Weiteren die erwarteten Zellhäufigkeiten \widehat{m}_{i^*} für die Datenmuster i^* , die sich wie folgt bestimmen lassen:

$$\widehat{m}_{i^*} = n_{u_{i^*}} \hat{f}(\mathbf{y}_{i^*} | \mathbf{z}_{i^*}) ,$$

wobei $n_{u_{i^*}}$ die Gesamtzahl der Fälle mit demselben Kovariatenmuster wie Datenmuster i^* ist und $\hat{f}(\mathbf{y}_{i^*} | \mathbf{z}_{i^*})$ die bedingte multinomiale Wahrscheinlichkeit für Datenmuster i^* bei dem entsprechenden Kovariatenmuster u_{i^*} .

Unter den genannten Annahmen und Bezeichnungen berechnet Latent GOLD die oben aufgelisteten χ^2 -Statistiken wie folgt:

$$L^2 = 2 \sum_{i^*=1}^{I^*} n_{i^*} \log \frac{n_{i^*}}{\widehat{m}_{i^*}} , \quad (16.5)$$

$$\chi^2 = \sum_{i^*=1}^{I^*} \frac{(n_{i^*})^2}{\widehat{m}_{i^*}} - n , \quad (16.6)$$

$$CR^2 = 1,8 \cdot \sum_{i^*=1}^{I^*} n_{i^*} \left(\left(\frac{n_{i^*}}{\widehat{m}_{i^*}} \right)^{2/3} - 1 \right) . \quad (16.7)$$

Die Anzahl der Freiheitgrade df ist:

$$df = \min \left(\sum_{u=1}^U \left(\prod_{t=1}^{T_u^*} M_{ut}^* - 1 \right), n \right) - m_k ,$$

wobei T_u^* die Gesamtzahl der beobachteten Indikatoren im Kovariatenmuster u darstellt und M_{ut}^* die Anzahl der Kategorien des t -ten beobachteten Indikators, der zu Kovariatenmuster u gehört. Der Ausdruck »min()« soll ausdrücken, dass die Freiheitsgrade df auf der Stichprobengröße n beruhen, wenn die Anzahl der unabhängigen Zellen in der hypothetischen Kreuztabelle größer ist als die Stichprobengröße. Mit Hilfe der χ^2 -Werte und den entsprechenden Freiheitsgraden können asymptotische p-Werte bestimmt werden, die Aufschluss darüber geben, ob ein Modell zu den Daten passt. Erwünscht sind nicht signifikante Ergebnisse. Der p-Wert sollte daher größer dem vorgegebenen Schwellenwert von zum Beispiel 0,05 sein. Everitt (1988) hat in einer Simulationsstudie darauf hingewiesen, dass die asymptotischen Annahmen dieses Tests jedoch nicht immer erfüllt sind. Hoijtink (2001) verweist darauf, dass der Test sensibel gegenüber Ausreißern ist und schlägt die Verwendung eines Pseudo-Likelihood-Ratio-Tests vor, bei dem nicht ein gesamtes Datenmuster betrachtet wird sondern immer nur Itempaare. Latent GOLD bietet auch die Möglichkeit, p-Werte über das *Bootstrap-Verfahren* zu bestimmen (siehe Abschnitt 16.4.4). Diese Option bietet sich insbesondere bei wenig besetzten Kreuztabellen an, bei denen die Annahme der χ^2 -Verteilung verletzt ist.

Die Likelihood-Ratio- χ^2 -Statistik L^2 beziffert das Ausmaß der durch das Modell *nicht* erklärten Beziehungen zwischen den Variablen. Je größer L^2 also ist, desto schlechter passt das Modell auf die Daten. Als eine Faustregel kann ein Modell als gut passend angesehen werden, wenn L^2 nicht wesentlich höher liegt als die Anzahl der Freiheitsgrade df (Vermunt und Magidson 2000, S. 61). Die Pearson- χ^2 - und die Cressie-Reed- χ^2 -Statistik stellen Alternativen zu der Likelihood-Ratio- χ^2 -Statistik dar und können analog interpretiert werden.

Basierend auf der Likelihood-Ratio- χ^2 -Statistik (Gleichung 16.5), stellt Latent GOLD die schon bekannten Informationskriterien **BIC** (*Bayesian Information Criterion*, Schwarz 1978), **AIC** (*Akaike's Information Criterion*, Akaike 1973), **AIC₃** (*Akaike's Information Criterion 3*, Bozdogan 1993) und **CAIC** (*Consistent Akaike Information Criterion*, Akaike 1987; Bozdogan 1987) zur Verfügung. Sie berücksichtigen als Erweiterung die Anpassungsgüte und – über die Freiheitsgrade bzw. Anzahl der Modellparameter – die *Sparsamkeit* eines Modells, wobei für alle vier Indizes gilt: Je niedriger die jeweiligen Werte sind, desto besser ist das Modell. Sie berechnen sich folgendermaßen (Vermunt und Magidson 2005b, S. 60) :

$$BIC_{L^2} = L^2 - df \cdot \log n , \quad (16.8)$$

$$AIC_{L^2} = L^2 - 2 \cdot df , \quad (16.9)$$

$$AIC_3{}_{L^2} = L^2 - 3 \cdot df , \quad (16.10)$$

$$CAIC_{L^2} = L^2 - df \cdot (\log n + 1) . \quad (16.11)$$

Diese Informationskriterien können alternativ auch auf Grundlage der Log-Likelihood-Statistik angegeben werden (siehe unten). Beide Berechnungsverfahren führen zu ähnlichen Ergebnissen im Hinblick auf die Beurteilung bei einem Modellvergleich, da die Unterschiede zwischen den Werten für unterschiedliche Modelle gleich sein sollten. Liegen die Freiheitsgrade allerdings extrem hoch, kann die Berechnung auf L^2 -Grundlage zu nicht verwertbaren Ergebnissen führen, dann muss auf die alternative Berechnung zurückgegriffen werden.

Über diese Informationskriterien hinaus kann als beschreibendes Maß ein Unähnlichkeitsindex »dissimilarity index« (DI) auf Grundlage der χ^2 -Statistik berechnet werden:

$$DI = \frac{\left(\sum_{i^*=1}^{I^*} |n_{i^*} - \hat{m}_{i^*}| \right) + \left(n - \sum_{i^*=1}^{I^*} \hat{m}_{i^*} \right)}{2n} .$$

Dieses Maß zeigt an, wie stark beobachtete und geschätzte Zellhäufigkeiten voneinander abweichen, was den Anteil der Stichprobe beziffert, der für eine perfekte Modellanpassung verändert werden müsste.

16.4.2 Log-Likelihood-Statistiken

Im Rahmen der *Log-Likelihood-Statistik* (LL-Statistik) stellt die Software als Output zunächst die für die oben erläuterte Parameterschätzung wichtigen Größen *Log-Likelihood* (LL), *Log-Prior* ($\log p(\boldsymbol{\vartheta})$) und *Log-Posterior* ($\log \mathcal{P}$) zur Verfügung:

$$\text{LL} = \sum_{i=1}^I w_i \log \hat{f}(\mathbf{y}_i | \mathbf{z}_i) ,$$

$$\log \mathcal{P} = \text{LL} + \log p(\hat{\boldsymbol{\vartheta}}) .$$

Zusätzlich zu den Informationskriterien, die in den Gleichungen 16.8 bis 16.11 vorgestellt wurden, wird für diese Indizes eine alternative Berechnungsmöglichkeit auf Grundlage der Log-Likelihood angeboten und ausgegeben, die auch in dem Fall von kontinuierlichen Indikatorvariablen zur Verfügung steht:

$$\text{BIC}_{\text{LL}} = -2\text{LL} + m_k (\log n) , \quad (16.12)$$

$$\text{AIC}_{\text{LL}} = -2\text{LL} + 2m_k , \quad (16.13)$$

$$\text{AIC3}_{\text{LL}} = -2\text{LL} + 3m_k , \quad (16.14)$$

$$\text{CAIC}_{\text{LL}} = -2\text{LL} + m_k ((\log n) + 1) . \quad (16.15)$$

Auch hier kann das Bootstrap-Verfahren eingesetzt werden, um einen p-Wert zu bestimmen. Dieser bezieht sich dann auf den -2LL -Differenzen-Test zwischen zwei Modellen und kann dahingehend interpretiert werden, ob ein Modell im Vergleich mit einem anderen signifikant besser ist (siehe Abschnitt 14.2.1). Verglichen werden kann ein Modell sowohl mit dem Nullmodell als auch mit einem beliebigen anderen Modell mit weniger Parametern.

16.4.3 Klassifikations-Statistiken

Im Bereich der Klassifikations-Statistiken werden Informationen zur Verfügung gestellt, die Aufschluss über die Klassenzugehörigkeitswahrscheinlichkeiten der einzelnen Fälle geben. Wie gut kann die Zugehörigkeit zu einer bestimmten latenten Klasse vorhergesagt werden, bei Kenntnis der beobachteten Werte der Indikatoren \mathbf{y} und Kovariaten \mathbf{z} ? Sind die Klassen trennscharf oder überlappen sie sich stark? Für das Antwortmuster i lassen sich diese (*A-posteriori*-)Klassenzugehörigkeitswahrscheinlichkeiten folgendermaßen bestimmen:

$$\hat{p}(x|\mathbf{z}_i, \mathbf{y}_i) = \frac{p(x|\mathbf{z}_i) \hat{f}(\mathbf{y}_i|x, \mathbf{z}_i)}{\hat{f}(\mathbf{y}_i|\mathbf{z}_i)} .$$

Zähler und Nenner stellen die Maximum-Likelihood-Schätzer für die entsprechenden Terme aus Gleichung 16.1 auf Seite 399 dar. Aus diesen Größen können der *Anteil der Klassifikationsfehler* (E) und drei R^2 -*Maßzahlen* für nominale Variablen berechnet werden:⁶ die proportionale Reduktion der Klassifikationsfehler $R_{x,\text{errors}}^2$, ein auf der Entropie basierendes $R_{x,\text{entropy}}^2$ und ein auf qualitativer Varianz basierendes Maß $R_{x,\text{variance}}^2$.

Der *Anteil der Klassifikationsfehler* E wird als

$$E = \frac{\sum_{i=1}^n w_i (1 - \max \hat{p}(x|\mathbf{z}_i, \mathbf{y}_i))}{n}$$

definiert, wobei $\sum w_i$ gleich n ist. Bei einer eindeutigen Zuordnung aller Fälle zu nur einem Cluster ist $\max \hat{p}(x|\mathbf{z}_i, \mathbf{y}_i)$ gleich 1 und der Klassifikationsfehler daher 0.

Allgemein gilt für das R^2 -*Maß*:

$$R_x^2 = \frac{\text{error}(x) - \text{error}(x|\mathbf{z}, \mathbf{y})}{\text{error}(x)},$$

mit $\text{error}(x)$ als Gesamtfehler bei der Vorhersage von x ohne jede Information von \mathbf{z} und \mathbf{y} . Der Fehler $\text{error}(x|\mathbf{z}, \mathbf{y})$ ist hingegen derjenige, der sich ergibt, wenn alle beobachteten Informationen einbezogen werden. Er wird definiert als gewichteter Durchschnitt der fallspezifischen Fehler $\text{error}(x|\mathbf{z}_i, \mathbf{y}_i)$:

$$\text{error}(x|\mathbf{z}, \mathbf{y}) = \frac{\sum_{i=1}^n w_i \text{error}(x|\mathbf{z}_i, \mathbf{y}_i)}{n}.$$

Die drei oben erwähnten R^2 -*Maßzahlen* unterscheiden sich in ihrer Definition von $\text{error}(x|\mathbf{z}_i, \mathbf{y}_i)$. Im Fall von $R_{x,\text{errors}}^2$ bestimmt sich der Fehler durch

$$\text{error}(x|\mathbf{z}_i, \mathbf{y}_i) = 1 - \max \hat{p}(x|\mathbf{z}_i, \mathbf{y}_i),$$

bei $R_{x,\text{entropy}}^2$ durch

$$\text{error}(x|\mathbf{z}_i, \mathbf{y}_i) = - \sum_{x=1}^K \hat{p}(x|\mathbf{z}_i, \mathbf{y}_i) \log \hat{p}(x|\mathbf{z}_i, \mathbf{y}_i),$$

und bei $R_{x,\text{variance}}^2$ durch

$$\text{error}(x|\mathbf{z}_i, \mathbf{y}_i) = 1 - \sum_{x=1}^K (\hat{p}(x|\mathbf{z}_i, \mathbf{y}_i))^2.$$

⁶ Ähnliche Maßzahlen können auch für die Vorhersage der Klassenzugehörigkeiten anhand der Kovariaten bestimmt werden. Diese Maßzahlen sind naturgemäß nur im Fall von aktiven Kovariaten sinnvoll.

Daneben kann die sogenannte *Klassifikations-Log-Likelihood* ($\log L^c$) ausgegeben werden:

$$\log L^c = \sum_{i=1}^n \sum_{x=1}^K \hat{w}_{xi} \log \hat{p}(x|z_i) \hat{f}(y_i|x, z_i).$$

Schließlich wird darauf aufbauend noch das *durchschnittliche Maß an Bedeutung* (»Average Weight of Evidence«, AWE) zur Verfügung gestellt, welches ähnlich dem BIC interpretiert werden kann (Banfield und Raftery 1993):

$$AWE = -2 \log L^c + 2 \left(\frac{2}{3} + \log n \right) m_k.$$

Hier gilt, ebenso wie bei den oben genannten Informationskriterien, je niedriger der Wert für das AWE ist, desto besser ist das jeweilige Modell.

Die mit den Klassifikations-Statistiken untersuchte Fragestellung ist vergleichbar mit jener der Überlappungsindizes, nämlich wie eindeutig die Objekte den Klassen zugeordnet sind. So zum Beispiel kommt auch im Index von DUNN der Ausdruck $(\hat{p}(x|z_i, y_i))^2$ vor (siehe Abschnitt 14.4).

16.4.4 Signifikanztests mit parametrischem Bootstrap

Wie oben schon angedeutet, bietet sich zur Bestimmung von p-Werten im Hinblick auf Signifikanztests die Option eines *parametrischen Bootstraps* an, wenn das Zutreffen der Verteilungsannahmen in Zweifel gezogen wird. Dies ist bei den oben genannten χ^2 -Statistiken und dem Test, ob ein Modell in Gänze zu den Daten passt, im Allgemeinen bei wenig besetzten Kreuztabellen der Fall. Angezeigt wird die Problematik durch negative Freiheitsgrade.

Im Rahmen der *Log-Likelihood-Statistiken* und dem LQ-Vergleichstest zwischen zwei Modellen mit unterschiedlicher Klassenanzahl kann ebenfalls nicht davon ausgegangen werden, dass diese Teststatistiken unter Gültigkeit der H_0 -Hypothese approximativ χ^2 -verteilt sind (McLachlan und Peel 2000; Nylund, Asparouhov u. a. 2007, S. 542, 543). Alternativ wurde von Lo, Mendell u. a. (2001) der sogenannte *Lo-Mendell-Rubin-Likelihood-Ratio-Test* (LMR-LRT) entwickelt, der eine approximative Likelihood-Ratio-Test-Verteilung nutzt, um die Verbesserung in der Anpassungsgüte zwischen zwei benachbarten Modellen zu testen. Dieser Test ist eher im Bereich des »Growth Mixture Modeling«⁷ (GMM) für Mischungen von normalverteilten, kontinuierlichen Variablen üblich und dort

⁷ Wachstumsmodelle (Reinecke 2005, S. 304–336) oder »Growth Mixture Models« stellen eine Kombination aus Strukturgleichungsmodellen und latenten Klassenanalysen dar.

getestet worden (Muthén 2004, S. 356, 357). Im Bereich von latenten Klassenanalysen ist er jedoch bislang nicht ausreichend geprüft worden, um ihn empfehlen zu können (Nylund, Asparouhov u. a. 2007, S. 538). Kritisch zu den asymptotischen Eigenschaften äußert sich auch Jeffries (2003), sie sind seiner Ansicht nach nur unter sehr restriktiven Bedingungen erfüllt. Zudem schneidet der LMR-LRT-Test im Vergleich mit dem hier vorgestellten Bootstrap-Verfahren schlechter ab (Nylund, Asparouhov u. a. 2007).

Die Idee des Bootstraps besteht darin, nicht von einer gegebenen Verteilung auszugehen, der die gesuchten Größen folgen, sondern aus den Daten (durch wiederholtes Stichprobenziehen mit »Zurücklegen« aus ein und demselben Datensatz) selbst die Verteilung zu bestimmen. Auf dieser Grundlage können dann p-Werte bestimmt werden, die Aufschluss über Signifikanzen geben.

Bei den p-Werten, die für die oben genannten χ^2 -Statistiken ausgegeben werden, wird beim Bootstrap-Verfahren nicht nur das spezifizierte Modell für den zugrunde liegenden Datensatz berechnet, sondern zusätzlich für eine Anzahl B an Replikationsstichproben, die wiederholt mit »Zurücklegen« aus den ursprünglichen Daten gezogen werden. Der zu berechnende Boostrap-p-Wert \hat{p}_{boot} ist definiert als der Anteil an Bootstrap-Stichproben mit einem größeren L^2 -Wert als der ursprüngliche Datensatz. Der Standardfehler von \hat{p}_{boot} entspricht

$$\sqrt{\frac{\hat{p}_{\text{boot}}(1 - \hat{p}_{\text{boot}})}{B}}.$$

Die Genauigkeit kann naturgemäß durch eine Erhöhung der Replikationszahl erreicht werden, wodurch sich jedoch auch die Rechenzeit – teils erheblich – erhöht (Vermunt und Magidson 2005b, S. 54). Stehen die χ^2 -Statistik und die darauf basierenden p-Werte nicht zur Verfügung, so entfällt auch die Möglichkeit, per Bootstrap diesbezügliche p-Werte zu berechnen. Dies ist der Fall, sobald eine oder mehrere *kontinuierliche* Indikatorvariablen verwendet werden, hier stehen aber die Log-Likelihood-Statistiken für die Signifikanztestung zur Verfügung.

Im Falle des LQ- bzw. -2LL -Differenzentests, bei dem die Log-Likelihood zweier Modelle mit unterschiedlichen Klassenanzahlen gegeneinander auf signifikante Unterschiede bzw. Verbesserungen getestet werden (siehe Abschnitt 14.2.1), zieht man ebenfalls B Replikationsstichproben auf Grundlage der durch die Maximum-Likelihood-Schätzer definierten Wahrscheinlichkeitsverteilung. Die -2LL -Differenz ist allgemein definiert als $-2 \cdot (\text{LL}_{H_0} - \text{LL}_{H_1})$, wobei das Modell H_0 das restriktivere Modell darstellt (beispielsweise mit $K - 1$ Klassen) und das Modell H_1 das allgemeinere (beispielsweise mit K Klassen). Die Replikationsstichproben werden unter der angenommen Gültigkeit des restriktiveren Modells H_0 gezogen und der berechnete p-Wert \hat{p}_{boot} ist dann als Anteil der Bootstrap-Stichproben definiert, die eine größere -2LL -Differenz aufweisen als die Originalstichprobe (Vermunt und Magidson 2005b, S. 54). Der -2LL -Differenztest

auf Bootstrap-Grundlage kann auch bei kontinuierlichen Indikatorvariablen verwendet werden.

16.4.5 Bivariate Residuen

Eine interessante Option bietet auch die Analyse der *bivariaten Residuen*. Eine der grundlegenden Annahmen des LC-Cluster-Modells ist die Annahme der lokalen Unabhängigkeit. Passt ein Modell mit einer bestimmten Klassenzahl nicht zu den Daten, beruht das meist auf der Verletzung dieser Annahme (Vermunt und Magidson 2005b, S. 24). Übliche Vorgehensweisen in solch einem Fall sind die Erweiterung um eine oder mehrere weitere Klassen oder die Eliminierung eines oder mehrerer Items, bis die lokale Unabhängigkeit erreicht ist. Latent GOLD ermöglicht allerdings auch eine alternative Verfahrensweise zur Verbesserung der Modellanpassung, die in der Zulassung von Beziehungen zwischen den Indikatoren bzw. direkten Effekten der Kovariaten auf die Indikatoren besteht (Magidson und Vermunt 2004, S. 182, 183). Das Programm berechnet die sogenannten bivariaten Residuen für $(y-y)$ - und $(z-y)$ -Paare, die anzeigen können, welche Beziehungen zwischen zwei Variablen durch das formulierte Modell nicht ausreichend erklärt werden. Informationen über diese Residuen können zur Beurteilung der Klassenanzahl eines Modells herangezogen werden. So sollten bei einem adäquaten Modell die paarweisen Residuen nicht größer als 3,84 sein, will man an der lokalen Unabhängigkeit festhalten. Treten größere Werte auf, besteht eine Lösungsmöglichkeit darin, die durch die Residuen angezeigten Beziehungen freizusetzen und die Annahme der lokalen Unabhängigkeit zu lockern. Die bivariaten Residuen sind definiert als:

$$\text{BR} = \frac{1}{P} \sum_{p=1}^P \left(\frac{\partial \log \mathcal{P}}{\partial \vartheta_p^{\text{local}}} \right)^2 \Bigg/ \left(\frac{\partial^2 \log \mathcal{P}}{\partial^2 \vartheta_p^{\text{local}}} \right),$$

wobei P die Anzahl der Parameter einer bestimmten lokalen Abhängigkeit ist, die durch $\vartheta_p^{\text{local}}$ bezeichnet wird. Für jeden der P Parameter werden nun die ersten und zweiten Ableitungen der Log-Posterior-Funktion $\log \mathcal{P}$ berechnet. Aufgrund des Quotienten $1/P$ kann das Maß als erwartete Modellverbesserung pro zusätzlichem Parameter interpretiert werden.

Die bivariaten Residuen können als untere Schranke für die Verbesserung der Modellanpassung (L^2 oder -2LL), die sich bei Freisetzung der jeweiligen Beziehungen (Aufgabe der Annahme der lokalen Unabhängigkeit) ergäbe, interpretiert werden (Vermunt und Magidson 2005b, S. 72–74). Insofern kann man sie auch als Modifikationsindizes betrachten, wie sie beispielsweise bei Strukturgleichungsmodellen bekannt sind (Reinecke 2005). Sie sollten nicht größer 3,84 sein. Dieser Schwellenwert ergibt sich aus folgender Überlegung:

Die durch die Residuen gemessene Verbesserung der LL-Funktion ist approximativ χ^2 -verteilt mit einem Freiheitsgrad. Der Schwellenwert für die χ^2 -Verteilung mit einem Freiheitsgrad für $p < 0,05$ ist gleich 3,84 (Vermunt und Magidson 2005a, S. 125).

16.4.6 Beurteilung und Auswahl von Modellen

Es wird wie bisher üblich vorgegangen. Zur Bestimmung der Klassenzahl definiert der Anwender eine Spannbreite, wobei auf jeden Fall das 1-Klassenmodell enthalten sein soll. Zur Auswahl eines geeigneten Modells wird auf die oben vorgestellten Statistiken und Informationskriterien zurückgegriffen. Üblicherweise wird hierbei der BIC, seltener AIC und CAIC herangezogen. Neueste Untersuchungen legen nahe, dass der AIC₃ das am geeignetste Kriterium darstellt, wenn die Klassenanzahl bestimmt werden soll (Andrews und Currim 2003; Dias 2004). Fonseca und Cardoso (2007) verweisen in ihrem Überblicksaufsatzz darauf, dass diese gute Performanz für AIC₃ für nominalskalierte Variablen gilt. Für quantitative Variablen erscheint dagegen BIC besser geeignet zu sein. Bei gemischten Variablen schneidet ICL-BIC⁸ (siehe Gleichung 15.8 auf Seite 384) besser ab. Eine Monte-Carlo-Simulationsstudie von Nylund, Asparouhov u. a. (2007) nennt den LQ-Test auf Grundlage von Bootstrap-p-Werten als bestes Entscheidungskriterium.

Das ausgewählte Modell wird auf inhaltliche Interpretierbarkeit geprüft, das heißt, es wird untersucht, ob den Klassen sinnvolle Namen gegeben werden können. Schließlich wird die formale und inhaltliche Gültigkeit untersucht. Die Vorgehensweise zur Lösung einer Klassifikationsaufgabe wird in der nachfolgenden Box zusammengefasst:

1. Berechnen einer Palette von Modellen mit unterschiedlicher Klassenzahl, die inhaltlich bzw. theoretisch angemessen erscheinen. Das 1-Klassenmodell sollte auf jeden Fall mituntersucht werden.
2. Vergleich der Modelle mit unterschiedlichen Klassenzahlen:
 - Als Entscheidungskriterium für die Wahl zwischen mehreren Modellen sollten die Informationskriterien BIC, AIC und AIC₃ eingesetzt werden. Der jeweils niedrigste Wert zeigt das am besten passende Modell an.
 - Als weitere wichtige Maßzahl können die prozentuellen Verbesserungen PVK verwendet werden.
 - Die Entscheidung kann zusätzlich durch L²- oder LL-Statistiken abgesichert werden.

⁸ Diese Maßzahl ist derzeit noch nicht in Latent GOLD implementiert, kann aber aus den Ergebnissen leicht berechnet werden.

- Bestehen Zweifel, dass die L²-Statistik χ^2 -verteilt ist, soll die Bootstrap-Option genutzt werden, um aussagekräftige p-Werte zu erhalten. Bei der LL-Statistik soll immer auf die Bootstrap-Werte zurückgegriffen werden.
 - Mit Hilfe des LQ-Tests auf Grundlage von Bootstrap-p-Werten können zwei Modelle gegeneinander auf signifikante Verbesserungen getestet werden.
3. Hat man sich auf dieser Grundlage für ein oder mehrere Modelle entschieden, kann die Signifikanz der einzelnen Indikatoren mit Hilfe des Wald-Tests geprüft werden. Die zugehörigen p-Werte sollten unter 0,05 oder einem anderen vorab definierten Schwellenwert liegen. Daneben kann man die erklärte Varianz (R^2) der einzelnen Indikatoren begutachten.
 4. Die bivariaten Residuen zwischen den Indikatoren und zwischen Kovariaten und Indikatoren sollten jeweils nicht größer 3,84 sein, sonst klärt das Modell die vorliegenden Beziehungen nur unzureichend auf. Gelöst werden kann das Problem entweder durch die Erhöhung der Klassenzahl oder die Freisetzung der jeweiligen Beziehungen.
 5. Die Zuordnungswahrscheinlichkeiten, Klassengrößen und -profile geben weiteren Aufschluss über die Modelle. Um eine endgültige Auswahl zu treffen, muss die Lösung zunächst inhaltlich interpretierbar sein.
 6. Des Weiteren sollte die inhaltliche Gültigkeit durch Spezifikation von Hypothesen untersucht werden. Sofern möglich, kann die inhaltliche Gültigkeitsprüfung direkt mit Latent GOLD durchgeführt werden, indem die bei der inhaltlichen Gültigkeitsprüfung verwendeten Variablen als Kovariaten in die Analyse einbezogen werden.
 7. Schließlich sollte die Stabilität und formale Gültigkeit begutachtet werden (siehe Kapitel 20).

Nochmals sei darauf hingewiesen, dass die inhaltlichen Kriterien ein Abweichen von den statistischen Kriterien rechtfertigen können, gerade wenn sich die Kennzahlen zwischen mehreren Modellen nur leicht unterscheiden. Ein statistisch nicht passendes Modell sollte allerdings in jedem Fall verworfen werden.

16.5 Ein Anwendungsbeispiel

16.5.1 Kontinuierliche Daten (latente Profilanalyse)

Analysiert werden analog zur Vorgehensweise in Abschnitt 14.2.1 die kontinuierlichen (quantitativen) Variablen mit den Gesamtpunktwerten für Materialismus (gMat) und

Tab. 16.8: Modellprüfgrößen für die LC-Cluster-Analyse der (kontinuierlichen) Wertedaten von Denz (Latent GOLD 4.5)

K	LL	PV_{OK} (%)	PV_K (%)	BIC_{LL}	AIC_{LL}	$AIC3_{LL}$	m_k	E
1	-374,1003	—	—	769,7932	756,2006	760,2006	4	0,0000
2	-330,4181	11,7	11,7	709,4196	678,8361	687,8361	9	0,1177
3	-316,1021	15,5	4,3	707,7784	660,2041	674,2041	14	0,2431
4	-303,6723	18,8	3,9	709,9096	645,3445	664,3445	19	0,2258
5	-292,8231	21,7	3,6	715,2021	633,6462	657,6462	24	0,2058
6	-287,6758	23,1	1,8	731,8983	633,3516	662,3516	29	0,1824
7	-279,4906	25,3	2,8	742,5187	626,9811	660,9811	34	0,1955
8	-277,0893	25,9	0,9	764,7069	632,1785	671,1785	39	0,2198
9	-273,0174	27,0	1,5	783,5540	634,0348	678,0348	44	0,2035
10	-271,5169	27,4	0,5	807,5438	641,0338	690,0338	49	0,2162

grau hinterlegt: Auf Grundlage von BIC, AIC und AIC3 ausgewählte Lösungen.

Anmerkung: Es wurde mit 100 zufälligen Startpartitionen gerechnet.

Postmaterialismus (gPmat), die einen Wertebereich von 1 (»sehr wichtig«) bis 5 (»sehr unwichtig«) aufweisen. Gerechnet werden die 1- bis 10-Klassenlösungen, deren Modellprüfgrößen in Tabelle 16.8 dargestellt sind.

Da mit kontinuierlichen Variablen gerechnet wird, stehen die χ^2 -Statistik und die darauf basierenden p-Werte nicht zur Verfügung. Die ausgewiesenen Informationskriterien werden auf Grundlage der Log-Likelihood berechnet. Der von Latent GOLD berechnete Wert der Log-Likelihood-Funktion stimmt beim 1-Klassenmodell mit dem in Tabelle 14.3 auf Seite 363 aufgelisteten, mit ALMO berechneten Wert überein (-374,1003). Die Werte für die weiteren Modelle weichen geringfügig voneinander ab, wobei die Unterschiede sich im Bereich von unter 5 Prozent der Gesamtwerte bewegen. Beispielsweise berechnet ALMO einen LL-Wert von -277,608 für das 10-Klassenmodell, Latent GOLD dagegen einen Wert von -271,5169. Die Abweichungen sind zum Teil durch die in Latent GOLD enthaltene PM-Schätzung bedingt, die »nicht-definierte« Randlösungen, bei der Parameterschätzungen von Wahrscheinlichkeiten von 0 bzw. 1 auftreten, vermeidet. Die Log-Likelihood sinkt von der 1- zur 10-Klassenlösung betragsmäßig monoton ab, was ein Hinweis ist, dass bei keiner Zahl ein lokales Minimum gefunden wird. Ein lokales Minimum für eine Klassenzahl führt zu einem Anstieg des LL-Wertes. Allerdings kann auch bei kontinuierlich sinkenden Log-Likelihood-Werten nicht ausgeschlossen werden, dass lokale Minima ermittelt wurden. Dies ist in dem Beispiel für einige Klassenlösungen der Fall. Wird das Programm erneut gerechnet, ergeben sich leicht abweichende Werte, da der Zufallszahlengenerator andere zufällige Startwerte erzeugt. Dies ist zum Beispiel für die 6-Klassenlösung der Fall, mit dem Effekt, dass die 7-Klassenlösungen teilweise signifikante Verbesserungen gegenüber der 6-Klassenlösung erzielt und zum Teil insignifikante (siehe dazu später).

Die *Informationskriterien* sind uneindeutig: Während der BIC das Minimum bereits bei der 3-Klassenlösung erreicht, zeigen AIC und AIC₃ mit ihrem Minimum die 7- bzw. 5-Klassenlösung als am besten passend an. Die Informationskriterien bzw. deren Minima berechnen sich nach den Gleichungen 16.12 bis 16.14 folgendermaßen:

$$\text{BIC}_{\text{LL},3} = -2\text{LL} + m_k(\log n) = -2 \cdot -316,1021 + 14 \cdot \log(221) = 707,7785 ,$$

$$\text{AIC}_{\text{LL},7} = -2\text{LL} + 2m_k = -2 \cdot -279,4906 + 2 \cdot 34 = 626,9811 ,$$

$$\text{AIC}_{\text{LL},5} = -2\text{LL} + 3m_k = -2 \cdot -292,8231 + 3 \cdot 24 = 657,6462 .$$

Die Ergebnisse stimmen auch mit in der Literatur berichteten Eigenschaften überein, dass der AIC zu einer größeren Clusterzahl tendiert bzw. überparametrisierte Modelle bevorzugt (Davier 1997, S. 56).

Betrachtet man die *prozentuale Verbesserung* einer jeden Lösung im Hinblick auf die vorausgegangene Lösung PV_K , so lässt sich nach zwei und fünf Klassen ein deutlicher Abfall beobachten. Dies spricht für die 2- oder 5-Klassenlösung. Ausrechnen lässt sich PV_K mit Hilfe der Gleichung 14.12 auf Seite 364:

$$\text{PV}_K = 1 - \frac{|\text{LL}_K|}{|\text{LL}_{K-1}|} \quad \text{bzw. in Prozent: } 100 \cdot \left(1 - \frac{|\text{LL}_K|}{|\text{LL}_{K-1}|} \right) .$$

Für den Wert der 5- im Vergleich zur 4-Klassenlösung also beispielsweise:

$$100 \cdot \left(1 - \frac{292,8231}{303,6723} \right) = 3,5727 .$$

In Bezug auf die prozentuale Verbesserung zum Nullmodell PVO_K , die sich analog zu PV_K berechnen lässt, wenn im Nenner immer der LL-Wert der 1-Klassenlösung eingesetzt wird, ist festzustellen, dass die Schwelle von 20 Prozent bei der 5-Klassenlösung überschritten wird.

Der *Likelihood-Quotienten-Test* (siehe Abschnitt 14.2.1), der zwei Modelle miteinander vergleicht und die Signifikanz der Verbesserung prüft, kann auch hier über die Bootstrap-Technik angewendet werden. Es ergeben sich die in Tabelle 16.9 dargestellten Prüfgrößen. Beispielsweise ergibt sich die Größe von 75,1899 beim Test der 5- gegen die 2-Klassenlösung durch die mit -2 multiplizierte Differenz der Log-Likelihoods aus Tabelle 16.8 auf der vorherigen Seite

$$-2 \cdot (\text{LL}_2 - \text{LL}_5) = -2 \cdot (-330,4181 - (-292,8231)) = 75,1899$$

und wird mit dem Bootstrap-Verfahren auf Signifikanz geprüft. Alle für die 2- und 5-Klassenlösung getesteten Paare von Klassenlösungen weisen eine signifikante Verbesserung der Klassenlösung mit der höheren Klassenzahl gegenüber derjenigen mit

Tab. 16.9: Likelihood-Quotiententest (Bootstrap)

gegen	Test von Klassenlösung K			
	3		5	
	-2LL Diff.	p-Wert	-2LL Diff.	p-Wert
1	115,9965	0,0000	162,5543	0,0000
2	28,6320	0,0000	75,1899	0,0000
3	—	—	46,5579	0,0000
4	—	—	21,6983	0,0040

niedrigerer Klassenzahl auf. Insofern ist der Test für die 2- und 5-Klassenlösung wenig hilfreich bei der Modellwahl. Allerdings erbringt die 7-Klassenlösung gegenüber der 6-Klassenlösung nur teilweise signifikante Verbesserungen, da sich beim erneuten Durchrechnen der LL-Wert der 6-Klassenlösung ändert. Die insignifikanten Befunde dominieren dabei.

Auf der Grundlage der vorgestellten Maßzahlen scheinen das 2-, 3- und das 5-Klassenmodell infrage zu kommen. Für das 2-Klassenmodell spricht »nur« der Abfall der prozentuellen Verbesserung; das 3-Klassenmodell ist interessant, da der BIC dort sein Minimum hat. Für die 5-Klassenlösung spricht, dass der AIC₃ dort sein Minimum hat, die prozentuale Verbesserung zum jeweiligen Vormodell danach deutlich absinkt und die prozentuale Verbesserung zum Nullmodell das erste Mal den Wert von 20 Prozent übersteigt. Für die 7-Klassenlösung spricht, dass der AIC sein Minimum erreicht und anschließend ebenfalls ein Abfall in der prozentuellen Verbesserung eintritt. Allerdings erbringt die 7-Klassenlösung gegenüber der 6-Klassenlösung nur teilweise signifikante Verbesserungen. Wir wollen nachfolgend daher die 3- und 5-Klassenlösung näher in Betracht ziehen.

Tab. 16.10: Wald-Teststatistiken für die 3- und 5-Klassenlösung

	W^2	p-Wert	R^2
<i>3-Klassenlösung</i>			
gMat	33,17	0,0000	0,30
gPmat	23,18	0,0000	0,34
<i>5-Klassenlösung</i>			
gMat	303,39	0,0000	0,57
gPmat	114,18	0,0000	0,44

Abkürzungen: gMat: Gesamtpunktwerte Materialismus, gPmat: Gesamtpunktwerte Postmaterialismus

Tab. 16.11: Klassenprofile (Klassengrößen und Mittelwerte der Indikatorvariablen, $n = 221$)

	Klasse 1	Klasse 2	Klasse 3
Größe (n)	117	86	18
Größe (p_k)	0,4679	0,4060	0,1262
gMat (\bar{x})	1,7037	2,2904	2,7155
gPmat (\bar{x})	1,6450	1,3470	2,2449

Abkürzungen: siehe Tabelle 16.10 auf der vorherigen Seite

(a) 3-Klassenlösung

	Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5
Größe (n)	144	50	12	10	5
Größe (p_k)	0,5418	0,3146	0,0736	0,0465	0,0235
gMat (\bar{x})	2,0421	1,6679	2,5016	3,3507	4,2936
gPmat (\bar{x})	1,4132	1,7727	2,4612	1,0379	1,9952

Abkürzungen: siehe Tabelle 16.10 auf der vorherigen Seite

(b) 5-Klassenlösung

Der *Wald-Test* bewertet sowohl in der 3- als auch in der 5-Klassenlösung beide Indikatorvariablen als signifikant in Bezug auf ihren Beitrag zur Trennung der Klassen (siehe Tabelle 16.10 auf der vorherigen Seite). Naturgemäß ist der Anteil der erklärten Varianz (R^2) bei der 3-Klassenlösung geringer: Bei gMat werden 30 und bei gPmat 34 Prozent der Varianz erklärt. In der 5-Klassenlösung ist mit 57 (gMat) und 44 Prozent (gPmat) der jeweilige Anteil größer. R^2 ist bei quantitativen Variablen (kontinuierliche Indikatoren oder Zählvariablen) formal gleich der bei K-Means eingeführten erklärten Varianz η^2 (siehe Abschnitt 12.2).

Das *bivariate Residuum* für die 3-Klassenlösung liegt für gMat und gPmat bei 2,4010. Für die 5-Klassenlösung ist es mit einem Wert von 1,4386 deutlich geringer. Beide Werte liegen unter dem Schwellenwert von 3,84, die Annahme der lokalen Unabhängigkeit ist daher erfüllt.

Die Klassen der 3-Klassenlösung können auf Grundlage der in Tabelle 16.11 ausgewiesenen Mittelwertprofile folgendermaßen beschrieben werden:

Klasse C₁: In dieser größten Klasse ($n = 117$ bzw. 47 Prozent) findet sich der schon bekannte Konsenstypus wieder, sowohl Postmaterialismus als auch Materialismus wird zugestimmt. Dieses Profil findet sich auch in der 5-Klassenlösung wieder (dort Klasse 2).

Klasse C₂: Die Mitglieder dieser Klasse können als Postmaterialisten eingestuft werden. Sie umfasst 86 Personen (41 Prozent).

Klasse C₃: Diese mit 18 Personen (13 Prozent) kleinste Klasse wird von den Nicht-Orientierten bestimmt, die allerdings den Postmaterialismus noch etwas stärker ablehnen als den Materialismus. Auch dieser Typus findet sich in der 5-Klassenlösung wieder (dort Klasse 3).

Natürgemäß kommt die 5-Klassenlösung mit mehr Klassen auch inhaltlich zu einem differenzierteren Ergebnis. Ist der Anwenderin die 3-Klassenlösung nicht »genau« genug, so sprächen auch inhaltliche Gründe für eine differenziertere Lösung.

16.5.2 Hinzunahme von Kovariaten

In einem nächsten Schritt sollen nun zusätzlich Kovariaten in das Modell aufgenommen werden. Geprüft wird das Modell der Abbildung 16.2 auf Seite 397. Das Modell geht von der Annahme aus, dass die sozio-demographischen Variablen »Geschlecht«, »besuchter Schultyp« und »Bildung der Mutter« einen Einfluss auf die latenten Klassen haben. Daneben wird ein direkter Effekt des Geschlechts auf den Materialismus aufgenommen (Begründung für das Auftreten dieses Effekts siehe Abschnitt 16.1). Zusätzlich wurde die Wahrnehmung einer gesellschaftlichen Bedrohung⁹ als inaktive Kovariate einbezogen. Dem liegt die Hypothese zugrunde, dass sich die latenten Klassen in der Wahrnehmung einer gesellschaftlichen Bedrohung unterscheiden. Die Hypothese ist kausal nicht gerichtet, daher wurde die Variable als inaktiv gesetzt.

Die Ergebnisse der Parameterschätzung gibt Tabelle 16.12 auf Seite 433 wieder. Den Wald-Statistiken für die Variablen Materialismus (gMat: 123,3430) und Postmaterialismus (gPmat: 26,2061) ist zu entnehmen, dass die latenten Klassen signifikant getrennt werden.

Als nächstes werden die Regressionskonstanten für die Indikatoren berichtet (gMat: 2,1710 und gPmat: 1,6981), dies sind bei der gewählten Effektkodierung (entspricht der Voreinstellung) die Gesamtmittelwerte. Anschließend werden die Fehlervarianzen ausgegeben. Auffallend ist, dass das dritte Cluster mit 0,6491 und 0,4978 relativ hohe Fehlerstreuungen aufweist.

Nach der Ausgabe der Ergebnisse für die Indikatoren werden die Ergebnisse für die latenten Klassen berichtet. Von den untersuchten Variablen mit einem angenommenen direkten Einfluss ist nur der Schultyp bei einem Fehlerniveau von $p < 0,05$ signifikant.

⁹ Das Item lautet: »Unsere Gesellschaft ist durch den Mangel an menschlichen Beziehungen 1 ›etwas bedroht‹, 2 ›sehr bedroht‹, 3 ›ernstlich bedroht‹ und 4 ›in ihrer Existenz bedroht‹.«

Lässt man mit $p < 0,10$ tendenzielle Zusammenhänge zu, so wirkt auch das Geschlecht signifikant mit einem Fehlerniveau von $p = 0,081$ auf die latenten Klassen ein. Die Effekte für den Schultyp sind wie folgt zu lesen: Der Effekt von $-0,6172$ des Schultyps »BHS«¹⁰ für die erste Klasse bedeutet, dass die Wahrscheinlichkeit des Auftretens der ersten Klasse sinkt, wenn der Schüler bzw. die Schülerin in eine BHS geht. Mit einem Wert von $-1,3865$ ist auch der Effekt der Ausprägung »AHS« negativ. Die Auftrittswahrscheinlichkeit der ersten latenten Klasse reduziert sich somit ebenfalls bei einem AHS-Besuch. Sie nimmt dagegen für die dritte Ausprägung zu, da der Effekt mit $2,0037$ positiv ist. Die anderen Effekte der Kovariaten auf die Klassenzugehörigkeiten lassen sich analog interpretieren.

Abschließend wird der direkte Effekt der Kovariate Geschlecht auf die Indikatorvariable Materialismus berichtet. Für junge Männer wird mit einem Wert von $0,1588$ ein positiver Wert ausgewiesen. Der Effekt ist signifikant. Da die Materialismuskala wie die Postmaterialismuskala von 1 »sehr wichtig« bis 5 »vollkommen unwichtig« geht, bedeutet der positive Effekt, dass junge Männer – ceteris paribus, also bei gleicher Klassenzugehörigkeit – Materialismus ablehnen. Dies ist durch das in der Skala enthaltene Item »eine starke Landesverteidigung« erklärbar, von dem junge Männer im Unterschied zu Mädchen direkt betroffen sind, da sie zum Heer einberufen werden.¹¹ Dies führt zu einer Ablehnung dieses Items und damit auch des Materialismus.

Aus den Parametern lassen sich mit den in Abschnitt 16.2 eingeführten Gleichungen die Auftrittswahrscheinlichkeiten berechnen. Bei kontinuierlichen Indikatoren und Zählvariablen als Indikatoren wird mit Dichtefunktionen gerechnet (siehe Abschnitt 16.2.4). Examplarisch soll die Berechnung für die Klasse C_1 im Materialismus verdeutlicht werden. Einsetzen in die Gleichung 16.1 auf Seite 399

$$f(\mathbf{y}_i | \mathbf{z}_i^{\text{cov}}) = \sum_{x=1}^K p(x | \mathbf{z}_i^{\text{cov}}) \prod_{h=1}^H f(y_{ih} | x, \mathbf{z}_i^{\text{cov}}),$$

ergibt für $f(y_{ih} | x, \mathbf{z}_i^{\text{cov}})$ für die erste latent Klasse (C_1) und für junge Männer

$$f(\text{gMat} | x = 1, \text{Geschlecht} = \text{männlich}) = \varphi((2,1710 + (-0,5761) + 0,1588) / 0,0801),$$

wobei $\varphi(\dots)$ die Dichtefunktion der Standardnormalverteilung ist. Für Mädchen und die erste latente Klasse (C_1) lautet die Funktion:

$$f(\text{gMat} | x = 1, \text{Geschlecht} = \text{weiblich}) = \varphi((2,1710 + (-0,5761) - 0,1588) / 0,0801).$$

¹⁰ Die Ausprägungsbezeichnungen wurden in die Latent-GOLD-Ausgabe nachträglich eingetragen.

¹¹ In Österreich werden alle männlichen Jugendlichen zum Bundesheer einberufen. Es kann ein Ersatzdienst (Zivildienst) geleistet werden.

Tab. 16.12: Ergebnisse für das LC-Modell mit Kovariaten (Parameterschätzung)

	overall	C_1	C_2	C_3	Wald	p-Wert	R^2
<i>Models for Indicators</i>							
gMat		-0,5761	0,1706	0,4055	123,3430	0,0000	0,3949
gPmat		-0,0960	-0,3370	0,4330	26,2061	0,0000	0,3262
<i>Intercepts</i>							
gMat	2,1710				1 107,7827	0,0000	
gPmat	1,6981				683,6607	0,0000	
<i>Error Variances</i>							
gMat		0,0801	0,2569	0,6491			
gPmat		0,0962	0,0567	0,4978			
<i>Model for Clusters</i>							
Intercept	-0,2670	0,2132	0,0538		0,2609	0,8800	
<i>Covariates</i>							
schulM		0,0713	0,0013	-0,0726		0,5555	0,7600
<i>schultyp</i>							
BHS	-0,6172	0,6522	-0,0350		20,7414	0,0004	
AHS	-1,3865	0,7664	0,6201				
BS	2,0037	-1,4186	-0,5850				
<i>Geschlecht</i>							
männlich	0,7075	-0,5153	-0,1922		5,0255	0,0810	
weiblich	-0,7075	0,5153	0,1922				
	gMat		Wald		p-Wert		
<i>Direct Effects</i>							
<i>Geschlecht</i>							
männlich		0,1588	18,0781		0,0000		
weiblich		-0,1588					

Da beim Postmaterialismus außer der latenten Klasse keine weiteren direkten Effekte vorliegen, vereinfachen sich die Gleichungen zu:

$$f(gPmat|x = 1, \text{Geschlecht} = \text{männlich}) = \varphi((1,6981 + (-0,0960))/0,0962)$$

und

$$f(gPmat|x = 1, \text{Geschlecht} = \text{weiblich}) = \varphi((1,6981 + (-0,0960))/0,0962).$$

Die Dichtefunktion ist für Männer und Frauen gleich, da das Modell keinen direkten Effekt vom Geschlecht auf den Postmaterialismus enthält.

Für die Auftrittswahrscheinlichkeiten der latenten Klasse k lauten die Gleichungen.

$$p(x = k | \text{schulm} = l^1, \text{schultyp} = l^2, \text{Geschlecht} = l^3) = \frac{\exp(\gamma_k)}{\sum_{x=1}^K \exp(\gamma_x)},$$

mit

$$\gamma_k = \gamma_{ko} + \gamma_{kl}^1 + \gamma_{kl}^2 + \gamma_{kl}^3.$$

Die Regressionskonstante für die latente Klasse k ist gleich γ_{ko} . Die Effekte der Kovariaten sind γ_{kl}^r . Für numerische Kovariate gilt: $\gamma_{kl}^r = \gamma_k^r \cdot l^r$ mit l^r als Variablenwert. Für die erste Klasse ergeben sich für die Ausprägungskombination »Schulbildung Mutter = 2, Schultyp = BHS, Geschlecht = männlich« folgende Funktionswerte:

$$p(x = 1 | \text{schulm} = 2, \text{schultyp} = \text{BHS}, \text{geschl} = \text{männlich}) = \frac{\exp(\gamma_1)}{\sum_{x=1}^K \exp(\gamma_x)},$$

mit

$$\gamma_1 = -0,2670 + 0,0713 \cdot 2 + (-0,6172) + 0,7075 = -0,0341,$$

$$\gamma_2 = 0,2132 + 0,0013 \cdot 2 + 0,6522 + (-0,5135) = 0,3527 \text{ und}$$

$$\gamma_3 = 0,0538 + (-0,0726) \cdot 2 + (-0,0350) + (-0,1922) = -0,3186.$$

Latent GOLD gibt die Auftrittswahrscheinlichkeiten unter der Überschrift »Profile« aus (siehe Tabelle 16.13 auf Seite 436). Für die kontinuierlichen Indikatoren werden die Klassenmittelwerte berichtet. Cluster 1 lässt sich als Konsensotypus bezeichnen, für Cluster 2 ist eine deutliche Präferenz für den Postmaterialismus erkennbar. Im Cluster 3 ist ebenfalls eine leichte Tendenz für den Postmaterialismus erkennbar, allerdings sind beide Werte nur wichtig bzw. unwichtig. Die Cluster unterscheiden sich von der im vorausgehenden Abschnitt berechneten 3-Klassenlösung geringfügig (siehe Tabelle 16.11 auf Seite 430). Dies ist zum einen darauf zurückzuführen, dass ein zusätzlicher Effekt des Geschlechts auf den Materialismus spezifiziert wurde. Zum anderen wurde in die Schätzung der Auftrittswahrscheinlichkeiten der latenten Klassen Kovariaten einbezogen.

Bei der Ausgabe der »Profile« werden die numerischen Kovariaten zu Kategorien zusammengefasst. Es wird die klassenweise Verteilung auf die zusammengefassten Kategorien berichtet sowie der Klassenmittelwert. Bei den nominalen Variablen wird nur die klassenweise Verteilung auf die Ausprägungen wiedergegeben. Die Werte sind wie folgt zu lesen: Der Mittelwert der ersten latenten Klasse im Materialismus ist gleich 1,6039. Jener des Postmaterialismus unterscheidet sich mit einem Wert von 1,6201 nicht davon. Der Mittelwert in der Bildung der Mutter ist gleich 3,7859. Es liegt also bei einer Skala von 1 bis 8 eine mittlere Bildung vor. Auf die erste Kategorie entfallen 16,67 Prozent, auf die zweite 34,08 Prozent usw. Da es sich um Spaltenanteilsraten handelt ist die Summe gleich 1

bzw. 100 Prozent. Berichtet werden auch die klassenweisen Verteilungen und Mittelwerte für die inaktiven Kovariaten. Für die erste Klasse ergibt sich ein Mittelwert von 2,2385, für die zweite von 2,8680 und für die dritte von 2,5714. In der Tendenz zeigt sich, dass in der Klasse 2 mehr gesellschaftliche Bedrohungen wahrgenommen werden als in den beiden anderen. Signifikanztests werden bei inaktiven Kovariaten nicht durchgeführt. Es lässt sich somit nicht sagen, ob in Cluster 2 signifikant mehr gesellschaftliche Bedrohungen wahrgenommen werden. Sollen Signifikanzprüfungen durchgeführt werden, müssen die Klassenzuordnungswahrscheinlichkeiten abgespeichert werden. Tests können dann mit den entsprechenden statistischen Verfahren durchgeführt werden (siehe Abschnitt 18.5).

Bei der Ausgabe »Profile« werden Spaltenanteils- und Spaltenmittelwerte berechnet. Mitunter sind auch Zeilenanteilswerte von Interesse. In dem Beispiel könnte eventuell bedeutsam sein, wie sich die jungen Männer auf die latenten Klassen verteilen. Zeilanteilswerte werden in Latent GOLD in der Ausgabe »Probmeans« ausgegeben (siehe Tabelle 16.14 auf Seite 437). Die Werte sind wie folgt zu interpretieren: 44 Prozent der jungen Männer gehören dem Cluster 1 an, 35 Prozent dem Cluster 2 und 21 Prozent dem Cluster 3.

Tab. 16.13: Ergebnisse für das LC-Modell mit Kovariaten (Profile)

	C_1	C_2	C_3
Clustergröße	0,4179	0,3965	0,1855
<i>Indicators</i>			
gMat (\bar{x})	1,6039	2,3228	2,5922
gPmat (\bar{x})	1,6021	1,3611	2,1310
<i>Kovariaten</i>			
<i>schulM</i>			
1 – 1	0,1667	0,1232	0,1635
2 – 2	0,3408	0,3065	0,3069
3 – 3	0,3045	0,1584	0,2039
4 – 5	0,1123	0,2053	0,1604
6 – 7	0,0756	0,2066	0,1653
\bar{x}	3,7859	4,4530	4,1903
<i>Schultyp</i>			
BHS	0,1898	0,5524	0,3719
AHS	0,0738	0,3785	0,4542
BS	0,7364	0,0691	0,1739
<i>Geschlecht</i>			
männlich	0,5292	0,4401	0,5499
weiblich	0,4708	0,5599	0,4501
<i>gesellschaftliche Bedrohung</i>			
1	0,2301	0,0457	0,1614
2	0,4066	0,2455	0,2891
3	0,2389	0,5037	0,3653
4	0,1135	0,2050	0,1836
.	0,0108	0,0000	0,0007
\bar{x}	2,2385	2,8680	2,5714

Abkürzung: ».: fehlender Wert

Tab. 16.14: Ergebnisse für LC-Modell mit Kovariaten (Probmeans)

	C_1	C_2	C_3
overall	0,4179	0,3965	0,1855
<i>Indicators</i>			
<i>gMat</i>			
1 – 1,500	0,8122	0,1388	0,0490
1,667 – 1,800	0,7541	0,1750	0,0709
1,833 – 2	0,4250	0,4107	0,1643
2,167 – 2,500	0,1388	0,6085	0,2526
2,667 – 4,833	0,0014	0,6203	0,3783
<i>gPmat</i>			
1 – 1,167	0,2573	0,6438	0,0989
1,200 – 1,333	0,3842	0,5669	0,0489
1,500 – 1,500	0,4186	0,5324	0,0490
1,667 – 1,833	0,6283	0,2371	0,1347
2 – 4,333	0,3433	0,0339	0,6228
<i>Covariates</i>			
<i>schulM</i>			
1 – 1	0,4680	0,3282	0,2038
2 – 2	0,4438	0,3787	0,1774
3 – 3	0,5583	0,2757	0,1660
4 – 5	0,2969	0,5148	0,1883
6 – 7	0,2191	0,5682	0,2127
<i>Schultyp</i>			
1	0,2159	0,5963	0,1878
2	0,1163	0,5659	0,3178
3	0,8376	0,0746	0,0878
<i>Geschlecht</i>			
1	0,4444	0,3506	0,2050
2	0,3917	0,4420	0,1663
<i>gesellschaftliche Bedrohung</i>			
1	0,6667	0,1257	0,2076
2	0,5295	0,3034	0,1671
3	0,2718	0,5437	0,1845
4	0,2914	0,4993	0,2092
.	0,9724	0,0013	0,0263

Abkürzung: ».: fehlender Wert

17 Weiterentwicklungen und Modifikationen

17.1 AutoClass

(gemeinsam mit Arne Bethmann)

Bereits bei Latent GOLD wurde darauf hingewiesen, dass Elemente des Bayes-Schätzverfahrens aufgenommen wurden, um nicht definierte Randlösungen zu vermeiden. In Latent GOLD müssen dazu vom Anwender bzw. der Anwenderin bestimmte Parameter definiert oder die angebotenen Default-Werte verwendet werden (siehe Abschnitt 16.3 auf Seite 411). AutoClass¹ enthält einen vollständig implementierten Bayes-Ansatz. Die erforderlichen Parameter werden automatisch gesucht und die besten Lösungen werden automatisch ausgewählt. Eine Vorab-Definition von Parameterwerten wie in Latent GOLD ist nicht erforderlich.

Die Grundidee der Bayes'schen Statistik besteht darin, auf der Grundlage von Vorwissen und Grundannahmen, die mit M bezeichnet werden sollen, bestimmte *A-priori-Wahrscheinlichkeiten* bzw. *Priori-Wahrscheinlichkeiten* $p(\theta|M)$ für zu schätzende Parameter θ anzugeben. Auf Grundlage dieser Priori-Verteilungen und vorhandener Daten X werden dann Schätzwerte für die Parameter berechnet. Die Schätzwerte sollen so bestimmt werden, dass ihr Auftreten unter Berücksichtigung des Vorwissens und der empirischen Daten maximal ist. Technisch heißt das, dass die *A-posteriori-Wahrscheinlichkeit* $p(\theta|M,X)$ ein Maximum werden soll. Die A-posteriori-Wahrscheinlichkeiten werden über das *Bayes-Theorem* bestimmt mit

$$p(\theta|M,X) = \frac{p(X|\theta,M)p(\theta|M)}{p(X|M)} .$$

¹ AutoClass wurde in den 1990er Jahren von Peter Cheeseman und John Stutz (Cheeseman und Stutz 1996; Hanson, Stutz u. a. 1991) bei der NASA entwickelt. Die Implementierung der von uns verwendeten, in der Programmiersprache C++ geschriebenen Version, realisierten Diane Cook und William Taylor. Mehr Informationen und Download können unter <http://ti.arc.nasa.gov/project/autoclass/> bezogen werden. AutoClass ist frei verfügbar.

Der entscheidende Unterschied zu den bisherigen Methoden besteht in der Spezifikation von Priori-Verteilungen und in der Tatsache, dass in die Parameterschätzung die empirischen Daten und die Priori-Verteilungen einfließen. Als allgemeine Vorteile der Bayes-Modellierung werden genannt (Dunson 2001; Frühwirth-Schnatter 2006):

1. Vorhandenes Vorwissen kann berücksichtigt werden.
2. Komplexe Modelle können geschätzt werden.
3. Die Ergebnisgrößen, wie zum Beispiel Vertrauensintervalle, sind einfacher zu interpretieren. Ein Rückgriff auf »wahre« Werte ist nicht erforderlich.

Bei AutoClass bzw. der automatischen Clusteranalyse mittels Bayes kommt vor allem das zweite Argument zum Tragen. Durch Bayes-Schätzungen können Randlösungen vermieden werden, wie sie bei einer ML-Schätzung auftreten können (siehe Abschnitt 16.3; Frühwirth-Schnatter 2006, S. 174). Die Autorin verweist ferner darauf, dass die Posteriori-Verteilungen vielfach regulärer sind als bei der ML-Methode. Vorteil 1 kommt kaum zum Tragen, da vielfach keine genauen Vorstellungen über die Priori-Verteilungen vorliegen. Daher werden häufig nicht-informative Priori-Verteilungen, zum Beispiel Gleichverteilungen der Parameter in einem bestimmten Intervall, angenommen. Der dritte Vorteil wird häufig noch nicht genutzt. Ausgegeben werden oft nur beste Werte, nicht aber ein Toleranzintervall oder ähnliche Informationen. Als Nachteil von Bayes-Verfahren gilt unter anderem, dass der Rechenaufwand beträchtlich ist. Auch die Abhängigkeit von den Priori-Verteilungen wird genannt (McLachlan und Peel 2000, S. 125–126), allerdings wird dieser Einwand durch die Wahl von nicht-informativen Priori-Verteilungen etwas abgeschwächt. Ein weiteres technisches Problem kann entstehen, wenn sich die Klassenlabels während der Iteration ändern (McLachlan und Peel 2000, S. 129–131).

Wir wollen zunächst die mathematischen Grundlagen von AutoClass vorstellen, bevor wir eine Anwendung mit den Denz-Daten zeigen. Mit simulierten Datensätzen, deren Clusterstruktur bekannt ist, können die Fähigkeiten von Klassifikationsalgorithmen genauer untersucht werden. Bei solchen Tests hat AutoClass gute Ergebnisse gezeigt. Diese werden gemeinsam mit den Ergebnissen anderer in diesem Buch vorgestellter Software in Abschnitt 17.3 auch für AutoClass dokumentiert.

17.1.1 Modell

AutoClass² geht wie die bisher behandelten probabilistischen Verfahren von einem Mischverteilungsmodell aus, das die Modellierung von kontinuierlichen und nominalen Variablen erlaubt. Das Mischverteilungsmodell lautet:

$$p(\mathbf{x}_g|\psi) = \sum_{k=1}^K \pi_k p(\mathbf{x}_g|\theta_k),$$

wobei \mathbf{x}_g den Vektor der Merkmalsausprägungen eines Objekts g , θ_k die Verteilungsparameter der latenten Klasse k und ψ den gesamten Parametervektor des Modells bezeichnen. Für die bedingte Dichtefunktion³ $p(\mathbf{x}_g|\theta_k)$ wird in der Regel lokale Unabhängigkeit angenommen:

$$p(\mathbf{x}_g|\theta_k) = p(x_{g1}|\theta_k) \cdot p(x_{g2}|\theta_k) \cdot \dots$$

Für nominale Variablen wird die Dichte mittels einer Multinomialverteilung modelliert, für kontinuierliche Variablen mittels einer Normalverteilung (siehe auch Gleichung 14.6 auf Seite 357):

$$\begin{aligned} p(x_j = l|\theta_k) &= \pi_{jl|k}, \\ p(x_j|\theta_k) &= \frac{1}{\sigma_{j|k}\sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_{j|k})^2}{2\sigma_{j|k}^2}\right). \end{aligned}$$

Verfügbar ist in AutoClass auch ein Modell, das für die kontinuierlichen Variablen eine Abschwächung der lokalen Unabhängigkeit erlaubt. Auch korrelierte nominale Variablen können modelliert werden. Für kontinuierliche Variablen können in der LISP-, nicht aber in der herunterladbaren C++-Version, auch Poissonverteilungen modelliert werden (Cheeseman und Stutz 1996, S. 66).

Für die Datenmatrix \mathbf{X} ergibt sich die Auftrittswahrscheinlichkeit $p(\mathbf{X}|\psi)$ mit:

$$p(\mathbf{X}|\psi) = \prod_g \sum_k \pi_k p(\mathbf{x}_g|\theta_k).$$

Bei der Bayes-Modellierung wird das Mischverteilungsmodell um Priori-Verteilungen erweitert zu:

$$p(\mathbf{X}|\psi, M_K) = p(\theta_K|M_K) \underbrace{\prod_k p(\theta_k|M_K)}_{\text{Priori-Verteilungen}} \underbrace{\prod_g \sum_k \pi_k p(\mathbf{x}_g|\theta_k, M_K)}_{\text{Likelihood}},$$

² Die nachfolgende Beschreibung folgt Hanson, Stutz u. a. (1991) und Cheeseman und Stutz (1996).

³ Wir verwenden im Folgenden $p(\dots)$ zur Kennzeichnung von Wahrscheinlichkeiten und Dichtefunktionen.

wobei $p(\theta_K|M_K)$ die Priori-Verteilung für die Klassenanteilswerte π_k für das Modell M_K mit K Klassen ist. Die Priori-Verteilungen der Parameter der Klasse k werden mit $p(\theta_k|M_K)$ bezeichnet. Angenommen wird, dass die Priori-Verteilungen der Parameter der latenten Klassen unabhängig sind. Auch innerhalb einer Klasse k wird Unabhängigkeit der Parameter gefordert: $p(\theta_k|M_K) = p(\theta_{k1}|M_K) \cdot p(\theta_{k2}|M_K) \cdot \dots$

Durch die Multiplikation der Likelihood-Funktion mit den Priori-Verteilungen wird ein Ausgleich zwischen einer guten Datenanpassung und der Komplexität des Modells erreicht. Komplexere Modelle mit einer größeren Parameterzahl, wie zum Beispiel Modelle mit einer größeren Klassenzahl oder mit korrelierenden Indikatoren, erzielen automatisch eine bessere Modellanpassung; die Likelihood-Funktion steigt. Umgekehrt besitzen sie mehr Modellparameter und daher mehr Priori-Verteilungen. Dadurch wird mit mehr Wahrscheinlichkeiten, die im Regelfall kleiner 1 sind, multipliziert und der Gesamtausdruck sinkt. Modelle mit mehr Parametern werden daher »bestraft«.

In AutoClass werden folgende Priori-Verteilungen angenommen:

- *Normierte Dirichlet (multiple Betaverteilung):*

$$p(\pi_k|M_K) = C \frac{\Gamma(K+1)}{\Gamma(1+1/K)^K} \prod_{k=1}^K \pi_k^{1/K}$$

für die Klassenanteilswerte π_k mit $C = (6/\pi^2 K^2) K!$ als Skalierungsfaktor. Die Berücksichtigung von $K!$ ist erforderlich, da die Klassenlabels beliebig austauschbar sind. $\Gamma(\dots)$ ist die Gammafunktion.

- *Dirichlet (multiple Betaverteilung):*

$$p(\pi_{jl|k}|M_K) = \frac{\Gamma(L_j + 1)}{[\Gamma(1 + 1/L_j)]^{L_j}} \prod_{l=1}^{L_j} \pi_{jl|k}^{1/L_j}$$

für die bedingten Auftrittswahrscheinlichkeiten $\pi_{jl|k}$ der Ausprägung l der nominalen Variablen j in der Klasse k .

- *Gleichverteilung:*

$$p(\mu_{j|k}|M_K) = \frac{1}{\mu_{j|k\max} - \mu_{j|k\min}}$$

für die Klassenmittelwerte $\mu_{j|k}$ einer Variablen j in der Klasse k im Intervall von $\mu_{j|k\min} = \min(x_j)$ bis $\mu_{j|k\max} = \max(x_j)$.

- *Gleichverteilung:*

$$p(\log \sigma_{j|k}|M_K) = \frac{1}{\log(\sigma_{j|k\max}) - \log(\sigma_{j|k\min})}$$

für den Logarithmus der Klassenstandardabweichung $\sigma_{j|k}$ im Intervall von $\log(\sigma_{j|k\min})$ bis $\log(\sigma_{j|k\max})$ mit $\sigma_{j|k\max}$ als größte Standardabweichung eines Clusters k in j und $\sigma_{j|k\min}$ als kleinste Standardabweichung eines Clusters k in j . $\sigma_{j|k\max}$ wird über die Spannweite der Variablen geschätzt. Für $\sigma_{j|k\min}$ wird programmintern ein unterer Wert definiert, um einen nicht definierten Logarithmus von 0 zu vermeiden.⁴

Alle gewählten Priori-Verteilungen sind flache Verteilungen, das heißt, dass alle möglichen Werte eines Parameters gleich wahrscheinlich sind. Damit soll das Problem vermieden werden, dass falsche Parameter für die Priori-Verteilungen gewählt werden, die zu falschen Schätzungen führen können (siehe 16.3 auf Seite 411). Die Priori-Verteilungen werden daher als nicht informativ bezeichnet.

17.1.2 Schätzverfahren und Bestimmung der Clusterzahl

Die Parameter des Mischverteilungsmodells werden im Bayes-Ansatz so geschätzt, dass die Posteriori-Verteilung für die Parameter

$$p(\psi|X, M_K) = \frac{p(\psi, X|M_K)}{p(X|M_K)} = \frac{p(X|\psi, M_K)p(\psi|M_K)}{p(X|M_K)}$$

ein Maximum wird. Die Schätzer werden als **MAP-Schätzer** (*Maximum-a-posteriori-Schätzer* bzw. *Maximum-posteriori-Schätzer*) bezeichnet.

Eine geschlossene Lösung der Maximierungsaufgabe existiert nicht. Zur Schätzung eingesetzt wird in AutoClass ein modifizierter EM-Algorithmus.⁵ Auf der Basis der vorgegebenen Priori-Verteilungen werden die Priori-Parameter mittels Zufallsprinzip gezogen, die in die weiteren Berechnungen einfließen. Dann wird der E-Schritt durchgeführt mit zufälligen Zuordnungswahrscheinlichkeiten im ersten Durchlauf. Anschließend wird der M-Schritt ausgeführt. Anstelle der ML-Schätzer werden die MAP-Schätzer berechnet. In die Berechnung fließen die Priori-Parameter ein. Die E- und M-Schritte werden so lange durchlaufen, bis Konvergenz vorliegt. Danach werden aus den vorgegebenen Priori-Verteilungen erneut Priori-Parameter gezogen und die Schätzung wird wiederholt. Dazu wird in den Voreinstellungen von 2, 3, 5, 7, 10, 15 und 25 Klassen ausgegangen⁶, in weiteren Schritten werden Zwischenlösungen berechnet. Die Bestimmung der Klassenzahl erfolgt bei AutoClass automatisch. Spezifiziert werden muss ein Abbruchkriterium. Dies kann

⁴ Bei korrelierten Indikatoren wird wie in Latent GOLD eine inverse Wishartverteilung verwendet (Hanson, Stutz u. a. 1991, S. 6).

⁵ Weitere Modellierungen und Schätzverfahren sind beschrieben in McLachlan und Peel (2000, S. 122–125) und in Frühwirth-Schnatter (2006, S. 125–167).

⁶ Die Zahlen können von der Anwenderin auch geändert werden.

die Gesamtzahl der Versuche, die insgesamt durchgerechnet werden, oder die Gesamtrechendauer sein. Um das Potential des Programms auszunutzen, ist es sinnvoll, hier mit einer großen Zahl an Versuchen zu arbeiten (zum Beispiel 50 000 oder 100 000 usw.). Danach werden die Modelle verglichen, um herauszufinden, welches am besten zu den Daten passt. Als Auswahlkriterium für die besten Lösungen verwendet AutoClass die *marginale Likelihood-Funktion*:

$$p(\mathbf{X}|M_K) = \int_{\psi} p(\mathbf{X}|\psi, M_K) p(\psi|M_K) d\psi(\psi).$$

Die marginale Likelihood-Funktion wird mittels einer speziellen von Cheeseman und Stutz (1996) entwickelten Approximation, die auf einer *Laplace-Transformation* basiert, berechnet mit:⁷

$$cs = \log(p(\mathbf{X}'|M_K)) - \log(p(\mathbf{X}'|\psi, M_K)) + \log(p(\mathbf{X}|\psi, M_K)).$$

Die Approximation wird in der Literatur (zum Beispiel Chickering und Heckerman 1997) nach den Autoren auch als *cs-Kriterium* oder *cs-Approximation* (cs für Cheeseman/Stutz) bezeichnet. Mit \mathbf{X}' werden die auf der Basis des EM-Algorithmus erwarteten erschöpfenden Statistiken (gewichtete Anteilswerte und Fallzahlen, gewichtete Klassenmittelwerte und -standardabweichungen, gewichtetes geometrisches Mittel der vorgegebenen Genauigkeit) bei bekannten Klassenzuordnungswahrscheinlichkeiten

$$w_{gk} = \frac{p(\mathbf{x}_g, g \in k | \psi, M_K)}{\sum_{k^*} p(\mathbf{x}_g, g \in k^* | \psi, M_K)}$$

der Objekte g zu den Klassen k bezeichnet. Die Dichtefunktion $p(\mathbf{X}'|\psi, M_K)$ ist die Wahrscheinlichkeit für das Auftreten der erwarteten erschöpfenden Statistiken, wenn die Parameter ψ und die Klassenzuordnungen bekannt sind. Die Dichtefunktion $p(\mathbf{X}'|M_K)$ ist die Wahrscheinlichkeit des Auftretens der erwarteten erschöpfenden Statistiken, wenn nur die Klassenzuordnungen bekannt sind. Über gute Ergebnisse bei der Modellauswahl mittels des cs-Kriteriums berichten Chickering und Heckerman (1997). Auch nach Frühwirth-Schnatter (2006, S. 139–143) ist die marginale Likelihood-Funktion zur Modellauswahl in vielen praktischen Fällen brauchbar. Sie berichtet aber auch über Beispiele, wo dies nicht der Fall ist.

17.1.3 Vergleichsrechnung mit den Denz-Daten

Für das schon in den vorangegangenen Kapiteln verwendete Beispiel der Denz-Daten weist AutoClass durchgehend eine 4-Clusterlösung bei den zehn besten Lösungen aus

⁷ Weitere Schätzmethoden für die marginale Likelihood-Funktion diskutiert ausführlich Frühwirth-Schnatter (2006, S. 139–166).

Tab. 17.1: Ausgabe der Modellprüfgröße (relative marginale Wahrscheinlichkeit) für die zehn besten Modelle bei AutoClass

Marginal-Likelihood	Klassenzahl	Versuch	Duplikat	Verwendung
$\exp(-1674,338)$	4	30 180	9	*saved*
$\exp(-1674,341)$	4	60 249		*saved*
$\exp(-1674,358)$	4	56 557	35	
$\exp(-1674,360)$	4	87 143	24	
$\exp(-1674,377)$	4	61 439	7	
$\exp(-1674,380)$	4	94 593		
$\exp(-1674,380)$	4	24 934	47	
$\exp(-1674,388)$	4	35 983	44	
$\exp(-1674,389)$	4	16 019	27	
$\exp(-1674,397)$	4	92 357		

(siehe Tabelle 17.1). Die Tabelle ist wie folgt zu lesen: Die beste Lösung hat einen Marginal-Likelihood-Wert von $\exp(-1674,338)$. Dieser Wert wird bei Versuch 30 180 gefunden. Die Lösung tritt mehrfach auf, das erste Duplikat wird bei Versuch 9 ermittelt. Für weitere Analysen wird diese Lösung abgespeichert. Bei der zweiten Lösung liegt der Marginal-Likelihood-Wert nur geringfügig niedriger. Auch diese Lösung wird abgespeichert. Die übrigen Lösungen werden nicht abgespeichert, da der Wert für die diesbezügliche Voreinstellung (zwei Lösungen werden abgespeichert) nicht geändert wurde.

Die Mittelwertprofile für die beste Lösung werden in Tabelle 17.2 ausgewiesen:

Klasse 1: Personen in dieser größten Klasse lassen sich als gemäßigte Postmaterialisten bezeichnen, da sie eine etwas stärkere Tendenz zum Postmaterialismus als zum Materialismus aufweisen.

Klasse 2: Sie beinhaltet Personen, die keine deutliche Präferenz zeigen, die sogenannten Nicht-Orientierten.

Klasse 3: Hier zeigt sich eine starke Zustimmung zum Postmaterialismus, während Materialismus verhältnismäßig stark abgelehnt wird.

Tab. 17.2: Klassenprofile der 4-Clusterlösung (Klassengrößen und Mittelwerte der Indikatorvariablen)

	Klasse 1	Klasse 2	Klasse 3	Klasse 4	gesamt
Größe (n)	166	28	23	4	221
Anteile	0,751	0,127	0,104	0,018	1,0
$g\text{Mat}(\bar{x})$	1,8369	2,2310	3,1217	4,5000	2,0688
$g\text{Pmat}(\bar{x})$	1,4952	2,5119	1,1826	1,9167	1,5991

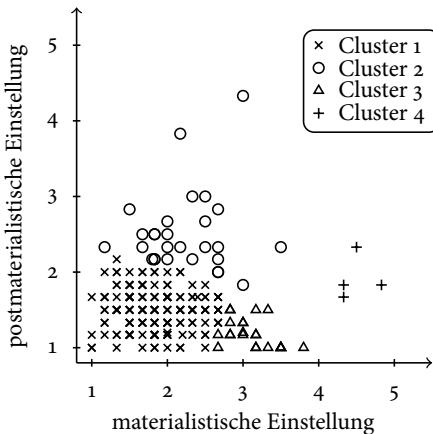


Abb. 17.1: Die Lösung von AutoClass für die Denz-Daten als Scatterplot

Klasse 4: Diese Klasse deutet in den Präferenzen auf ein anti-materialistisches Profil hin, ist mit vier Fällen aber eher als Restkategorie zu interpretieren.

Diese Ergebnisse zeigen starke Ähnlichkeiten mit den bisher durchgeführten Analysen. Das erklärt sich daraus, dass die Modellierung in AutoClass einer latenten Profilanalyse entspricht, wenn – wie in diesem Beispiel – ausschließlich kontinuierliche Merkmale verwendet werden. Wie sich die im zweidimensionalen Raum befindlichen Fälle gruppieren, zeigt Abbildung 17.1.

17.2 TwoStep-Cluster

Der Algorithmus TwoStep-Cluster wurde von Chiu, Fang u. a. (2001) entwickelt und ist in IBM-SPSS verfügbar (spss Inc. 2000). Er ist für große Datensätze geeignet, da er eine Art »Vorclusterung« durchführt und daher sehr schnell ist. Weiterhin verarbeitet die Prozedur gleichzeitig metrische (»continuous«) und nominale (»categorical«) Variablen. Die Clusterzahl wird auf Grundlage des BIC oder AIC automatisch bestimmt.

17.2.1 Allgemeiner Ansatz

Der Algorithmus besteht aus zwei Schritten:

Schritt 1: Vorgruppierung der Fälle

Schritt 2: Clustern der Gruppen

Vorgruppierung der Fälle: In diesem ersten Schritt werden alle Datensätze der Reihe nach vorgruppiert. Diese Gruppen (*pre-cluster*), die im Schritt 2 anhand ihrer Eigenschaften den endgültigen Clustern zugeordnet werden, können als dichte Bereiche im analysierten Eigenschaftsraum verstanden werden. Man kann sich diese Vorgehensweise graphisch als einen Baum vorstellen, wobei sich die Gesamtzahl der Fälle vom Ursprung (Wurzel) her immer weiter in Äste verzweigt und schließlich als Gruppen in Blättern mündet (Brosius 2008, S. 747). Die Anzahl dieser Gruppen wird mit der Höchstzahl an Verzweigungen pro Blattknoten MXBRANCH (voreingestellter Wert: 8) und der maximalen Baumtiefe MXLEVEL (3) gesteuert, die angezeigte maximale Anzahl⁸ ($\sum_{i=1}^{\text{MXLEVEL}} \text{MXBRANCH}^i$) – 1 liegt für die Voreinstellung bei 585.

Je höher der Wert für den Schwellenwert für anfängliche Distanzänderung (Voreinstellung: 0), desto geringer wird die Anzahl der Gruppen. Weil das Ergebnis von der Reihenfolge der Fälle abhängen kann – die Fälle werden ja nacheinander den Gruppen zugeordnet –, empfiehlt SPSS Inc. (2000, S. 4) eine zufällige Sortierung zu verwenden.⁹

Clustern der Gruppen: Die Gruppen aus Schritt 1 werden mit Hilfe eines modellbasierten hierarchischen Verfahrens auf die endgültigen Cluster (siehe Abschnitt 1.3) verschmolzen. Im Modell wird angenommen, dass im Cluster k die metrischen Variablen x_j ($j = 1, \dots, p$) unabhängig normalverteilt sind (Mittelwert μ_{kj} , Standardabweichung σ_{kj}) und die nominalen Variablen a_j ($j = 1, \dots, q$) multinomialverteilt sind mit der Wahrscheinlichkeit π_{kjl} für Kategorie l ($l = 1, \dots, m_l$) der Variable a_j . Die Log-Likelihood-Distanz (die ebenfalls verfügbare euklidische Distanz kann nur metrische Variablen verarbeiten) zwischen zwei Clustern K und K^* ist definiert als

$$d(K, K^*) = \xi_K + \xi_{K^*} - \xi_{\{K, K^*\}},$$

wobei

$$\xi_{\bullet} = \underbrace{-n_{\bullet} \sum_{j=1}^p \frac{1}{2} \log(\sigma_{\bullet j}^2 + \sigma_j^2)}_{\text{Beitrag der metrischen Variablen}} + \underbrace{n_{\bullet} \sum_{j=1}^q \sum_{l=1}^{m_j} \pi_{\bullet jl} \log \pi_{\bullet jl}}_{\text{Beitrag der nominalen Variablen}}.$$

Der Term ξ wird für die Cluster K , K^* und das fusionierte Cluster $\{K, K^*\}$ berechnet. Anstelle des Punktes kann in der Formel K , K^* und $\{K, K^*\}$ eingesetzt werden. Dabei kann ξ als Streuungsmaß innerhalb eines Clusters interpretiert werden. Wie bei den hierarchisch-agglomerativen Methoden werden nun nacheinander diejenigen Cluster

⁸ Spätestens mit Version 17 wurde die Berechnung der maximalen Anzahl von $\text{MXLEVEL}^{\text{MXBRANCH}} = 8^3 = 512$ (für die voreingestellten Werte) auf die im Text angegebene Formel umgestellt.

⁹ Brosius (2008, S. 747–751) gibt weitere Informationen zu diesem Schritt, der auf einem BIRCH-Verfahren beruht (Zhang, Ramakrishnan u. a. 1996, 1997).

Tab. 17.3: Modellprüfgrößen für die TwoStep-Analyse mit den Wertedaten von Denz

Anzahl der Cluster	Bayes-Kriterium nach Schwarz (BIC)	BIC-Änderung ^a	Verhältnis der BIC-Änderungen ^b	Verhältnis der Distanzmaße ^c	Strafterm ^d b_K
1	326,963	—	—	—	21,59
2	258,555	-68,407	1,000	1,544	43,19
3	221,870	-36,685	0,536	2,053	64,78
4	215,070	-6,800	0,099	1,432	86,37
5	216,840	1,770	-0,026	1,342	107,96
6	223,666	6,826	-0,100	1,035	129,56
7	230,997	7,332	-0,107	1,155	151,15
8	240,240	9,243	-0,135	1,346	172,74
9	252,659	12,419	-0,182	1,632	194,33
10	268,630	15,971	-0,233	1,118	215,93

- a) Die Änderungen wurden von der vorherigen Anzahl an Clustern in der Tabelle übernommen.
- b) Die Änderungsquote ist das Verhältnis der Änderung der BIC-Werte zweier aufeinander folgender Lösungen zur Änderung des BIC-Wertes zwischen 1- und 2-Klassenlösung. Grau hinterlegt ist der Wert, bei dem erstmals gilt: $BIC_K - BIC_{K+1} < 0,04 \cdot (BIC_1 - BIC_2)$.
- c) Die Quoten für die Distanzmaße beruhen auf der aktuellen Anzahl der Cluster im Vergleich zur vorherigen Anzahl der Cluster. Grau hinterlegt ist das höchste Verhältnis.
- d) Der Strafterm ist nicht Bestandteil der Ausgabe, die Berechnung erfolgt nach Formel 17.1 mit $p = 2$, $q = 0$ und $n = 221$.

mit der jeweils kleinsten Distanz $d(K, K^*)$ verschmolzen bis nur noch ein Cluster übrig ist.

Anzahl der Cluster: Die Bestimmung der Clusterzahl erfolgt auf Grundlage des bayesianischen Informationskriteriums BIC:¹⁰

$$BIC_K = -2 \sum_{v=1}^K \xi_v + \underbrace{K(2p + \sum_{v=1}^q (m_l - 1))}_{\substack{\text{Anzahl zu schätzender Parameter}} \cdot \log n}_{\substack{\text{Strafterm für Modellkomplexität } b_K}} . \quad (17.1)$$

Falls $BIC_1 < BIC_2$ bricht der Algorithmus ab und die 1-Clusterlösung wird ausgegeben. Ansonsten wird die maximale Clusterzahl K_{max} dort angenommen, wo erstmals gilt $BIC_K - BIC_{K+1} < 0,04 \cdot (BIC_1 - BIC_2)$. Damit wird die Stelle gesucht, wo der Abfall des BIC im Vergleich zum Abfall zwischen der 1- und 2-Clusterlösung sehr klein wird bzw. der BIC wieder ansteigt, was an negativen BIC-Änderungen in der von IBM-SPSS ausgegebenen Tabelle 17.3 zu erkennen ist.

¹⁰ Möglich ist hier auch die Verwendung des AIC.

In der Tabelle findet man K_{\max} dort, wo in der Spalte »Verhältnis der BIC-Änderungen« der Wert 0,04 zum ersten Mal unterschritten wird. Die darüber stehende Clusterlösung ist dann die Lösung mit der maximalen Clusterzahl. Eine negative BIC-Änderung – und damit ein Wert kleiner 0,04 – tritt mit -0,026 bei der 5-Clusterlösung auf, daher ist in dem Beispiel die maximale Clusterzahl gleich 4.

Ausgehend von dieser Obergrenze wird mit Hilfe des Verhältnisses der Distanzmaße $R_2(K)$ die endgültige Anzahl der Cluster bestimmt:

$$R_2(K) = \frac{d_{\min}(C_K)}{d_{\min}(C_{K+1})},$$

wobei C_K das Clustermodell mit K Clustern und $d_{\min}(C)$ das Minimum der Distanz zwischen den Clustern eines Modells C sind. Für alle Clustermodelle mit höchstens K_{\max} Clustern wird nun $R_2(K)$ berechnet. Als Lösung in Betracht kommen nun das höchste (2,053 für die 3-Clusterlösung) oder das zweithöchste (1,544 bei der 2-Clusterlösung) Verhältnis. Diese beiden Größen werden erneut miteinander in Beziehung gesetzt. Es wird also das Verhältnis von 2,053 : 1,544 berechnet – es beträgt 1,330. Ist das Verhältnis der zwei höchsten Distanzverhältnisse größer als ein Schwellenwert¹¹, der mit 1,15 festgelegt ist, wird die Lösung mit dem höchsten Distanzverhältnis als die beste ausgewählt. In dem Beispiel ist diese Bedingung erfüllt, daher wird die 3-Clusterlösung als die beste ausgewählt. Andernfalls wäre die Clusterlösung mit der größten Clusterzahl ausgewählt worden.

Während in der Ausgabe von IBM-SPSS die Tabelle »Automatische Clusterbildung«¹², die hier in Tabelle 17.3 dokumentiert ist, die Spalten »BIC-Änderung« und »Verhältnis der BIC-Änderungen« leicht nachvollzogen werden können,¹³ trifft dies nicht ohne weiteres auf das »Verhältnis der Distanzmaße« $R_2(K)$ zu. Ausgehend vom BIC und dem Strafterm b_K , der zusätzlich zur Ausgabe von IBM-SPSS in Tabelle 17.3 wiedergegeben ist, kann $R_2(K)$ wie folgt berechnet werden:

$$R_2(K) = \frac{\text{BIC}_{K-1} - \text{BIC}_K + b_K - b_{K-1}}{\text{BIC}_K - \text{BIC}_{K+1} + b_{K+1} - b_K}.$$

¹¹ Wie die Grenze 0,04 bei Bestimmung von K_{\max} wurde auch diese Grenze mit Hilfe von Simulationsrechnungen ermittelt.

¹² In Version 18 wird bei der Anzeige der Tabelle die Warnung ausgegeben, dass der entsprechende Unterbefehl künftig wegfallen und die Tabelle nicht mehr zur Verfügung stehen wird.

¹³ Die Notation der Dokumentation (SPSS Inc. 2008, S. 817) weicht insofern vom ausgegebenen Ergebnis ab, dass die Inhalte der Ausgabe in den zwei Spalten rechts vom BIC-Wert um eine Zeile nach oben verschoben werden müssten, wenn man der Notation folgt.

Tab. 17.4: Klassenprofile der 3-Clusterlösung (Klassengrößen und Mittelwerte der Indikatorvariablen)

	Klasse 1	Klasse 2	Klasse 3	gesamt
Größe (n)	59	37	125	221
Anteile	0,267	0,167	0,566	1,0
gMat (\bar{x})	2,7610	2,3459	1,6600	2,0688
gPmat (\bar{x})	1,3480	2,3919	1,4829	1,5991

Zuordnung der Fälle zu Clustern: Jedes Objekt wird auf Grundlage des verwendeten Distanzmaßes deterministisch zu einem Cluster zugeordnet. Dies führt bei überlappenden Clustern zu verzerrten Schätzungen (Abschnitt 12.1 und Bacher 2000).

Behandlung von Ausreißern: Das Verfahren kann auf Grundlage eines vom Anwender anzugebenden Rausch-Anteils, zum Beispiel 5 Prozent, alle Gruppen aus Schritt 1, deren Größe unterhalb dieses Anteils liegt, als Ausreißer behandeln und im zweiten Schritt ignorieren.

Ausgabe: Die Ausgabe erfolgt seit Version 18 mit einer sogenannten Modellanzeige, die mit Doppelklick auf die Modellübersicht im Ausgabefenster erreicht werden kann. Damit ist eine sehr komfortable Exploration der erzeugten Clusterlösung möglich. Weiterhin kann auch die Größe des Einflusses von Variablen auf das Modell statistisch mit der Prozedur AIM getestet werden (SPSS Inc. 2008, S. 9–10; Schendera 2010, S. 105–115).

17.2.2 Ergebnis für die metrischen Denz-Daten und für gemischte Skalenniveaus

Der Algorithmus ermittelt für die Denz-Daten eine 3-Clusterlösung (siehe Tabelle 17.4). Identifiziert werden die gemäßigten Postmaterialisten als größte Klasse 3 mit relativ niedrigen Werten auf beiden Skalen. Als zweitgrößtes Cluster werden in Klasse 1 die Postmaterialisten zusammengefasst. Das kleinste Cluster ist Klasse 2 – die Gruppe der Nicht-Orientierten.

Wie bereits angedeutet, kann IBM-SPSS TwoStep-Cluster grundsätzlich auch Datensätze mit *Variablen gemischter Skalenniveaus* klassifizieren. Bei ordinalen Variablen muss die Entscheidung getroffen werden, ob sie nominal oder metrisch in die Analyse eingehen sollen.

Eine von Bethmann und Wenzig (2010) vorgenommene Reanalyse von Beispielen aus einschlägigen SPSS-Lehrbüchern und der Originaldokumentation kommt jedoch zum Ergebnis, dass die gefundenen Cluster Auffälligkeiten in der Verteilung der nominalen

Variablen haben. Es findet regelmäßig eine Entmischung der Daten bezüglich der (Ausprägungskombinationen der) nominalen Variablen statt und die Information der metrischen Variablen spielt bei der Clusterlösung eine stark untergeordnete Rolle. Bacher, Wenzig u. a. (2004, S. 7) stellen nach einer Analyse des Distanzmaßes fest, dass ein Unterschied in einer nominalen Variable gleichbedeutend mit einem Unterschied von 2,45 Einheiten bei einer standardisierten metrischen Variable ist. Es müssen hiermit in einer quantitativen Variablen sehr große Unterschiede vorliegen, damit sie gleichbedeutend sind mit unterschiedlichen nominalen Ausprägungen. Ihre Simulationsexperimente legen auch die Empfehlung nahe, den Algorithmus nur unter Vorbehalt für gemischtskalierte Variablen zu benutzen.

17.3 Vergleich ausgewählter Software (*gemeinsam mit Arne Bethmann*)

In Experimenten mit simulierten Datensätzen kann die Effektivität von verschiedenen Algorithmen getestet werden, weil die »richtigen« Ergebnisse bereits bekannt sind. Hierfür wurden von Bacher, Wenzig u. a. (2004) Datensätze von unterschiedlich komplexen »Gesellschaften« vorgestellt (vgl. Tabelle 17.5 auf Seite 453). Diese sollen zur Prüfung ausgewählter Clusterverfahren verwendet werden.

17.3.1 Simulationsmodelle

Im einfachsten Modell 0 enthält der Datensatz keine Clusterstruktur. Die drei Zufallsvariablen (normalverteilt mit Standardabweichung $\sigma = 1$), die als Einkommen, Bildung und Beruf interpretiert werden können, haben für alle Fälle den gleichen Mittelwert ($\mu = 0$). In den Modellen 2a und 2b sind jeweils zwei Klassen simuliert, in denen die Mittelwerte der Variablen entweder hoch (Oberschicht) oder gering (Unterschicht) sind. Da bei Modell 2b die Mittelwerte weiter auseinanderliegen ($\pm 1,5$), unterscheiden sich die Fälle stärker und damit sind die Gruppen einfacher zu identifizieren als in Modell 2a ($\mu = \pm 0,75$) mit zwei überlappenden Clustern.

In Modell 3 kommt eine Mittelschicht hinzu, die durch jeweils mittlere Durchschnittswerte in den Schichtungsvariablen charakterisiert ist. Im komplexesten Fall (Modell 5) wird eine Gesellschaft mit einer Unter-, Mittel- und Oberschicht sowie mit Selbständigen und Intellektuellen simuliert. Die Variablen sind in den ersten drei genannten Schichten wie in Modell 3 konsistent und nehmen im Durchschnitt niedrige, mittlere bzw. hohe

Werte an. Für die Gruppe der Selbständigen werden hohe Einkommenswerte und mittlere Berufs- und Bildungswerte simuliert, für die Gruppe der Intellektuellen umgekehrt mittlere Einkommenswerte und hohe Bildungs- und Berufswerte. Damit ist das Modell 5 mit fünf Clustern, die zum Teil inkonsistent sind, sicher die größte Herausforderung für die Algorithmen.

Für jedes Modell liegen jeweils Datensätze in sechs Stichprobengrößen (500, 1 000, 2 000, 5 000, 10 000, 20 000) vor. Für Clusteranalysen mit gemischtskalierten Variablen wurden zwei Variablen (diejenigen, die als Bildung und Beruf interpretiert werden können) in 0, 1 und 2 trichotomisiert (an den Grenzen -1 und $+1$ der von $-\infty$ bis $+\infty$ gehenden Variablenwerte der Standardnormalverteilung) und als ordinalskaliert behandelt. Wenn Algorithmen ordinale Merkmale nicht als solche angemessen modellieren können, muss entschieden werden, ob diese Variablen als kontinuierlich oder nominal in die Analysen eingehen sollen. Dabei wurde von folgender Regel ausgegangen: Die ordinalen Variablen werden als nominale Variablen behandelt, wenn das Programm Modelle für nominalskalierte Variablen enthält, sonst als quantitative Variablen.

Um den Algorithmen mehr Information zur Verfügung zu stellen, wurden Datensätze mit drei und sechs Variablen erzeugt. Letztere lassen sich zum Beispiel als Schichtzugehörigkeit von Kindern interpretieren, die aus den Informationen der beiden Eltern ermittelt wird. Jeder Datensatz wurde außerdem fünfmal simuliert, um zufällige Verzerungen zu minimieren. Somit werden insgesamt 300 Experimente durchgeführt, um die Leistungsfähigkeit der Algorithmen zu untersuchen.¹⁴

17.3.2 Ergebnisse

In Tabelle 17.6 sind exemplarisch die detaillierten Ergebnisse unserer Berechnungen mit AutoClass für alle Modelle dargestellt.¹⁵ Sie sind nach Stichprobengrößen (Spalte N) und Variablenanzahl (3 bzw. 6) differenziert. Die grau hinterlegten Werte geben den Anteil der korrekt klassifizierten Objekte an, wenn die korrekte Anzahl von Clustern gefunden wird. Ist das nicht der Fall, gibt der Wert in Klammern die von AutoClass als optimal ausgegebene Zahl von Clustern an.

¹⁴ Diese Experimente haben Bacher, Wenzig u. a. (2004, S. 12) mit dem TwoStep-Clusteralgorithmus von SPSS (heute IBM-SPSS) durchgeführt, dabei wurde von einer etwas anderen Datenkonstellation ausgegangen: eine bzw. zwei Variablen wurden als quantitativ betrachtet, eine bzw. zwei als ordinal und eine bzw. zwei als nominal.

¹⁵ Für die hier vorgestellten Ergebnisse haben wir die ordinalen Merkmale als nominalskaliert in die Berechnungen aufgenommen, da in AutoClass keine Möglichkeit besteht, die Ordinalität der Merkmale zu berücksichtigen.

Tab. 17.5: Übersicht über die Datensatzstruktur der Simualtionsexperimente mit den Mittelwerten der simulierten Variablen (normalverteilt, Standardabweichung $\sigma = 1$). Für Experimente mit gemischtskalinierten Variablen wurden zwei Variablen (Bildung, Beruf) bei ± 1 ordinalisiert.

Modell	Cluster	Anteile in %	$\mu_{\text{Einkommen}}$	μ_{Bildung}	μ_{Beruf}	Interpretation
Mo	1	100	0	0	0	keine Clusterstruktur
M2a	1	50	-0,75	-0,75	-0,75	Unterschicht
	2	50	0,75	0,75	0,75	Oberschicht
M2b	1	50	-1,5	-1,5	-1,5	Unterschicht
	2	50	1,5	1,5	1,5	Oberschicht
M3	1	25	-1,5	-1,5	-1,5	Unterschicht
	2	50	0	0	0	Mittelschicht
	3	25	1,5	1,5	1,5	Oberschicht
M5	1	15	-1,5	-1,5	-1,5	Unterschicht
	2	50	0	0	0	Mittelschicht
	3	15	1,5	1,5	1,5	Oberschicht
	4	10	1,5	0	0	Selbständige
	5	10	0	1,5	1,5	Intellektuelle

Tab. 17.6: Übersicht über die Simulationsergebnisse für AutoClass

Modell	N	3 Variablen		6 Variablen	
Modell o	20 000	100,100,100,100,100		100,100,100,100,100	
	5000	100,100,100,100,100		100,100,100,100,100	
	500	100,100,100,100,100		100,100,100,100,100	
Modell 2a	20 000	88,88,88,88,89		95,95,95,95,95	
	5000	88,88,88,88,88		95,95,95,95,96	
	500	89,87,88,88,87		96,94,94,94,94	
Modell 2b	20 000	99,99,99,99,99		100,100,100,100,100	
	5000	99,99,99,99,99		100,100,100,100,100	
	500	99,99,99,100,100		100,100,100,100,100	
Modell 3	20 000	83,83,83,83,84		94,93,93,93,93	
	5000	84,83,84,83,83		93,93,93,93,93	
	500	82,83,79,82,80		94,93,93,90,92	
Modell 5	20 000	(3),(4),(3),(3),(3)		82,82,82,82,83	
	10 000	(3),(4),(3),(3),(3)		82,82,82,82,83	
	5000	(3),(3),(3),(3),(3)		82,83,82,82,83	
	2000	(3),(3),(3),(3),(3)		82,(4),82,(4),80	
	1000	(3),(3),(3),(3),(3)		(3),(3),(3),(3),(4)	
	500	(2),(3),(3),(4),(2)		(3),(3),(3),(3),(4)	

grau hinterlegt: Prozentanteile der korrekt klassifizierten Fälle
in Klammern: Anzahl der ermittelten Cluster (bei nicht korrekt ermittelter Clusteranzahl)

Dabei zeigt sich, dass bei den 2- und 3-Cluster-Modellen immer die korrekte Anzahl von Clustern identifiziert wird. In den meisten Experimenten werden dabei weit über 80 Prozent der Fälle richtig zugeordnet. Daneben fällt auf, dass die zunehmende Komplexität der Clusterstruktur – wie zu vermuten war – zu schlechteren Klassifikationsergebnissen führt. In Modell 2b mit zwei sauber getrennten Clustern werden fast alle Objekte korrekt zugeordnet. Im Fall der zwei überlappenden Cluster (Modell 2a) nimmt die Güte der Klassifikation dagegen bereits ab. Noch ein wenig schlechter werden die Ergebnisse, wenn das Modell aus drei überlappenden Clustern besteht (Modell 3). Ein weiterer wichtiger Einflussfaktor für die Güte der Zuordnung ist die Menge an Informationen, die in einem Datensatz zur Verfügung steht. Besonders für das Modell mit fünf Clustern zeigt sich, dass der Klassifikationserfolg stark von der Anzahl der verwendeten Variablen abhängig ist. Während es AutoClass bei den Datensätzen mit nur drei Variablen nie gelingt, auch nur die korrekte Anzahl von Clustern zu bestimmen, sehen die Ergebnisse bei sechs Variablen zumindest ein wenig besser aus. Die Güte der Zuordnungen ist auch in diesen Experimenten sehr hoch, sobald die richtige Anzahl der Cluster identifiziert wird. Die Informationsmenge spiegelt sich auch in der Anzahl der zu klassifizierenden Objekte wieder. Über alle Experimente hinweg zeigt sich, dass die Ergebnisse mit steigender Fallzahl in der Tendenz besser werden. AutoClass ist außerdem in der Lage, eine fehlende Clusterstruktur (Modell 0) korrekt zu erkennen. In keinem der Experimente hat das Programm fälschlicherweise eine Clusterstruktur angenommen, unabhängig von der Stichprobengröße und von der Anzahl verwendeter Variablen.

Bei den weniger komplexen Modellen 0 bis 3 für drei Variablen zeigt sich, dass viele der eingesetzten Programme recht gut damit umgehen können (Bacher, Wenzig u. a. 2004; Bethmann und Wenzig 2008). Daher wird nachfolgend das 5 Clustermodell ausführlicher behandelt, welches allen Programmen Probleme bereitet. Lediglich Latent GOLD kann hier teilweise bereits für drei Variablen korrekte Ergebnisse liefern (siehe Tabelle 17.7). Stellt man den Algorithmen sechs Variablen zur Verfügung, verbessern sich die Ergebnisse deutlich. Vollständig korrekte Lösungen (siehe Tabelle 17.7) liefert aber auch dann keines der Programme: Am besten schneiden Latent GOLD und AutoClass ab. Latent GOLD als beste Software benötigt mindestens 2 000 Fälle, um die 5-Clusterstruktur in allen fünf Versuchen zu entdecken. Auch bei 1 000 Fällen wird die Klassenzahl dreimal korrekt aufgefunden. AutoClass benötigt mindestens 5 000. Die anderen Softwarepakete liefern in keinem Fall vollständig korrekte Ergebnisse.

Die mangelhaften Ergebnisse von MCLUST sind vermutlich darauf zurückzuführen, dass für die ordinalskalierten Variablen eine Normalverteilung angenommen werden musste, da keine anderen Verteilungsannahmen zur Verfügung stehen. Dies schränkt die Verwendbarkeit von MCLUST für gemischtskalierte Variablen ein. Für metrische Merkmale scheint MCLUST allerdings sehr gute Ergebnisse zu liefern (Fraley und Raftery 2006). Das

Tab. 17.7: Vergleich der Ergebnisse verschiedener Algorithmen für das Modell 5

Algorithmus	N	3 Variablen	6 Variablen
ALMO 12	20 000	4,4,4,4,4	6,6,6,6,6
	10 000	4,4,4,4,4	5,5,5,4,4
	5 000	3,4,4,3,3	4,4,5,5,4
	2 000	3,3,3,3,3	4,4,4,4,4
	1 000	2,2,2,3,2	3,3,4,3,4
	500	2,2,2,2,2	3,3,3,3,3
AutoClass ^d	20 000	3,4,3,3,3	5,5,5,5,5
	10 000	3,4,3,3,3	5,5,5,5,5
	5 000	3,3,3,3,3	5,5,5,5,5
	2 000	3,3,3,3,3	5,4,5,4,5
	1 000	3,3,3,3,3	3,3,3,3,4
	500	2,3,3,4,2	3,3,3,3,4
Latent GOLD 4.5 ^b	20 000	5,5,5,5,5 ^a	5,5,5,5,5
	10 000	5,6,5,3,3 ^a	5,5,5,5,5
	5 000	3,4,4,3,3	5,5,5,5,5
	2 000	3,3,3,3,3	5,5,5,5,5
	1 000	3,3,3,3,3	4,5,5,4,5
	500	2,3,3,3,2	3,3,3,3,4
MCLUST ^c	20 000	12,12,4,12,12	5,6,6,7,7
	10 000	12,12,12,12,12	5,6,6,6,6
	5 000	12,12,12,12,12	3,6,6,6,4
	2 000	12,12,4,12,12	6,5,5,5,6
	1 000	11,4,4,12,3	9,5,5,4,5
	500	12,12,12,3,12	3,5,3,3,8
SPSS TwoStep ^d	20 000	7,7,7,7,7	6,5,6,4,4
	10 000	7,7,7,7,7	6,5,6,3,3
	5 000	7,7,7,7,7	6,3,3,4,2
	2 000	7,7,7,7,7	6,7,2,6,6
	1 000	7,7,7,7,7	3,3,7,4,2
	500	7,7,7,7,7	6,2,3,6,3
WEKA ^d	20 000	3,3,3,3,3	8,8,6,10,9
	10 000	3,3,3,3,3	8,8,6,10,9
	5 000	3,3,3,3,3	6,6,10,7,9
	2 000	3,4,4,3,3	8,5,8,7,5
	1 000	3,3,4,3,4	4,3,4,4,5
	500	2,3,3,3,3	3,3,3,3,4

a) teilweise nicht konvergiert

b) kleinster BIC für Lösungen mit ein bis sieben Clustern

c) ordinal als quantitativ

d) ordinal als nominal

grau unterlegt: richtige Anzahl der Cluster ermittelt

schlechte Abschneiden von ALMO ist dadurch bedingt, dass ordinale Variablen analysiert werden. Bei gemischten Variablen (quantitativ, ordinal und nominal) werden bessere Ergebnisse erzielt (Bacher, Wenzig u. a. 2004).

Die Anwendung von Mischverteilungsmodellen stellt einen gangbaren Weg zur automatischen Bestimmung von Clusteranzahlen dar. Die tatsächliche Güte der Ergebnisse ist aber – wie sich in den Simulationen zeigt – von verschiedenen Faktoren abhängig. Zum einen spielt die Menge der Information, die für die Modellierung der Clusterstruktur zur Verfügung steht, eine große Rolle: Sowohl mit steigender Anzahl der zu klassifizierenden Objekte als auch mit der Anzahl der verwendeten Variablen werden die Ergebnisse besser. Zum anderen ist die konkrete Implementierung der Schätzung in der Software ausschlaggebend. Die Softwarepakete unterscheiden sich dabei unter anderem in den schätzbaren Modellparametern (Verteilungsannahmen etc.), in der Implementierung des Schätzverfahrens (zum Beispiel EM-Algorithmus mit Maximum-Likelihood- oder Maximum-a-posteriori-Schätzung) und in den Gütekriterien für den Vergleich der geschätzten Modelle (häufig BIC oder AIC, bei AutoClass aber zum Beispiel die Marginal-Likelihood-Funktion). Nochmals sei aber betont, dass formal in der Regel mehrere Clusterlösungen geeignet sind und eine eindeutige (automatische) Bestimmung der Clusterzahl in dieser Phase der Analyse häufig nicht möglich ist.

Teil IV

Spezielle Anwendungsfragen

18 Häufig gestellte Anwendungsfragen

18.1 Welches Verfahren?

Welches Verfahren eingesetzt werden soll, hängt von der untersuchten Fragestellung, den vorhandenen Daten und der verfügbaren Software ab. Nachfolgend sollen folgende Fragestellungen unterschieden werden:

1. Bildung abgeleiteter Variablen,
2. räumliche Darstellung von Variablen oder Objekten,
3. hierarchische Ähnlichkeitsdarstellung von Variablen oder Objekten,
4. Klassifikation von Variablen,
5. Klassifikation von Objekten,
6. Klassifikation von Verläufen (siehe Kapitel 19).

18.1.1 Bildung abgeleiteter Variablen

Wenn möglich empfehlen wir die Analyse mit *abgeleiteten Variablen* (Faktor- oder Gesamtpunktwerte). Dadurch werden Messfehler und somit der Einfluss irrelevanter Variablen reduziert, was insbesondere für die deterministische Clusteranalyse von Bedeutung ist. Ferner wird Übersichtlichkeit und damit leichtere Interpretierbarkeit ermöglicht. Als Standardverfahren zur Bildung von abgeleiteten Variablen empfehlen wir die Faktorenanalyse (Hauptkomponenten- oder Hauptachsenanalyse mit anschließender Rotation). Die Faktorenanalyse setzt quantitative oder – der Forschungspraxis folgend – ordinalskalierte Variablen voraus. Auch bei dichotomen Variablen kann sie eingesetzt werden, wenn die Items nicht zu unterschiedlich schwierig sind (siehe Abschnitt 5.3). Die Forderung ist entsprechend den Simulationsergebnissen von Abschnitt 5.3.1 erfüllt, wenn die Anteilswerte der Ja-Antworten zwischen 25 und 75 Prozent variieren. Die Faktorenanalyse ist nicht geeignet für:

- nominale Items mit mehr als zwei Ausprägungen,
- dichotome Items mit sehr unterschiedlichen Schwierigkeitsgraden (hierarchische Items),
- Präferenzdaten mit fester Auswahl,
- mehrdimensionale Items.

Für diese Datenkonstellationen empfehlen wir folgende Verfahren:

- *nominalskalierte Variablen mit mehr als zwei Ausprägungen*: hier empfehlen wir den Einsatz der nominalen Faktorenanalyse nach McDonald oder der multiplen Korrespondenzanalyse (siehe Abschnitt 5.4).
- *Präferenzdaten mit fester Auswahl*: ein- oder mehrdimensionales Unfolding (Sixtl 1982, S. 414–431). ALMO bietet hier eine probabilistische Verallgemeinerung für den eindimensionalen Fall an. Bei Präferenzdaten mit variabler Auswahl eignet sich die Faktorenanalyse.
- *mehrdimensionale Items*: (konfirmatorische) mehrdimensionale Skalierung, zum Beispiel mittels nichtmetrischer Skalierung nach Kruskal (siehe Kapitel 4 und Abschnitt 4.6.2).
- *hierarchische Items*: Item-Response-Modelle, wie zum Beispiel Guttman-Skalierung, Mokkenskalierung oder Raschverfahren (siehe Abschnitt 5.3.1).

18.1.2 Räumliche Darstellung von Objekten oder Variablen

Wird eine *räumliche Darstellung von Objekten oder Variablen* in einem niedrigdimensionalen Raum gesucht, kommen die in Teil I behandelten Verfahren in Frage, insbesondere die Hauptkomponentenmethode ohne anschließende Rotation, die multiple Korrespondenzanalyse und die nichtmetrische mehrdimensionale Skalierung oder die hier nicht behandelten metrischen Varianten (Borg 1981).

Liegt eine empirisch erhobene (Un-)Ähnlichkeitsmatrix vor, ist die mehrdimensionale Skalierung zu empfehlen, da sowohl die Korrespondenzanalyse als auch Hauptkomponentenmethode ein spezielles Ähnlichkeitsmaß voraussetzen. Bei der Korrespondenzanalyse sind dies die standardisierten Residuen, bei der Hauptkomponentenmethode Korrelationen oder Kovarianzen. Beim Vorliegen einer rechteckigen Datenmatrix, bestehend aus Objekten (zum Beispiel Personen) und Variablen (Objekt-Variablen-Datenmatrix), ist es zunächst sinnvoll, die multiple Korrespondenzanalyse oder die Hauptkomponentenmethode einzusetzen, da sie – abgesehen von ganz selten Grenzfällen – eine formal eindeutige Lösung besitzen und daher lokale Minima vermeiden. Liefern sie keine brauchbaren Ergebnisse, empfehlen wir die mehrdimensionale Skalierung. Ihr Einsatz empfiehlt

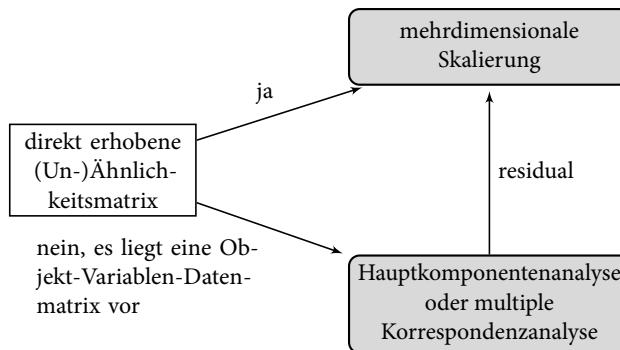


Abb. 18.1: *Entscheidungsbaum zur Auswahl eines geeigneten Verfahrens für die räumliche Darstellung von Objekten oder Variablen*

sich auch dann, wenn die Hauptkomponentenmethode oder die multiple Korrespondenzanalyse zu keinen brauchbaren Ergebnissen führen. Bei der mehrdimensionalen Skalierung sind allerdings lokale Minima nicht ausgeschlossen, da Startwerte erforderlich sind. Reduzieren lässt sich diese Gefahr dadurch, dass mehrere Startkonfigurationen gerechnet werden. Im vorausgehenden Abschnitt wurden zwei Voraussetzungen für die multiple Korrespondenzanalyse und die Hauptkomponentenmethode genannt: Es müssen Ratingdaten vorliegen und die Items dürfen nicht mehrdimensional sein. Diese beiden Voraussetzungen sind nur erforderlich, wenn abgeleitete Variablen (Faktorwerte) gebildet werden sollen. Zum Auffinden einer Darstellung in einem niedrigdimensionalen Raum sind sie nicht erforderlich. Allerdings ist bei diesen Datenkonstellationen eine Interpretation der Dimensionen nicht mehr möglich und sinnvoll. Als Orientierung für die Praxis kann der Entscheidungsbaum der Abbildung 18.1 dienen.

18.1.3 Auffinden einer hierarchischen Ähnlichkeitsstruktur

Die hierarchisch-agglomerativen Verfahren (siehe Abschnitt 6.4) eignen sich auch zum Auffinden einer hierarchischen Ähnlichkeitsstruktur in Form eines Baums (Dendrogramms). Als Standardverfahren empfehlen wir hierfür den Weighted-Average-Linkage, da bei diesem Verfahren das Verschmelzungsniveau eine klare inhaltliche Interpretation besitzt. Gegenüber dem Single- und Complete-Linkage hat das Verfahren den Vorteil, dass alle (Un-)Ähnlichkeiten in die Berechnung eingehen und durch die Verwendung von Mittelwerten Fehler ausgeglichen werden, während Complete- und Single-Linkage nur auf dem unähnlichsten bzw. ähnlichssten Paar basieren. Der Einsatz des Ward-, Zentroid- oder Medianverfahrens ist nicht sinnvoll, da sie mit der quadrierten euklidischen Distanz

ein spezielles Distanzmaß erfordern und ihr Ziel das Auffinden von Clusterprofilen ist. Der Within-Average-Linkage hat den Nachteil, dass Inversionen auftreten können.

Complete- und Single-Linkage sollten dann eingesetzt werden, wenn die gesuchte Ähnlichkeitsstruktur invariant gegenüber monotonen Transformationen sein soll, das heißt, wenn sich die hierarchische Ähnlichkeitsstruktur des Dendrogramms nicht ändern soll, wenn die empirischen (Un-)Ähnlichkeiten quadriert, logarithmiert usw. werden.

Steht ein Verfahren zur Bildung eines Konsensus-Diagramms zur Verfügung, kann auch dieses eingesetzt werden. Dies hat den Vorteil, dass sich die Anwenderin für kein bestimmtes Verfahren entscheiden muss. Zudem wird ein »Durchschnitt« aus mehreren Lösungen berechnet und damit werden Fehler ausgeglichen.

18.1.4 Räumliche oder hierarchische Darstellung?

Zur Analyse der Ähnlichkeiten von Objekten oder Variablen bestehen drei Möglichkeiten:

1. Ermittlung zugrunde liegender gemeinsamer Dimensionen bzw. Faktoren mittels Faktoren- oder multipler Korrespondenzanalyse bzw. nominaler Faktorenanalyse nach McDonald.
2. Darstellung in einem niedrigdimensionalen Raum mittels mehrdimensionaler Skalierung, multipler Korrespondenzanalyse oder Hauptkomponentenmethode ohne Rotation.
3. Darstellung als Baum in Form eines Dendrogramms mittels eines hierarchisch-agglomerativen Verfahrens.

Die erste Variante setzt voraus, dass zumindest eine Teilmenge der Variablen (oder Objekte) eindimensional misst, also nur auf einen Faktor lädt. Zudem muss es sich um Ratingdaten handeln (siehe oben). Sind die Voraussetzungen erfüllt, empfehlen wir diese Variante. Die Faktoren- und multiple Korrespondenzanalyse sind robuste Verfahren. Sie haben eine formal eindeutige Lösung, die Gefahr eines lokalen Minimums oder Maximums besteht nicht. Sind die Voraussetzungen nicht erfüllt, ist die erste Variante nicht anwendbar. Eingesetzt werden kann entweder ein räumliches Verfahren oder eine hierarchische Clusteranalyse.

Bei der räumlichen Darstellung wird angenommen, dass sich die Variablen oder Objekte als Punkte in einem niedrigdimensionalen Raum darstellen lassen. Häufig gesucht wird eine zweidimensionale Repräsentation.

Bei der dritten Option wird vermutet, dass jedes Objekt Element eines Baumes ist. Mit der Frage, ob den Daten ein räumliches Modell oder eine Hierarchie als Baum angemessen ist, hat sich theoretisch in den 1970er Jahren Holman (1972) auseinandergesetzt. Er konnte zeigen, dass sich beide Modelle ausschließen. Einer Menge von p Objekten oder Variablen ist entweder ein hierarchisches Clustermodell angemessen oder die Darstellung in einem euklidischen Raum mit weniger als $p - 1$ Dimensionen. Die Simulationsstudien von Pruzansky, Tversky u. a. (1982) bestätigen diese Schlussfolgerungen. Sie weisen nach, dass das für Daten geeigneteren Verfahren (MDS oder hierarchische Clusteranalyse) auch bei »Fehlerüberlagerungen« bessere Modellanpassungen erbringt. Bei zunehmendem Fehlerausmaß verschwinden aber die Unterschiede der beiden Verfahren. Dies dürfte ein Grund sein, warum in der Forschungspraxis oft beide Verfahren zu gleich guten Ergebnissen führen. Wir empfehlen daher – sofern inhaltliche Zielsetzungen nicht den Einsatz eines bestimmten Verfahrens erfordern – beide Verfahren einzusetzen und jenes Modell (räumlich oder hierarchisch) mit dem besten Modellfit auszuwählen. Zur Beurteilung der Modellgüte sollte eine Maßzahl Anwendung finden, die nicht von einem der beiden Verfahren zur Schätzung benutzt wird. Der Stress-Koeffizient scheidet daher aus. Benutzt werden können kophenetische Korrelationskoeffizienten, beispielsweise γ oder r . Auch die Schiefe der untersuchten Unähnlichkeiten gibt nach Pruzansky, Tversky u. a. (1982) Hinweise auf das geeignete Modell. Für Daten, für die eine räumliche Darstellung geeignet ist, nimmt die Schiefe einen positiven Wert an, für hierarchische Daten einen negativen.

Beim Einsatz der mehrdimensionalen Skalierung ist auf die Anwendungsvoraussetzungen Acht zu geben. Für das Auffinden einer stabilen Punktkonfiguration ist eine bestimmte Variablenzahl bzw. eine bestimmte Zahl empirischer (Un-)Ähnlichkeiten erforderlich. Sixtl (1982) nennt für die nichtmetrische mehrdimensionale Skalierung ein Verhältnis von $1 : 5$. Das heißt für eine zweidimensionale Lösung sind zehn Variablen erforderlich (siehe Abschnitt 4.3). Die hierarchisch-agglomerativen Verfahren treffen dagegen keine derartigen Annahmen und können daher auch für eine kleinere Objektmenge eingesetzt werden.

18.1.5 Klassifikation von Variablen

Zur Klassifikation von Variablen empfehlen wir den Einsatz des Weighted-Average-Linkage aus folgenden Gründen:

- Im Unterschied zum Single- und Complete-Linkage wird die gesamte vorhandene Information genutzt und alle (Un-)Ähnlichkeiten gehen in die Berechnung ein. Es wird mit »Mittelwerten« gerechnet und damit werden Fehler ausgeglichen.

- Im Unterschied zum Average-Linkage, der wie der Weighted-Average-Linkage den Mittelwertverfahren angehört und alle (Un-)Ähnlichkeiten verwendet, besitzt das Verschmelzungsniveau beim Weighted-Average-Linkage eine klare inhaltliche Interpretation.
- Im Unterschied zum Within-Average-Linkage treten keine Inversionen auf.
- Ward-, Zentroid- und Median- und K-Means-Verfahren scheiden aus, da sie die quadrierte euklidische Distanz als Unähnlichkeitsmaß voraussetzen. Zur Messung der Ähnlichkeit von Variablen sind aber Korrelationsmaße sinnvoll.
- Hinzukommt, dass beim Ward-, Zentroid-, Median- und K-Means-Verfahren die Konstruktion von Clusterprofilen (Mittelwerte der Cluster in den Variablen) im Vordergrund steht. Für eine Klassifikation von Variablen sind Clusterprofile nicht sinnvoll. Abgebildet würden die Mittelwerte der zu Clustern zusammengefassten Variablen je Person. Bei $n = 1\,000$ Befragten und $K = 3$ Variablenclustern würde sich eine $1\,000 \times 3$ Matrix der Clustermittelwerte ergeben.
- Da Clusterprofile nicht sinnvoll sind, scheiden auch probabilistische Verfahren aus.

Soll die gesuchte Klassifikation invariant gegenüber monotonen Transformationen sein, scheiden die Mittelwertverfahren aus und der Complete- oder Single-Linkage sind zu wählen, wobei wegen der strengeren Homogenitätsvorstellungen der Complete-Linkage zu bevorzugen ist, wenn es um das Auffinden einer Klassifikation geht.

18.1.6 Klassifikation von Objekten

Zur Klassifikation von Objekten eignen sich – abhängig von den verfügbaren Daten – folgende Verfahren:

- Mittelwertverfahren (Weighted-Average-Linkage), wenn eine direkt erhobene (Un-)Ähnlichkeitsmatrix vorliegt oder bei einer Objekt-Variablen-Datenmatrix die Verwendung der quadrierten euklidischen Distanz als Unähnlichkeitsmaß nicht angebracht ist.
- Ward-Verfahren, wenn eine Objekt-Variablen-Datenmatrix vorliegt und quadrierte euklidische Distanzen angewendet werden können. Die Zahl der zu klassifizierenden Objekte sollte nicht zu groß sein, da sonst die Gefahr von Bindungen in frühen Verschmelzungsschritten besteht, die die Ergebnisse beeinflussen. Gegenüber dem Zentroid- und Medianverfahren hat das Ward-Verfahren den Vorteil, dass keine Inversionen im Verschmelzungsschema auftreten.
- K-Means-Verfahren, wenn eine Objekt-Variablen-Datenmatrix vorliegt und quadrierte euklidische Distanzen angewendet werden können. Die Zahl der zu klassifizierenden Objekte sollte eine bestimmte Mindestgröße überschreiten.

- Probabilistische Verfahren, wenn eine Objekt-Variablen-Datenmatrix vorliegt. Auch hier ist eine gewisse Mindestgröße des analysierten Samples erforderlich.

Als Standardverfahren empfehlen wir für die Klassifikation von Objekten beim Vorliegen einer bestimmten Mindestgröße der Untersuchungspopulation probabilistische Verfahren. Sie haben folgende Vorteile:

- Das Problem der Inkommensurabilität der Variablen wird automatisch bei der Modellierung gelöst.
- Die Auswahl eines bestimmten (Un-)Ähnlichkeitsmaßes entfällt. Dies ist allerdings auch beim Ward-, Zentroid- und Medianverfahren sowie dem K-Means-Verfahren nicht erforderlich.
- Es können unterschiedliche Variablentypen modelliert werden.
- Es können Messfehler modelliert werden.
- Die Verfahren sind weniger sensibel gegenüber irrelevanten Variablen als die deterministischen Verfahren.
- Es stehen formale Maßzahlen zur Auswahl einer geeigneten Clusterzahl zur Verfügung.
- Die Clustermittelwerte werden unverzerrter geschätzt als bei den deterministischen Verfahren. Das K-Means-Verfahren als deterministisches Verfahren berechnet eine optimale Partition. Bei Überlappungen führt dies zu verzerrten Schätzungen der Clustermittelwerte.

Probabilistische Verfahren setzen eine bestimmte Mindestgröße der Untersuchungspopulation voraus. Richtwerte hierfür gibt es nicht. Die erforderliche Mindestgröße hängt von der Zahl der Variablen und deren Qualität sowie von dem zugrunde liegenden Clustermodell ab. Liegen sehr viele Variablen vor, die zudem sehr gut gemessen werden, ist auch eine kleine Stichprobe ausreichend. Umgekehrt wird eine sehr große Stichprobe benötigt, wenn sehr wenige und zudem noch fehlerbehaftete Variablen vorhanden sind. Gleicher gilt für die zugrunde liegende Clusterstruktur. Sind die Cluster gut getrennt, reicht eine kleine Stichprobe aus. Überlappen sich die Cluster oder sind kleine Cluster vorhanden, ist eine größere Untersuchungspopulation erforderlich. Bei den in Abschnitt 17.3 berichteten Simulationsstudien erbringen fast alle probabilistischen Verfahren bereits für $n = 500$ befriedigende Ergebnisse für drei Variablen, wenn zwei gut getrennte Cluster vorliegen. Dagegen sind sechs Variablen und mindestens 2 000 Fälle für eine 5-Clusterstruktur mit zwei sehr kleinen Clustern erforderlich, wenn mit Latent GOLD, dem Programm mit der besten Performanz, gerechnet wird.

Da es keine allgemeinen Richtwerte für die erforderliche Stichprobengröße gibt, empfehlen wir die Durchführung von Stabilitätstests. Erweist sich die gefundene Clusterlösung als instabil, sollte ein deterministisches Verfahren eingesetzt werden. Hier empfehlen wir, zunächst mit den K-Means-Verfahren zu arbeiten. K-Means-Verfahren haben gegenüber

den zur Analyse verbleibenden hierarchischen Clusterverfahren den Vorteil, dass sie für eine gegebene Clusterzahl K eine optimale Lösung berechnen. Bei den hierarchischen Verfahren kann die Lösung dagegen suboptimal sein, da in einem vorausgehenden Verschmelzungsschritt eine »falsche« Entscheidung getroffen wurde. Auch die K-Means-Verfahren setzen eine bestimmte Mindeststichprobengröße, für die keine allgemeinen Schwellenwerte existieren, voraus. Daher sollte wie bei den probabilistischen Verfahren die Stabilität untersucht und bei Instabilität ein hierarchisches Verfahren eingesetzt werden. Hier empfehlen wird das Ward-Verfahren. Das Verschmelzungsniveau hat eine klare Interpretation und Inversionen im Verschmelzungsniveau treten im Unterschied zum Zentroid- und Medianverfahren nicht auf.

Soll ein bestimmtes (Un-)Ähnlichkeitsmaß verwendet werden und handelt es sich dabei um ein Distanzmaß, das von K-Means-Verfahren nicht unterstützt wird,¹ dann empfehlen wir den Einsatz des Weighted-Average-Linkage bzw. – wenn Invarianz gegenüber monotonen Transformationen erwünscht ist – den Complete- oder Single-Linkage, wobei der Complete-Linkage wegen der strenger Homogenitätsvorstellungen zu bevorzugen ist.

Die Auswahl eines Clusteranalyseverfahrens zur Klassifikation von Objekten kann sich an dem in Abbildung 18.2 dargestellten Entscheidungsbaum orientieren. Zunächst ist zu prüfen, ob der Einsatz eines speziellen, durch K-Means- oder dem Ward-Verfahren nicht unterstützten (Un-)Ähnlichkeitsmaßes erwünscht ist, was beispielsweise bei dichotomen Variablen der Fall sein kann. Bei »ja« sollte – abhängig, ob Invarianz gegenüber monotonen Transformationen erwünscht ist – mit Weighted- oder Within-Average-Linkage oder mit Complete- oder Single-Linkage gerechnet werden. Bei »nein« sollte ein probabilistisches Verfahren eingesetzt werden, außer die Stichprobe ist offensichtlich zu klein und umfasst zum Beispiel nur $n = 10$, $n = 20$ oder $n = 30$ Fälle. Bei »Unbrauchbarkeit« der probabilistischen Verfahren, zum Beispiel wegen mangelnder Stabilität, sollte K-Means eingesetzt werden. Ist auch K-Means »unbrauchbar«, wird das Ward-Verfahren empfohlen. Neben der fehlenden Stabilität kann eine Clusterlösung auch unbrauchbar sein, wenn sie inhaltlich nicht interpretierbar oder formal oder inhaltlich invalide ist.

18.2 Verwendung aller Variablen?

Die Möglichkeit, sowohl bei den deterministischen als auch bei probabilistischen Clusteranalyseverfahren Objekte mit gemischten Variablen zu clustern, darf nicht dahingehend

¹ Standard-Implementationen des K-Means-Verfahrens verwenden nur quadrierte euklidische Distanzen. Mitunter gibt es aber auch Software, bei der mit der City-Block-Metrik gerechnet werden kann (siehe Abschnitt 12.14).

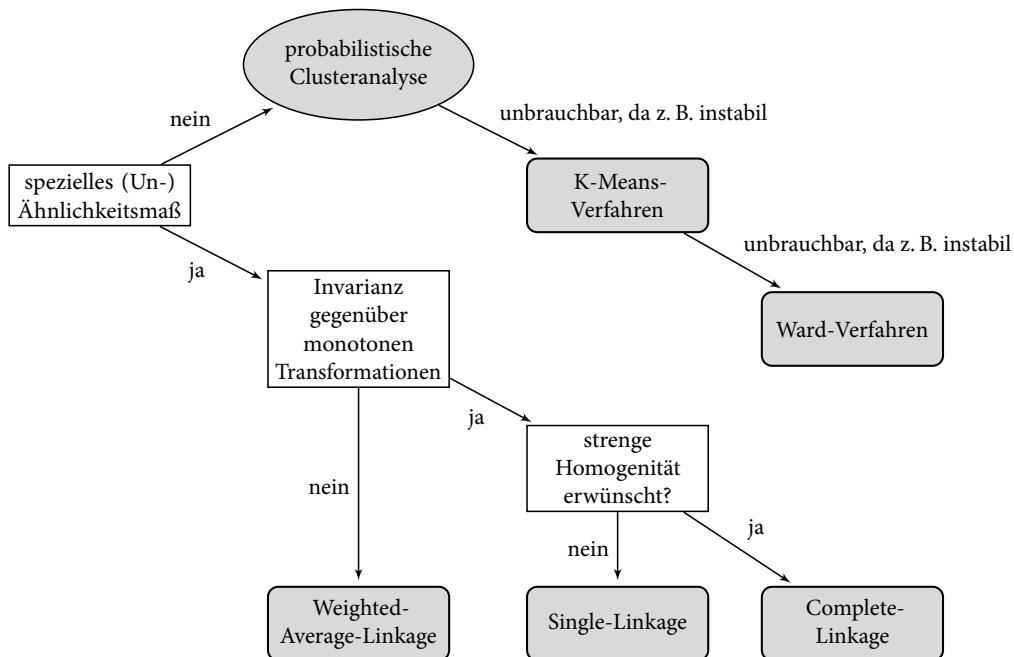


Abb. 18.2: Entscheidungsbaum zur Auswahl eines geeigneten Verfahrens für eine objektorientierte Clusteranalyse

missverstanden werden, in einer Analyse alle erhobenen Variablen für eine Klassifikation der untersuchten Objekte zu verwenden. Ein derartiges Vorgehen ist in der Regel wenig sinnvoll, da das Klassifikationsziel nicht definiert ist, eine kaum interpretierbare Anzahl von Clustern oder Klassen entstehen würde und bei deterministischen Verfahren die Gefahr besteht, dass irrelevante Variablen die Clusterstruktur zerstören.

Allgemein empfehlen wir daher:

1. Anstelle der ursprünglichen Variablen sollten abgeleitete Variablen verwendet werden (siehe Abschnitt 18.1).
2. Es sollten inhaltlich zusammengehörende Variablen oder Variablengruppen analysiert werden.

Die Bildung von abgeleiteten Variablen wurde bereits ausführlich erörtert. Bei der Auswahl von inhaltlich zusammengehörenden Variablen kann man sich am Modell² in Abbildung 18.3 auf der nächsten Seite, das den meisten sozialwissenschaftlichen Erhebungen

² Rückkoppelungen wurden nicht eingetragen, da diese bei Querschnittsdaten nur unter bestimmten Bedingungen geschätzt werden können.

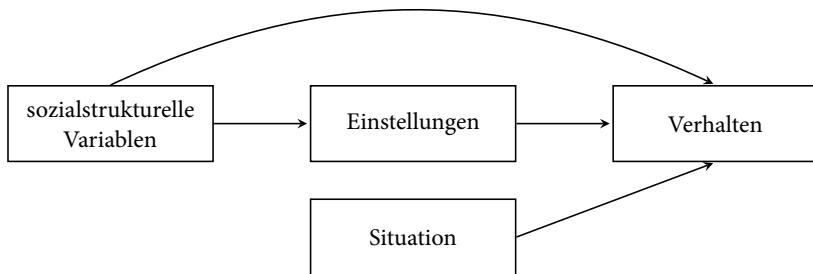


Abb. 18.3: Modell sozialwissenschaftlicher Erhebungen

zugrunde liegt, orientieren. Das Modell enthält vier Variablengruppen: sozialstrukturelle Variablen, Einstellungsvariablen, Situationsvariablen und Verhaltensvariablen. Mit Hilfe dieser vier Variablengruppen ist auf der Individualebene die Bildung von vier unterschiedlichen Klassifikationen möglich:

1. *Klassifikation der Befragten (Objekte) aufgrund ihrer sozialstrukturellen Merkmale:* Das Klassifikationsziel ist die Bestimmung sozialstruktureller Typen. Die Einstellungs-, Situations- und Verhaltensvariablen können zur Beschreibung und inhaltlichen Validitätsprüfung der sozialstrukturellen Typen verwendet werden.
2. *Klassifikation der Befragten aufgrund ihrer Einstellungsvariablen:* Das Klassifikationsziel ist hier die Bestimmung von Einstellungstypen. Dabei wird man zunächst die Einstellungselemente faktorenanalytisch untersuchen und in die Clusteranalyse nicht die Items selbst, sondern die gebildeten Gesamtpunkt- oder Faktorwerte einbeziehen, sofern die Annahme der Einfachstruktur zutrifft (siehe dazu Abschnitt 5.3.1). Die Variablen der anderen Variablengruppen können wiederum zur Deskription und Validitätsprüfung verwendet werden.
3. *Klassifikation der Befragten aufgrund ihrer Situationsvariablen:* Das Klassifikationsziel ist hier die Bestimmung von Situationstypen. Die Variablen der anderen Variablengruppen können wiederum zur Deskription und inhaltlichen Validitätsprüfung verwendet werden.
4. *Klassifikation der Befragten aufgrund ihres Verhaltens:* Das Klassifikationsziel ist hier die Bestimmung von Verhaltenstypen. Die Variablen der anderen Variablengruppen können wiederum zur Deskription und inhaltlichen Validitätsprüfung verwendet werden.

Die hier dargestellte Vorgehensweise ist als Orientierungshilfe zu verstehen. Entscheidend sind letztlich inhaltliche Überlegungen. Besteht diese – wie etwa in der Lebensstilforschung oft der Fall – darin, dass eine bestimmte Typologie durch sozialstrukturelle Variablen sowie durch bestimmte Einstellungen und Verhaltensweisen charakterisiert ist, wird man selbstverständlich die entsprechenden Variablen in die Clusteranalyse einbe-

ziehen. Steht eine ausreichend große Zahl an Variablen zur Verfügung, ist es ratsam, sich bestimmte Variablen für die inhaltliche Validitätsprüfung aufzuheben. Zu empfehlen ist auf jeden Fall eine inhaltlich gut begründete Variablenauswahl, am besten wäre ohnedies die Durchführung einer konfirmatorischen Clusteranalyse.

18.3 Welches Un- bzw. Ähnlichkeitsmaß?

Die Frage nach der Auswahl eines geeigneten (Un-)Ähnlichkeitsmaßes stellt sich:

1. beim Auffinden einer hierarchischen Ähnlichkeitsmatrix beim Vorliegen einer Objekt-Variablen-Datenmatrix,
2. bei der Klassifikation von Variablen,
3. bei der Klassifikation von Objekten, wenn die in K-Means- und im Ward-Verfahren vorgesehenen Distanzmaße aus inhaltlichen Überlegungen oder anderen Gründen nicht geeignet sind, was den Ausnahmefall darstellt.

Für die Auswahl entscheidend ist, ob Variablen oder Objekte geclustert werden. Bei Objekten sollten Distanzmaße oder daraus abgeleitete Maße eingesetzt werden, außer eine Elimination der Profilhöhe und Profilstreuung ist erwünscht (siehe Abschnitt 7.6). Im letzteren Fall können Korrelationsmaße verwendet werden. Für eine variablenorientierte Clusteranalyse sind dagegen Korrelationsmaße und daraus abgeleitete Maße geeignet, da hier die Profilhöhe (Mittelwerte der Variablen) und die Profilstreuung (Varianzen der Variablen) uninteressant sind. Sind letztere von Interesse, sollten Distanzmaße eingesetzt werden.

Aufgabenstellung 1 kann eine variablen- oder objektorientierte Aufgabe sein. Das Ziel kann im Auffinden einer hierarchischen Ähnlichkeitsbeziehung von Variablen oder von Objekten bestehen. Bei einer variablenorientierten Problemstellung wird man Korrelations- oder daraus abgeleitete Maßzahlen verwenden, bei einer objektorientierten Distanzmaße oder daraus abgeleitete Maßzahlen.

Die sich aus der allgemeinen Minkowskimetrik (Formel 8.12 auf Seite 219) ergebenden Distanzmaße hängen von den Metrikparametern r und q ab. Der Metrikparameter r bewirkt eine Gewichtung der Unterschiede in den Variablen, der Metrikparameter q eine Reskalierung. Je größer r gewählt wird, desto größeres Gewicht haben größere Unterschiede in einer Variablen. Ist diese Gewichtung nicht erwünscht, sollte die City-Block-Metrik eingesetzt werden. Andernfalls empfohlen wird die euklidische Distanz oder die quadrierte euklidische Distanz, da sie wie die City-Block-Metrik inhaltlich gut

interpretierbar sind.³ Die euklidische Distanz ist die geradlinige Entfernung (»Fluglinie«) zwischen zwei Objekten. Die quadrierte euklidische Distanz hat einen klaren Bezug zur Varianz. Bei Anwendung von hierarchischen Verfahren ist zu beachten, dass diese mit Ausnahme von Single- und Complete-Linkage nicht invariant gegenüber monotonen Transformationen sind. Das Ward-Verfahren führt also zu abweichenden Ergebnissen, wenn mit der euklidischen Distanz oder der quadrierten euklidischen Distanz gerechnet wird. Die quadrierte euklidische Distanz ist das korrekte Distanzmaß, damit im Verschmelzungsniveau der Zuwachs der Fehlerstreuung abgebildet wird.

Bei den Korrelationsmaßen ist, der Forschungspraxis folgend, die Anwendung der Produkt-Moment-Korrelation oder eines anderen Korrelationskoeffizienten empfehlenswert. Eine Sonderstellung nehmen dichotome Variablen ein. Für sie wurde eine Vielzahl von (Un-)Ähnlichkeitsmaßen entwickelt, die sich sowohl für eine objekt- als auch für eine variablenorientierte Analyse eignen. Die Maßzahlen unterscheiden sich darin, wie

- der gemeinsame Besitz (das gemeinsame Auftreten der 1-Ausprägungen),
- der gemeinsame Nicht-Besitz (das gemeinsame Auftreten der 0-Ausprägungen) und
- die Nichtübereinstimmung (das Auftreten der Merkmalskombinationen (0,1) und (1,0))

gewichtet werden. Bei einer Inhaltsanalyse kann es beispielsweise sinnvoll sein, den gemeinsamen Besitz stärker zu gewichten (siehe dazu Abschnitt 8.2). Auch bei ordinalen Variablen besteht die Möglichkeit, die Ausprägungen unterschiedlich zu gewichten. Ist dies erwünscht, sollte die Canberra-Metrik oder der Jaccard-II-Koeffizient Anwendung finden. Zusammenfassend kann man den in Abbildung 18.4 dargestellten Entscheidungsbaum konstruieren und befolgen.

18.4 Wie viele Cluster?

Die Bestimmung der Zahl der Cluster ist nach wie vor die »Achillesverse« der Clusteranalyse. In der probabilistischen Clusteranalyse stehen mit den Informationsmaßen zwar formal begründete Entscheidungskriterien zur Verfügung. Allerdings liegen die Werte der Informationsmaße für unterschiedliche Lösungen oft sehr nahe beieinander, so dass es schwer fällt, die Unterschiede als bedeutsam zu erachten. Hinzukommt, dass unterschiedliche Informationsmaße verschiedene Lösungen nahe legen können. Auch für die K-Means-Verfahren liegen Maßzahlen zur Bestimmung der Clusterzahl vor. Aber auch

³ Gut interpretierbar ist auch noch die sogenannte *Chebychev-Metrik*, bei der nur der maximale Unterschied in einer Variablen in die Distanzen eingeht. Allerdings geht hier nur ein kleiner Teil der verfügbaren Informationen in die Berechnung ein.

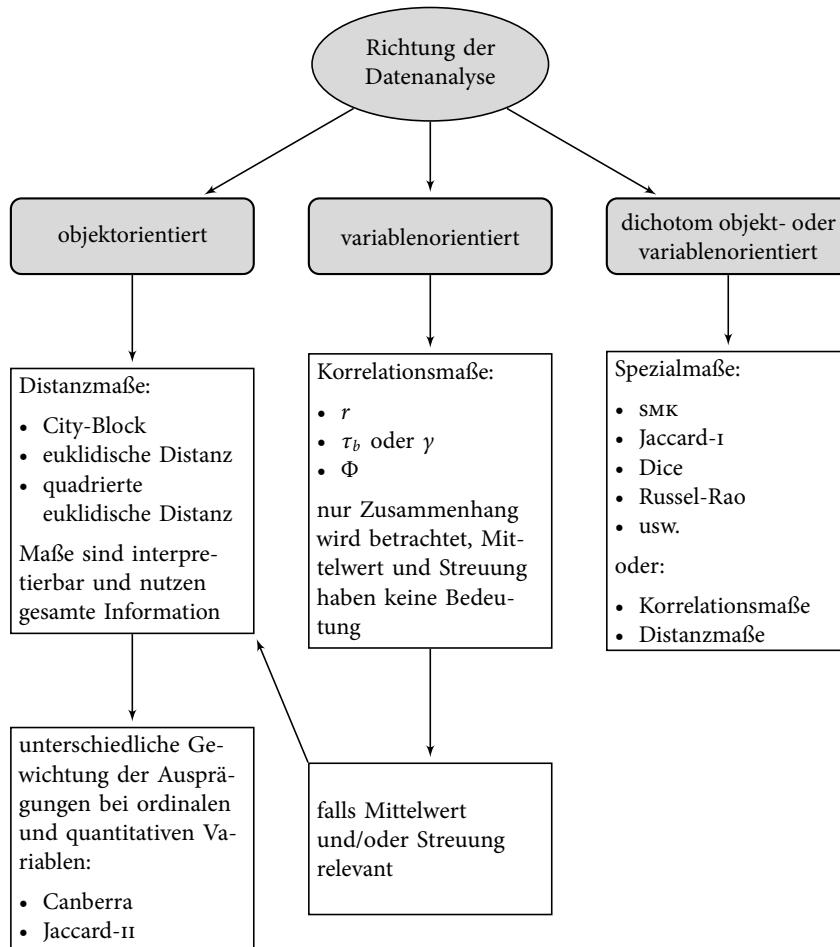


Abb. 18.4: Entscheidungsbaum für Auswahl eines Un- oder Ähnlichkeitsmaßes

hier tritt häufig der Fall ein, dass mehrere Lösungen als brauchbar erscheinen. Gleichermaßen gilt für die hierarchischen Verfahren.

Relativ gute Erfahrungen haben wir mit der prozentuellen Veränderungen PV_k gegenüber einer vorausgehenden Lösung gemacht (siehe Abschnitte 12.2, 14.2 und 16.5) bzw. bei der hierarchischen Clusteranalyse auch mit dem Distanzzuwachs (siehe Abschnitt 6.6). Aber auch dieses Kriterium führt häufig zu dem Resultat, dass mehrere Lösungen formal zulässig sind. Wir empfehlen daher, in diesem Schritt nicht zu restriktiv zu sein und alle formal zulässigen Clusterlösungen weiterzuverfolgen und zu untersuchen, ob die anderen Kriterien an eine gute Klassifikation, nämlich inhaltliche Interpretierbarkeit,

Stabilität sowie formale und inhaltliche Gültigkeit erfüllt sind. Ein Verfahren zur Prüfung der formalen Gültigkeit, das die Stabilität mit einschließt, wird in Kapitel 20 behandelt.

18.5 Modale Klassenzugehörigkeit oder Zuordnungswahrscheinlichkeiten?

Soll mit den Ergebnissen der probabilistischen Clusteranalyse weitergerechnet werden, bestehen drei Möglichkeiten:

1. Verwendung der modalen Clusterzugehörigkeit
2. Verwendung der Zuordnungswahrscheinlichkeiten
3. Gewichtung der Fälle mit den Zuordnungswahrscheinlichkeiten

Bei der modalen Clusterzugehörigkeit wird jedes Objekt g jenem Cluster k zugeordnet, zu dem die Zuordnungswahrscheinlichkeit $\pi_{k|g}$ ein Maximum ist. Dass das Maximum bei zwei Klassen auftritt, ist extrem unwahrscheinlich, da mit vielen Dezimalstellen gerechnet wird, auch wenn diese im Output nicht angezeigt werden.⁴ Die modale Clusterzugehörigkeit ist eine nominalskalierte Variable V mit Ausprägungen von 1 bis K . Der erste Zugang führt also dazu, dass mit einer nominalen Variablen gerechnet wird. Bei der zweiten Option liegt bei K Clustern für jeden Fall eine quantitative Variable V_k für jedes Cluster k vor, die als Werte die Zuordnungswahrscheinlichkeiten $\pi_{k|g}$ enthält. Pro Fall gibt es also K quantitative Variablen, während bei der ersten Methode nur eine nominale Variable vorliegt. Zwischen den Variablen bestehen lineare Abhängigkeiten, da die Summe der Zuordnungswahrscheinlichkeiten gleich 1 ist. Bei multivariaten Analysen mit der Klassenzugehörigkeit als unabhängiger Variablen, in denen die Zuordnungswahrscheinlichkeiten einbezogen werden, muss daher eine Variable als Referenzvariable bzw. -kategorie definiert und aus der Analyse ausgeschlossen werden. Dabei ist es sinnvoll, ein Cluster als Referenzkategorie auszuwählen, das inhaltlich gut interpretierbar, stabil und valide ist, da die Ergebnisse in Bezug auf dieses Cluster interpretiert werden (zum Beispiel: »Im Vergleich zum Referenzcluster wirkt Cluster k^* stärker auf die Variable z ein«). Die Verwendung eines nicht interpretierbaren Restclusters als Referenzcluster⁵ ist nicht sinnvoll, da dann auch alle Ergebnisse, bei denen andere Cluster mit diesem in Beziehung gesetzt werden, nicht interpretierbar sind. Eine dritte Möglichkeit besteht in der Gewichtung der Fälle. Gebildet wird eine nominale Variable V mit K Ausprägungen.

⁴ Tritt dieser unwahrscheinliche Fall dennoch ein, kann zufällig entschieden oder der Fall aus der Analyse ausgeschlossen werden.

⁵ Dies kann etwa als Folge eines automatischen Vorgehens, bei dem die erste und die letzte Ausprägung automatisch als Referenzkategorie ausgewählt werden, eintreten.

Jeder Fall wird bei K Clustern K -mal in die Analyse einbezogen, zuerst mit der Ausprägung $k = 1$ und der Zuordnungswahrscheinlichkeit $\pi_{k=1|g}$, dann mit der Ausprägung $k = 2$ und der Zuordnungswahrscheinlichkeit $\pi_{k=2|g}$ usw. Diese Methode wird in Bacher und Vermunt (2010) beschrieben. Daher sollen nachfolgend die beiden zuerst genannten Ansätze näher behandelt werden.

Für welche der ersten beiden Varianten man sich entscheidet, ist zum Teil Geschmackssache. Sollen Tabellenanalysen gerechnet werden, ist die Verwendung der modalen Clusterzugehörigkeit eleganter, da nur mit einer Variablen gerechnet werden muss. Dies ist auch bei der dritten Variante der Fall. Aber auch hier ist es möglich, die Zuordnungswahrscheinlichkeiten als Gewichte zu nutzen (Bacher und Vermunt 2010). Sollen dagegen Korrelationen oder weiterführende Analysen gerechnet werden, ist der Einsatz von Zuordnungswahrscheinlichkeiten zu empfehlen. Die nominale Variable »modale Clusterzugehörigkeit« muss ohnedies in K Dummies aufgelöst werden. Im Unterschied zu den Dummies enthalten die Zuordnungswahrscheinlichkeiten aber mehr Information, da sie Werte zwischen 0 und 1 haben können, während die Dummies nur gleich 0 oder 1 sind.

Wir empfehlen daher allgemein die Verwendung der Zuordnungswahrscheinlichkeiten außer bei Tabellenanalysen, wo durch Verwendung einer nominalen Variablen eine übersichtliche Information erzielt wird und die Zahl der zu berechnenden Tabellen deutlich reduziert werden kann. Bei K Clustern und q externen Variablen, die mit den Clustern kreuztabelliert werden sollen, müssen anstelle von $K \cdot q$ Tabellen nur q Tabellen berechnet werden. Zum Beispiel für $K = 3$ und $q = 10$ müssen bei Verwendung der Zuordnungswahrscheinlichkeiten 30 Tabellen berechnet werden, bei der modalen Zuordnung nur 10.⁶

Das Wechseln der Methoden ist legitim. Empirisch korrelieren die modale Clusterzugehörigkeit und die Zuordnungswahrscheinlichkeiten in der Regel relativ hoch miteinander, so dass sowohl das Wechseln als auch die Verwendung einer Methode gerechtfertigt ist. Für die Denzdaten ergeben sich folgende Korrelationen:

- *2-Clusterlösung: $r = 0,936$ zwischen modaler Clusterzugehörigkeit (0/1- Dummy) und Zuordnungswahrscheinlichkeit für beide Cluster,*
- *3-Clusterlösung: $r = 0,947; r = 0,952; r = 0,954$ zwischen modaler Clusterzugehörigkeit (0-1 Dummy) und Zuordnungswahrscheinlichkeit für die drei Cluster,*
- *5-Clusterlösung: $r = 0,915; r = 0,924; r = 0,918; r = 0,947; r = 1,00$ zwischen modaler Clusterzugehörigkeit (0/1-Dummy) und Zuordnungswahrscheinlichkeit für die drei Cluster.*

⁶ Werden die Zuordnungswahrscheinlichkeiten als Gewichte verwendet (Variante 3), müssen ebenfalls nur q Tabellen berechnet werden.

Die Korrelationen sind deutlich höher 0,90, so dass sich die Ergebnisse der beiden Methoden nur geringfügig unterscheiden.

Allerdings zeigt sich in dem Beispiel auch, dass sich für die Zuordnungswahrscheinlichkeiten etwas höhere Korrelationen mit externen Kriterienvariablen ergeben als für die modalen Zuordnungswahrscheinlichkeiten. Für die 3-Clusterlösung ergeben sich folgende Beziehungen:

$$\begin{aligned} r_{1x} &= 1,2165 \cdot r_{1x}^m + 0,0082; & R^2 &= 0,9518 \text{ für Cluster 1 ,} \\ r_{2x} &= 1,0240 \cdot r_{2x}^m + 0,0010; & R^2 &= 0,8580 \text{ für Cluster 2 ,} \\ r_{3x} &= 1,1722 \cdot r_{3x}^m - 0,0055; & R^2 &= 0,9918 \text{ für Cluster 3 .} \end{aligned}$$

Die Größe r_{kx} ist die Korrelation der Zuordnungswahrscheinlichkeiten des Clusters k mit einer externen Variablen x , r_{kx}^m ist die Korrelation des Dummy der modalen Clusterzugehörigkeit für Cluster k mit x . Für alle Cluster liegt ein starker linearer Zusammenhang vor. Die Korrelationen mittels Zuordnungswahrscheinlichkeiten mit externen Variablen und jener der modalen Clusterzugehörigkeit mit externen Variablen lassen sich durch eine lineare Gleichung ineinander überführen. Aufgrund der starken Korrelationen zwischen den beiden Methoden (siehe oben) in dem Beispiel sind auch keine anderen Ergebnisse zu erwarten. Die Gleichung für das erste Cluster bedeutet, dass sich bei einer Korrelation von beispielsweise $r_{kx}^m = 0,40$ für die modale Clusterzugehörigkeit eine Korrelation von $r_{1x} = 1,2165 \cdot 0,40 + 0,0082 = 0,4948$ für die Zuordnungswahrscheinlichkeit ergibt. Umgekehrt resultiert eine Korrelation von $r_{kx}^m = -0,40$ in $r_{1x} = -0,4948$. Da der unstandardisierte Regressionskoeffizient für alle drei Cluster immer größer 1,00 ist, bedeutet dies, dass sich bei Verwendung der Zuordnungswahrscheinlichkeiten stärkere Korrelationen ergeben. Erklärbar ist dies vermutlich durch die feinere Skalierung bei Verwendung der Zuordnungswahrscheinlichkeiten bzw. durch einen Kategorisierungseffekt beim Einsatz der modalen Klassenzugehörigkeiten.

19 Klassifikation von Verläufen mittels Optimal Matching

(Heinz Leitgöb)

19.1 Einführung

Ausgehend von Entwicklungen in der quantitativ orientierten Lebensverlaufsforschung lässt sich gegenwärtig ein verstärktes sozialwissenschaftliches Interesse an der *Klassifizierung von Verläufen* konstatieren. Dieser »Trend« gilt unter anderem als Reaktion auf die Kritik an der Übergangsfokussierung der Lebensverlaufsforschung (trotz konzeptionslogischer Priorität des Verlaufs begriffs), deren Ursache vornehmlich in der langjährigen Verfügbarkeit eines elaborierten Verfahrens zur Modellierung von Statusübergängen in Form der Ereignisdatenanalyse begründet ist (Blossfeld, Golsch u. a. 2007; Sackmann und Wingens 2001).¹

Verläufe werden in diesem Zusammenhang als *Sequenzen von Statuskonfigurationen* (Kluge und Kelle 2001, S. 5) verstanden, das heißt, sie repräsentieren eine zeitlich geordnete Abfolge von Ausprägungen einer in aller Regel nominalen² Zustandsvariablen (Abbott 1995). Die Zeit³ fungiert somit als das die Verläufe strukturierende Moment. Zudem wird ihr ein diskreter Charakter zugeschrieben, der sich über die Abfolge von realisierten Beobachtungs- bzw. Messpunkten manifestiert. Folglich können Ereignisse – Wechsel zwischen zwei Zuständen eines definierten Zustandsraumes – nicht unmittelbar zum Zeitpunkt ihres Vollzuges erfasst werden, sondern emergieren erst im Zuge der auf das

1 Demgegenüber erscheint eine holistische analytische Erfassung von (Lebens-)Verläufen erst seit der Implementierung der Sequenzdatenanalyse in den Sozialwissenschaften möglich.

2 Grundsätzlich können Verläufe ebenso für nichtnominal skalierte Variablen entwickelt werden. So zieht etwa Aisenbrey (2000) für ihre exemplarische Analyse von schulbezogenen Leistungsverläufen mit Schulnoten eine ordinale Variable heran. Dennoch beschränkt sich die vorliegende Einführung auf die sequenzdatenanalytische Behandlung von nominalen Variablen.

3 Theoretisch kann die Ordnung der Zustände innerhalb eines Verlaufs auch über ein anderes Merkmal als die Zeit erfolgen (zum Beispiel die Anordnung der vier Basen bei der DNA).

Ereignis folgenden Beobachtung⁴. Diese Diskrepanz relativiert sich, da – im Gegensatz zur Übergangsforschung – bei der Analyse von Verläufen explizite Informationen über die zeitliche Verweildauer in Zuständen nicht quantifiziert werden bzw. von keiner Relevanz sind. Vielmehr liegt der Fokus auf der Detektion von Verlaufsmustern und deren Gruppierung.

19.2 Methodische Grundlagen

Die Spezifikation von Verläufen als Sequenzen bedingt zunächst die Definition des (diskreten) Zustandsraumes einer die Sequenz erzeugenden Variable Z , der die Gesamtheit aller M möglichen Ausprägungen (Zustände) von Z repräsentiert: $Z = \{z_1, \dots, z_m, \dots, z_M\}$. Die Determinierung des Zustandsraumes sollte theorie- bzw. forschungsfragengeleitet erfolgen. In diesem Zusammenhang empfehlen Brüderl und Scherer (2004), diesen nur so differenziert wie nötig zu gestalten, da mit steigender Anzahl der Zustände die Komplexität der Analyse erheblich zunimmt (siehe diesbezüglich auch Buchmann und Sacchi 1995): Die theoretisch mögliche Anzahl an Sequenzen (s) der Länge T bei M Zuständen beträgt M^T , das heißt, erhöht sich M um a Zustände ($a > 0$) auf $M + a$, so steigt s um $(M + a)^T - M^T$. Relativierend muss allerdings festgehalten werden, dass in der Regel die Anzahl der empirisch beobachtbaren Sequenzen ohnehin erheblich unter s liegt. Zudem kann eine zu restriktive Definition des Zustandsraumes dazu führen, dass die Unterschiedlichkeit von relevanten Verlaufsmustern verborgen bleibt und schlussendlich die »wahre« Anzahl an Verlaufsclustern unterschätzt wird. Eine Sammlung von realisierten Anwendungen zur Definition des geeigneten Zustandsraumes stellen MacIndoe und Abbott (2009) zur Verfügung.

Die Länge einer Sequenz (die Summe der enthaltenen Zustände, die als Elemente der Sequenz bezeichnet werden) ergibt sich aus der Anzahl der T Beobachtungszeitpunkte. Somit kann eine Sequenz \mathbf{a} formal als Zeilenvektor angeschrieben werden: $\mathbf{a} = (a_1, \dots, a_t, \dots, a_T)$, wobei für jedes Element a_t von \mathbf{a} die Bedingung $a_t \in Z$ zutreffen muss. Analog kann eine Sequenz $\mathbf{b} = (b_1, \dots, b_t, \dots, b_T)$ formuliert werden, für die ebenfalls $b_t \in Z$ gilt. Die Entwicklung einer Klassifikation bedingt nunmehr die adäquate Ermittlung der (Un-)Ähnlichkeit zwischen Sequenzen, das heißt, es gilt ein (Un-)Ähnlichkeitsmaß zu wählen, das geeignet ist, die sequenzielle Datenstruktur von \mathbf{a} und \mathbf{b} im

⁴ Die Länge der Zeitintervalle zwischen den Beobachtungszeitpunkten determiniert die Präzision der erhobenen Verläufe: Bei einer engen zeitlichen Abfolge der Beobachtungen können Zustandswechsel verhältnismäßig zeitnah erfasst werden. Eine solche Vorgehensweise minimiert die Wahrscheinlichkeit, dass ein Zustand mit einer geringen Verweildauer keine Berücksichtigung findet, da der Wechsel in und aus diesem Zustand aufgrund der gewählten Intervallbreite zwischen zwei Beobachtungszeitpunkten liegt. Daraus lässt sich ableiten, dass eine engmaschige Durchführung der Beobachtungen angestrebt werden sollte.

Paarvergleich entsprechend zu berücksichtigen. Diese Anforderung können sogenannte »naive« Distanzmaße (Brüderl und Scherer 2004, S. 333) auf Basis der verallgemeinerten Minkowski-Metrik (zum Beispiel die City-Block-Metrik, die (quadrierte) euklidische Distanz) lediglich in einem unzureichenden Ausmaß erfüllen. Diesbezüglich können die folgenden beiden Gründe angeführt werden:

1. Die genannten Distanzmaße setzen Sequenzen von gleicher Länge T voraus.
2. Die angeführten herkömmlichen Distanzmaße ignorieren den ganzheitlichen Charakter der Sequenzen.

Die herkömmlichen Distanzmaße setzen Sequenzen von gleicher Länge T voraus (Kruskal 1983), wie sich anhand einer Darstellung der analytischen Vorgehensweise zeigen lässt: Bei übereinstimmender Länge bestehen die Sequenzen **a** und **b** jeweils aus insgesamt T aufeinander folgenden Realisierungen der nominalen Variablen Z , was die Bildung von insgesamt $T \cdot M$ Dummies erfordert. Diese werden im Zuge der Distanzermittlung der Reihe nach abgearbeitet, das heißt, es erfolgt ein Abgleich von Dummy ($t = 1; m = 1$) bis Dummy ($t = T; m = M$) zwischen den beiden Sequenzen. Bei einer Übereinstimmung der Elemente an der Stelle t beträgt die Distanz in diesem Segment der Sequenzen 0, da alle m Dummies die gleichen Werte aufweisen. Unterscheiden sich die Elemente allerdings, ergibt sich für die City-Block-Metrik und die quadrierte euklidische Distanz⁵ jeweils ein Wert von 2, da die Werte von zwei Dummies an der Stelle t divergieren. Demzufolge entsprechen die beiden angeführten Distanzmaße der doppelten Anzahl der identifizierten Nicht-Übereinstimmungen zwischen den Sequenzen.⁶

Liegen im Gegensatz dazu nun ungleich lange Sequenzen vor, existiert für die Sequenzen eine differierende Anzahl von Dummies. Besitzt Sequenz **a** etwa die Länge t und Sequenz **b** die Länge $t - q$ (es gilt $q \neq 0$), so besteht eine Differenz von $|q| \cdot M$ Dummies, welche die Berechnung dieser Distanzmaße bei listenweisem Fallausschluss gänzlich unmöglich macht. Die Anwendung des paarweisen Ausschlussverfahrens führt zu einer Nicht-Berücksichtigung jener $|q| \cdot M$ Dummies, die ausschließlich in der längeren Sequenz (**a** wenn $q > 0$; **b** wenn $q < 0$) vorkommen, das heißt, diese Option resultiert in einer methodisch bedingten Rechtszensierung der längeren Sequenz und somit in der Anpassung ihrer Länge an jene der kürzeren Sequenz. In der Folge gehen lediglich jene Dummies in die Berechnung des Distanzmaßes ein, die in beiden Sequenzen vorliegen.

Die Möglichkeit der Analyse von ungleich langen Sequenzen wirft allerdings vorab die Frage auf, ob bzw. wie die Differenz der Elemente zwischen den Sequenzen in der Berechnung des Unähnlichkeitsmaßes ihre Berücksichtigung finden soll. Handelt es sich um so

⁵ Zur Identität der City-Block-Metrik und der quadrierten euklidischen Distanz bei der Analyse von nominalen Variablen siehe Abschnitt 8.2.

⁶ Durch eine Gewichtung mit 0,5 kann die »Doppelzählung« vermieden werden (siehe Abschnitt 7.5).

genannte »ganzheitliche« (»entire«) Sequenzen (beispielsweise DNA, abgeschlossene Erwerbsbiographien, Schritte von historischen Tanzzeremonien), so kann deren Länge als ein Merkmal ihres inhärenten Charakters verstanden werden. Aus diesem Grund muss im vorliegenden Fall die Längendifferenz der Sequenzen in das Distanzmaß eingehen, da diese einen Teil der Unähnlichkeit der Sequenzen abbildet. Demgegenüber treten im sozialwissenschaftlichen Kontext jedoch nahezu ausschließlich »empirische« Sequenzen auf, deren Länge sich durch die Beobachtungsdauer (die Anzahl der realisierten Beobachtungszeitpunkte) bestimmt und infolge von temporärer Nichtbeobachtbarkeit, Panelmortalität oder einem zeitlich versetzten Beginn der Beobachtung der Sequenzen Variationen, unterworfen ist. Längenunterschiede reflektieren folglich nicht eine unterschiedliche »Natur« der Sequenzen, sondern sind auf erhebungstechnische Gegebenheiten zurückzuführen. Eine Verrechnung der Längendifferenz im Distanzmaß hat demzufolge zur Konsequenz, dass dieser Teil der Gesamtdistanz fast ausnahmslos inadäquat ermittelt wird oder auf methodischen Artefakten beruht, da kein Wissen über den weiteren Verlauf von unvollständig erhobenen Sequenzen vorhanden ist.⁷ Brüderl und Scherer (2004) schlagen zur Vermeidung dieser Problematik eine »künstliche« Zensierung vor, das heißt, alle Sequenzen werden auf die Länge der kürzesten Sequenz gebracht. Diese Vorgehensweise kann jedoch – vor allem bei einer großen Längendifferenz der Sequenzen – mit einem erheblichen Informationsverlust verbunden sein. Eine alternative Option stellt die Entfernung aller Sequenzen mit weniger als T Elementen dar, was allerdings ebenfalls in einem Informationsverlust resultiert und zudem bei einem systematischen Auftreten der Längenunterschiede zu einer verzerrten Lösung führt. Gemäß den Ausführungen verliert dieses erste Argument gegen die Anwendung von gebräuchlichen Distanzmaßen für sozialwissenschaftliche Sequenzen, deren Länge durch die Beobachtungsdauer determiniert wird, erheblich an Gewicht. Demgegenüber wiegt das anschließende zweite Argument allerdings umso schwerer.

Die angeführten herkömmlichen Distanzmaße vergleichen – wie bereits erläutert – die Sequenzen elementweise (Kruskal 1983; Brüderl und Scherer 2004) und ignorieren infolgedessen den ganzheitlichen Charakter der Sequenzen: In der Konsequenz können identische, allerdings verschobene (an unterschiedlichen Stellen auftretende) Subsequenzen in **a** und **b** nicht identifiziert und entsprechend berücksichtigt werden, das heißt, teilsequentielle Übereinstimmungen wirken sich – je nach der Heterogenität ihres Musters und der Distanz zwischen den Stellen ihres Auftretens in den Sequenzen – nicht in hinreichend reduzierender Weise auf die Distanzmaße aus. Im Extremfall kann dies bei annähernd

⁷ Die Analyse von ungleich langen Sequenzen bedingt überdies eine Standardisierung der ermittelten Distanzen. In diesem Zusammenhang schlagen Abbott und Forrest (1986, S. 482) vor, die im Zuge von Optimal Matching (om) ermittelte Levenshtein-Distanz durch die Länge der längeren Sequenz im Paarvergleich zu dividieren. Alternativ kann die Länge der längsten Sequenz im vorliegenden Datensatz zur Standardisierung herangezogen werden (vgl. Brzinsky-Fay, Kohler u. a. 2006).

identischen Sequenzen sogar zu beinahe maximalen Werten der Distanzmaße führen, wie an späterer Stelle noch exemplarisch verdeutlicht werden soll.

Die bisherigen Ausführungen offenbaren die limitierte Anwendbarkeit von konventionellen clusteranalytischen Distanzmaßen auf Basis der verallgemeinerten Minkowski-Metrik zur Ermittlung der Unähnlichkeit von Sequenzen. Aus diesem Grund erscheint eine Orientierung an Methoden der Sequenzdatenanalyse, die in der Molekularbiologie⁸, der Informatik und der Linguistik auf eine bereits jahrzehntelange Entwicklungs- und Anwendungstradition verweisen können, vielversprechend. Diesbezüglich kommt dem von Abbott und Forrest (1986) in den 1980er Jahren in die Sozialwissenschaften implementierten Verfahren des Optimal Matchings (OM)⁹ eine besondere Aufmerksamkeit zu. Aus diesem Grund konzentrieren sich die weiteren Ausführungen auf die Darstellung von OM. Zunächst soll jedoch in aller Kürze die sogenannte *Hamming-Distanz* (Hamming 1950) erläutert werden, da sie gewissermaßen die Basis der im Zuge von OM berechneten *Levenshtein-Distanz* (Levenshtein 1966) repräsentiert.

19.3 Die Hamming-Distanz

Die *Hamming-Distanz* basiert auf dem Abgleich der Elemente zweier Sequenzen a und b an jeder Stelle $t = 1, \dots, T$ und summiert die Anzahl der Nicht-Übereinstimmungen auf, die jeweils mit einem Gewicht von 1 (Kosten) in die Distanz eingehen. Liegt also $a_t \neq b_t$ vor, so werden Kosten von 1 vergeben. Demgegenüber betragen die Kosten 0, wenn eine Übereinstimmung der beiden Elemente vorliegt, das heißt $a_t = b_t$ der Fall ist. Bei einer Identität der beiden Sequenzen ($a = b$) fallen folglich keine Kosten an und die Hamming-Distanz beträgt 0. Im Gegensatz dazu weist die Hamming-Distanz bei maximaler Unähnlichkeit (an jeder Stelle t liegt eine Nicht-Übereinstimmung zwischen den beiden Sequenzen vor) einen Wert von T auf.

⁸ Ein Abriss über die Verwendung von sequenzdatenanalytischen Verfahren in den Naturwissenschaften (Biologie) und den Sozialwissenschaften sowie eine kurze Abhandlung über die Anwendungsunterschiede in den Wissenschaftsdisziplinen ist bei Lesnard (2006) zu finden.

⁹ Optimal Matching stellt lediglich eines aus einer Vielzahl von sequenzdatenanalytischen Verfahren dar. Einen klassischen Überblick zu diesem Thema stellt das Standardwerk von Sankoff und Kruskal (1983) zur Verfügung. Des Weiteren diskutiert Gusfield (1997) alternative Algorithmen zur Ermittlung der Ähnlichkeit von Sequenzen. Zudem können deskriptive Verfahren, wie etwa die Ermittlung der Auftretenshäufigkeit einzelner Sequenzen, ebenfalls unter dem Begriff der Sequenzdatenanalyse subsumiert werden. Im sozialwissenschaftlichen Kontext wurden – nicht zuletzt aufgrund von substanzialer Kritik an OM (Levine 2000; L. L. Wu 2000) hinsichtlich der Anwendbarkeit auf Lebensverläufe – von Dijkstra und Taris (1995) sowie Elzinga (2003) alternative sequenzdatenanalytische Algorithmen entwickelt, die aus Platzgründen nicht näher ausgeführt werden können.

Dieser Prozess der Distanzermittlung kann, einer alternativen – in stärkerem Ausmaß sequenzdatenanalytischen – Logik folgend, auch folgendermaßen konzipiert werden: Versuche die Sequenz \mathbf{a} (Quellsequenz) in die Sequenz \mathbf{b} (Zielsequenz) zu überführen und substituiere zu diesem Zweck an jeder Stelle t das Element a_t durch das Element b_t , falls $a_t \neq b_t$ gegeben ist. Verrechne des Weiteren für jede Substitution Kosten von 1 ($c_s = 1$) und bilde die Summe aller anfallenden Substitutionskosten. Die Hamming-Distanz kann demnach als die Summe der für eine Übereinstimmung der beiden Sequenzen benötigten Substitutionen von Elementen der Sequenz \mathbf{a} durch Elemente der Sequenz \mathbf{b} definiert und wie folgt angeschrieben werden:

$$\text{HAMMING}(\mathbf{a}, \mathbf{b}) = \sum_{a_t \neq b_t} 1, t = 1, \dots, T.$$

Unter Verwendung der Substitutionskostenfunktion

$$c_{st} = \begin{cases} 0, & a_t = b_t \\ 1, & a_t \neq b_t \end{cases}$$

ist die Hamming-Distanz gleich

$$\text{HAMMING}(\mathbf{a}, \mathbf{b}) = \sum_{t=1}^T c_{st}.$$

Nachdem die Hamming-Distanz offensichtlich denselben – bereits genannten – Limitierungen unterliegt wie die gängigen clusteranalytischen Distanzmaße, kann diese ebenfalls als »naive« Distanz bezeichnet werden. Zudem stellen die City-Block-Metrik und die quadrierte euklidische Distanz (identische) Linearkombinationen der Hamming-Distanz dar:

$$\text{HAMMING}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \cdot \text{CITY}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \cdot \text{QEUKLID}(\mathbf{a}, \mathbf{b}).$$

Der Unterschied zwischen diesen Distanzmaßen kann allerdings in der Logik ihrer Ermittlung ausgemacht werden: Während die City-Block-Metrik sowie die quadrierte euklidische Distanz – aufgrund ihres primären Anwendungsbereichs auf metrisch skalierte Variablen – eine Orientierung an den Differenzen zwischen den Elementen der Sequenzen aufweisen, bildet das Prinzip der Überführung der Sequenz \mathbf{a} in die Sequenz \mathbf{b} unter Anwendung der Transformationsoperation »Substitution« aus der Hamming-Distanz das Fundament für die Berechnung der Levenshtein-Distanz im Zuge von OM¹⁰.

¹⁰ Als Überblicksartikel bzw. Einführungsliteratur zu OM in den Sozialwissenschaften können exemplarisch Abbott (1995); Abbott und Tsay (2000); Aisenbrey (2000); Brüderl und Scherer (2004) sowie MacIndoe und Abbott (2009) genannt werden.

19.4 Die Levenshtein-Distanz

Die *Levenshtein-Distanz* (auch als Edit-Distanz bezeichnet) kann als eine Erweiterung der Hamming-Distanz erachtet werden, da neben der Substitution noch zwei weitere Operationen, nämlich »Einfügen« (füge b_t ein) und »Löschen« (lösche a_t), im Zuge der Sequenztransformation von \mathbf{a} zu \mathbf{b} zur Verfügung stehen. Diese werden als Indeloperationen bezeichnet und ermöglichen die Analyse von ungleich langen Sequenzen sowie eine Identifikation und adäquate Berücksichtigung von in den Sequenzen verschobenen identischen Subsequenzen, das heißt, ihr Einsatz führt zur Überwindung der analytischen Schwächen der »naiven« Distanzmaße. Allerdings ist es nicht mehr möglich, die Levenshtein-Distanz als die Summe der für ein Alignment benötigten Transformationen festzusetzen, da eine Substitution (ersetze a_t durch b_t) gleichsam durch die Ausführung von zwei Indeloperationen realisiert werden kann (lösche a_t und füge b_t ein). In der Konsequenz müssen für die Transformationsoperationen unterschiedliche Gewichte spezifiziert werden, welche als Transformationskosten bezeichnet werden. Um den metrischen Axiomen zu genügen (Kruskal 1983), muss für die Levenshtein-Distanz des Weiteren gelten, dass die Transformationskosten für Einfügen (c_e) jenen für Löschen (c_l) entsprechen, da ansonsten eine Verletzung der symmetrischen Eigenschaft $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$ vorliegen kann. Aus diesem Grund wird in weiterer Folge ausschließlich von Indelkosten (c_i) gesprochen, für die somit $c_i = c_e = c_l$ gilt. Bei variierende Substitutionskosten (eine Ausführung erfolgt an späterer Stelle), muss für eine Erfüllung der symmetrischen Eigenschaft der Levenshtein-Distanz ebenfalls $c_s(z_m, z_n) = c_s(z_n, z_m)$ gegeben sein.

Ziel von OM ist es nun, jene Kombination (oftmals sind auch mehrere Kombinationen möglich) der erlaubten Operationen zu ermitteln, für welche die für ein Alignment insgesamt benötigten Transformationskosten minimal sind, das heißt, die Levenshtein-Distanz kann als die Summe der für eine Angleichung der Sequenz \mathbf{a} an die Sequenz \mathbf{b} (»Alignment«) benötigten minimalen Transformationskosten verstanden werden. Die Ähnlichkeit der beiden Sequenzen hängt somit von der Art und Anzahl der für eine Angleichung benötigten Transformationen ab. Die Levenshtein-Distanz wird rekursiv berechnet:

$$\text{LEVENSHTEIN}(\mathbf{a}_t, \mathbf{b}_{t^*}) = d(\mathbf{a}_t, \mathbf{b}_{t^*})$$

$$= \min \begin{cases} d(\mathbf{a}_{t-1}, \mathbf{b}_{t^*}) + c_i(a_t, \varphi) & \rightarrow \text{Löschen von } a_t \\ d(\mathbf{a}_{t-1}, \mathbf{b}_{t^*-1}) + c_s(a_t, b_{t^*}) & \rightarrow \text{Ersetzen von } a_t \text{ durch } b_{t^*} \\ d(\mathbf{a}_t, \mathbf{b}_{t^*-1}) + c_i(\varphi, b_{t^*}) & \rightarrow \text{Einfügen von } b_{t^*} \end{cases}$$

mit φ als Platzhalter.

Die angeführte Rekursionsgleichung der Levenshtein-Distanz (jedes Folgenglied ist eine Funktion der vorausgegangenen Folgenglieder) offenbart, dass die Distanz sich nicht in einem einzigen Schritt errechnen lässt, sondern ein iterativer Prozess abgearbeitet werden muss, der über den Ansatz der dynamischen Programmierung (Day und McMorris 1994) gelöst wird. Die Bedeutung der vorliegenden Rekursionsgleichung kann in Anlehnung an Kruskal (1983, S. 26) folgendermaßen formuliert werden: Das mit den minimalen Kosten verbundene Alignment zwischen der Sequenz a bis zur Stelle t und der Sequenz b bis zur Stelle t^* basiert auf der kostengünstigsten Wahl aus den drei Alternativen:

1. Verwende das mit den geringsten Kosten verbundene Alignment zwischen a_{t-1} und b_{t^*} und lösche a_t .
2. Verwende das mit den geringsten Kosten verbundene Alignment zwischen a_{t-1} und b_{t^*-1} und substituiere a_t durch b_{t^*} .
3. Verwende das mit den geringsten Kosten verbundene Alignment zwischen a_t und b_{t^*-1} und füge b_{t^*} ein.

Die Realisierung der Levenshtein-Distanzermittlung erfolgt unter Anwendung des *Needleman-Wunsch-Algorithmus* (Needleman und Wunsch 1970), dessen Ablauf und Funktionsweise etwa von Abbott und Hrycak (1990); Erzberger (2001); Erzberger und Prein (1997) sowie Kruskal (1983) explizit dargestellt werden.

19.5 Ein theoretisches Beispiel

Nachdem das Konzept der Levenshtein-Distanz in den Grundzügen dargelegt wurde, soll nun ein konstruiertes, elementares Beispiel¹¹ die Logik der Distanzermittlung, die Inadäquatheit von »naiven« Distanzmaßen anhand der Hamming-Distanz sowie die Notwendigkeit der Vergabe von unterschiedlichen Kosten für Substitutions- und Indeloperationen verdeutlichen. Zu diesem Zweck werden drei Paarvergleiche von Sequenzen angestellt, denen die folgenden Annahmen zugrunde liegen: Die Panelstichprobe einer Geburtskohorte wird ab dem 18. Lebensjahr halbjährlich bezüglich ihres Ausbildungs-/Erwerbsstatus befragt. Der Zustandsraum dieses Merkmals beinhaltet die folgenden sechs Elemente ($M = 6$): Schule (S), Präsenzdienst (P), tertiäre Ausbildung (T), Erwerbstätigkeit (E), Arbeitslosigkeit (A) und (Bildungs-)Karens (K). Zum Analysezeitpunkt wurden bereits sechs Panelwellen ($T = 6$) realisiert und es liegt ein vollständiger Datensatz (ohne fehlende Werte) vor. Die Verläufe von sechs interessierenden Individuen werden in Tabelle 19.1 in sequentieller Form dargestellt. Angeführt werden neben der Hamming-Distanz

¹¹ Das vorliegende Beispiel ist an jenes von Brüderl und Scherer (2004, S. 333) angelehnt.

Tab. 19.1: Vergleich der Distanzmaße bzw. der Auswirkung einer unterschiedlichen Gewichtung der Transformationsoperationen

	Vergleich 1	Vergleich 2	Vergleich 3
Sequenz 1	TTTTTT	EEETTT	SPAEK
Sequenz 2	EEEEEE	TTTEEE	PAEKTE
Hamming-Distanz	6	6	6
Levenshtein-Distanz 1 ($c_s = 1; c_i = 1$)	6	6	2
Levenshtein-Distanz 2 ($c_s = 1; c_i = 0,5$)	6	3	1

zwei Levenshtein-Distanzmaße (LD_1 und LD_2). Bei der LD_1 entsprechen die Substitutionskosten den Indelkosten, das heißt, alle Transformationsoperationen werden gleich gewichtet. Bei LD_2 hingegen werden die Substitutionskosten gleich 1 gesetzt, die Indelkosten gleich 0,5. Damit wird der Tatsache Rechnung getragen, dass sich eine Substitution aus zwei Indeloperationen zusammensetzt.

Im Vergleich 1 weisen beide Sequenzen lediglich einen – allerdings divergierenden – Zustand auf, das heißt, die beiden Sequenzen sind – unter den gegebenen Annahmen – einander maximal unähnlich. Folglich wird für die Hamming-Distanz ein Wert von 6 ausgewiesen, da zu allen 6 Zeitpunkten eine Nichtübereinstimmung der Elemente der Sequenzen 1 und 2 vorliegt. Im Zuge der Berechnung der LD_1 werden 6 Substitutionen (ersetze 6-mal T durch E) durchgeführt, die ebenfalls zu aufsummierten Gesamtkosten von 6 führen. Zur Realisierung eines Alignments auf Basis der LD_2 erweist es sich als arbiträr, ob 6 Substitutionsoperationen oder 12 Indeloperationen (lösche 6-mal T; füge 6-mal E ein) zur Anwendung kommen, beide Strategien führen zu einer Distanz von 6, da die Indelkosten auf 0,5 festgelegt wurden. Im vorliegenden Fall von einander maximal unähnlichen Sequenzen – durchgehende Absolvierung einer tertiären Ausbildung bzw. ununterbrochene Erwerbstätigkeit – führen somit alle drei Distanzmaße auch konsistent zu den maximalen Distanzen, das heißt, für die Hamming-Distanz sowie für die LD_1 und LD_2 sind bei $T = 6$ keine höheren Kosten als 6 möglich.

Im Paarvergleich 2 sind die Elemente in den beiden Sequenzen ebenfalls an jeder Stelle ungleich. Beide Sequenzen bestehen allerdings vollständig aus 2 identischen Subsequenzen gleicher Länge (EEE und TTT), welche in umgekehrter Reihenfolge angeordnet sind. Tertiäre Ausbildung und Erwerbstätigkeit werden von beiden Individuen im gleichen zeitlichen Ausmaß, jedoch in unterschiedlichen, einander nicht überlappenden, Zeitintervallen wahrgenommen. Die Hamming-Distanz nimmt wiederum den maximalen Wert von 6 ein, da abermals an keiner Stelle eine Übereinstimmung der Elemente der beiden Sequenzen besteht. Die undifferenzierte Transformationskostenfestsetzung ($c_s = c_i = c = 1$) der LD_1 führt ebenfalls zu einem Wert von 6, da entweder 6 Substitutionen (ersetze

3-mal E durch T; ersetze 3-mal T durch E) oder 6 Indeloperationen (lösche 3-mal E am Beginn der Sequenz; füge 3-mal E am Ende der Sequenz ein) für eine Überführung der Sequenz 1 in die Sequenz 2 notwendig sind. Die Hamming-Distanz sowie die LD_1 sind folglich nicht in der Lage, die identischen Subsequenzen zu berücksichtigen und führen analog zu Vergleich 1 zu den Maximalkosten. Demgegenüber ist bei der LD_2 die kostenminimale Alignmentstrategie (lösche 3-mal E am Beginn der Sequenz; füge 3-mal E am Ende der Sequenz ein) mit Gesamtkosten von 3 verbunden, da die Indelkosten mit 0,5 festgelegt wurden. Die gewählte Transformationskostenstruktur ermöglicht somit den effektiven Einsatz der Indeloperationen, mit dem eine entsprechende Beachtung der übereinstimmenden Subsequenzen EEE und TTT einhergeht, das heißt, die LD_2 bringt für die beiden Sequenzen des Vergleichs 2 eine größere Ähnlichkeit (in der Form einer geringeren Distanz) hervor als für jene des Vergleichs 1.

Die beiden Sequenzen des Vergleichs 3 besitzen die identische Subsequenz PAEKT, diese tritt allerdings zeitlich versetzt (an unterschiedlichen Stellen in den Sequenzen) auf. So besucht Individuum 1 zu Beginn der Panelerhebung noch eine maturaführende¹² Schule mit einer fünfjährigen Dauer auf der Ebene der Sekundarstufe II, während Individuum 2 das Gymnasium zum selben Zeitpunkt bereits beendet und den Präsenzdienst (P) angetreten hat. Beide Individuen durchleben nach dem Präsenzdienst eine Phase der Arbeitslosigkeit (A), gefolgt von Eintritt in das Erwerbsleben (E), der Aufnahme von Bildungskarenz (K) und dem Beginn einer tertiären Ausbildung (T). Während der Beobachtungszeitraum für Individuum 1 an dieser Stelle endet, kann für Individuum 2 zum letzten Erhebungszeitpunkt eine Rückkehr in die Erwerbstätigkeit (E) ausgemacht werden. Aufgrund der Zensierung der Daten kann nicht festgestellt werden, ob die zeitversetzte Übereinstimmung der beiden Verläufe ihren Fortgang findet, das heißt, ob Individuum 1 zum Zeitpunkt $T + 1$ ebenfalls die tertiäre Ausbildung abbricht und analog zu Individuum 2 beginnt, wiederum einer Erwerbstätigkeit nachzugehen. Die zwischen den vorliegenden Sequenzen ermittelte Hamming-Distanz beträgt nun ein weiteres Mal den Maximalwert von 6, das heißt, aufgrund der wiederholten Nichtübereinstimmung der Elemente zu allen Beobachtungszeitpunkten werden die beiden Sequenzen neuerlich als einander maximal unähnlich erachtet. Im Gegensatz dazu offenbaren sich bei den Levenshtein-Distanzen an dieser Stelle die Vorzüge der Indeloperationen in beachtlicher Weise: So werden für eine kostenoptimale Überführung der Sequenz 1 in die Sequenz 2 lediglich zwei Indeloperationen benötigt (lösche 1-mal S am Beginn der Sequenz; füge

¹² Die Matura repräsentiert das österreichische Äquivalent zum Abitur. Die berufsbildenden höheren Schulen (BHS) stellen etwa maturaführende Schulen mit einer fünfjährigen Sekundarstufe II dar. Im Gegensatz dazu weist das (Real-)Gymnasium lediglich eine vierjährige Dauer der Sekundarstufe II auf. Die unterschiedliche zeitliche Dauer der Schulen zur Realisierung des formal gleichen Bildungsabschlusses (Matura) kann somit zu Verschiebungen von späteren, an sich identischen Ausbildungs-/Erwerbsverläufen von Individuen führen.

1-mal E am Ende der Sequenz ein), resultierend in Gesamtkosten von 2 (LD_1) bzw. 1 (LD_2). Die durch den Besuch von im zeitlichen Ausmaß divergierenden maturaführenden Schulen hervorgerufene Diskrepanz der Startzeitpunkte der identischen Subsequenz der beiden Individuen kann im Zuge der Berechnung der LD_1 und LD_2 durch die Anwendung der Indeloperationen entsprechend verarbeitet werden, so dass die Ähnlichkeit der beiden Sequenzen – basierend auf ihrer identischen Subsequenz – hinreichend zum Ausdruck kommt.

In der Konsequenz legt das angeführte Beispiel die Überlegenheit der Levenshtein-Distanz gegenüber der Hamming-Distanz und den anderen genannten »naiven« Distanzmaßen für die Ermittlung der Ähnlichkeit zwischen Sequenzen dar. Zudem konnte die Relevanz der differenzierten Vergabe von Transformationskosten demonstriert werden, da ausschließlich die LD_2 mit ihrer spezifischen Kostenstruktur in der Lage war, die zunehmende Ähnlichkeit der Sequenzen von Paarvergleich 1 zu Paarvergleich 2 entsprechend durch eine Verringerung der Distanzwerte abzubilden. Aus diesem Grund soll nun der Festsetzung der Transformationskosten – eine der zentralen Problematiken der Methode des OM – die ungeteilte Aufmerksamkeit gewidmet werden.

19.6 Die Festsetzung der Transformationskosten

Bislang wurde nur auf die Notwendigkeit der Äquivalenz der Indelkosten eingegangen. Die Substitutionskosten wurden konstant gehalten, das heißt, es erfolgte die Vorgabe, dass für alle möglichen Substitutionen zwischen nicht-identischen Zuständen aus Z mit den gleichen Kosten verbunden sind. Diese Annahme soll nun aufgegeben werden, da sie dahingehend als restriktiv erachtet werden kann, dass manche Zustände aus Z einander aus inhaltlich-theoretischer Sicht ähnlicher sind als andere Zustände und diesem Umstand bei der Distanzermittlung Rechnung getragen werden sollte. Werden etwa die Zustände »Teilzeiterwerb« und »Arbeitslosigkeit« dem »Vollzeiterwerb« gegenüber gestellt, so wird mit ziemlicher Sicherheit ein Konsens darüber herrschen, dass »Vollzeiterwerb« und »Teilzeiterwerb« eine höhere konzeptlogische Ähnlichkeit aufweisen als »Vollzeiterwerb« und »Arbeitslosigkeit«. In der Konsequenz sollte im ersten Fall eine Substitution der beiden Zustände mit geringeren Kosten versehen werden als im zweiten Fall. Variierende Substitutionskosten bedingen zur Erfüllung der symmetrischen Eigenschaft der Levenshtein-Distanz allerdings $c_s(z_m, z_{m^*}) = c_s(z_{m^*}, z_m)$, das heißt, eine Substitution von »Vollzeiterwerb« durch »Teilzeiterwerb« muss mit denselben Kosten verbunden sein, wie die Ersetzung von »Teilzeiterwerb« durch »Vollzeiterwerb«. Bei einer diesbezüglich differenzierten Kostenvergabe würde ansonsten die Angleichung der Sequenz **a** an die Sequenz **b** zu anderen Gesamtkosten führen, als eine Überführung von **b** in **a** (falls

im Zuge des Alignmentprozesses überhaupt Substitutionsoperationen zur Anwendung gelangen).

In seiner substantiellen Kritik an OM beanstandet L. L. Wu (2000) unter anderem die Äquivalenz der mit den Substitutionen von z_m durch z_{m^*} und von z_{m^*} durch z_m verbundenen Kosten. Im Kern wird bemängelt, dass etwa ein Übergang von der »Arbeitslosigkeit« in die »Erwerbstätigkeit« einen (erheblichen) individuellen Aufwand erfordert (Jobsuche), während die Transition von der »Erwerbstätigkeit« in die »Arbeitslosigkeit« entweder fremdbestimmt ist (Entlassung bzw. Kündigung durch den Arbeitgeber bzw. die Arbeitgeberin) oder lediglich einen Formalakt darstellt (Kündigung durch den Arbeitnehmer bzw. die Arbeitnehmerin) und somit in aller Regel mit einem ungleich geringeren Aufwand realisiert werden kann. Folglich erscheint die Gleichheit der Substitutionskosten aus theoretischer Sicht nicht haltbar. In diesem Zusammenhang verweist Halpin (2010) jedoch auf die irrtümliche Anwendung des Übergangskonzepts im Zuge der Substitutionsoperation: Nicht der Aufwand der Transitionen zwischen zwei Zuständen soll in den Substitutionskosten erfasst werden, sondern das Ausmaß der Ähnlichkeit zwischen den beiden Zuständen. Dies bedeutet in weiterer Folge, dass der Äquivalenz der Substitutionskosten wieder ihre (theoretisch) Legitimation zugesprochen werden kann.

Gauthier, Widmer u. a. (2009) stellen eine Typologie von bislang in empirischen OM-Anwendungen realisierten Strategien zur Festsetzung der Substitutionskosten zur Verfügung. Neben der trivialen Lösung der Gleichsetzung aller möglichen Substitutionskosten auf \bar{c}_s identifizieren sie vier weitere grundlegende Taktiken zur Ermittlung von differenzierten Substitutionskosten¹³, welche das gemeinsame Ziel verfolgen, die Ähnlichkeiten zwischen den Ausprägungen einer qualitativen Zustandsvariablen zu quantifizieren. Die Substitutionskosten werden in diesem Fall durch eine symmetrische ($M \times M$)-Substitutionskostenmatrix S repräsentiert. Da die Substitution eines Zustandes z_m durch sich selbst klarerweise mit keinen Kosten verbunden ($c_s(z_m, z_m) = 0$) ist bzw. die Äquivalenz von zwei Zuständen keine Substitution erfordert, ist die Hauptdiagonale von S ausschließlich mit Nullen besetzt:

$$S = \begin{pmatrix} 0 & c_s(z_1, z_2) & \dots & c_s(z_1, z_M) \\ c_s(z_2, z_1) & 0 & \dots & c_s(z_2, z_M) \\ \vdots & \vdots & \ddots & \vdots \\ c_s(z_M, z_1) & c_s(z_M, z_2) & \dots & 0 \end{pmatrix}.$$

¹³ Aus Platzgründen muss auf eine Erläuterung der bei Gauthier, Widmer u. a. (2009) angeführten Herangehensweisen zur Ermittlung der Substitutionskostenmatrix S verzichtet werden. Sie sollen an dieser Stelle lediglich erwähnt werden: (1) Kostenfestsetzung auf Grundlage von theoretischen Annahmen, (2) Anwendung von empirisch begründbaren Werten, (3) Verwendung der empirischen Übergangswahrscheinlichkeiten zwischen den Zuständen als Proxy für die Ähnlichkeit der Zustände und (4) Kombinationen der soeben genannten Strategien.

Um eine optimale Ausschöpfung aller Transformationsoperationen im Zuge der Ermittlung der Levenshtein-Distanz zu gewährleisten, muss überdies eine geeignete Relation zwischen Substitutions- und Indelkosten gegeben sein. Ausgehend von der restriktierten Substitutionskostenstruktur $c_s(z_m, z_{m^*}) = \bar{c}_s$ muss für eine Durchführung von Substitutionsoperationen zunächst $\bar{c}_s < 2c_i$ gelten, da in allen anderen Fällen eine Substitution kostengünstiger ($\bar{c}_s > 2c_i$) bzw. zu gleichen Kosten ($\bar{c}_s = 2c_i$) über die beiden Indeloperationen (Löschen und Einfügen) realisiert werden könnte und die Transformationsoperation der Substitution somit obsolet wird. Für differenzierte Substitutionskosten muss in der Konsequenz $c_s^{\max} < 2c_i$ vorliegen, das heißt, die maximalen Substitutionskosten für die beiden einander unähnlichsten Zustände aus Z müssen – zumindest marginal – unter den doppelten Indelkosten liegen. Zudem sollte die Intervalluntergrenze für \bar{c}_s über c_i liegen ($c_i < \bar{c}_s$), um den Nutzen der Einführung der Indeloperationen nicht erheblich abzuschwächen (vgl. Brüderl und Scherer 2004; siehe auch die Ausführungen von Abbott und Tsay 2000, S. 12, 13). Analog gilt für variable Substitutionskosten, dass die minimalen Substitutionskosten generell die Indelkosten übersteigen sollten ($c_i < c_s^{\min}$). Zusammenfassend lässt sich die Transformationskostenstruktur nun allgemein subsumieren:

$$c_i < c_s^{\min} \leq c_s^{\max} < 2c_i .$$

Werden die Indelkosten etwa auf 1 fixiert, so liegt der Wertebereich der Substitutionskosten im kontinuierlichen Intervall $]1, 2[$ (vgl. Brüderl und Scherer 2004), das heißt, die Werte der Elemente der Matrix S sollen (mit Ausnahme der Elemente der Hauptdiagonale) diesem Intervall entstammen. Liegt kein vernünftiger theoriebezogener Grund für eine Differenzierung der Substitutionskosten vor, werden die Substitutionskosten mehrheitlich so gewählt, dass sie den doppelten Indelkosten entsprechen. Diese Vorgehensweise erscheint einerseits gerechtfertigt, da – wie bereits an anderer Stelle angeführt – eine Substitution über die Anwendung von zwei Indeloperationen realisiert werden kann. Andererseits zieht diese Transformationskostenstruktur jedoch die Redundanz der Substitutionsoperation nach sich.

Prinzipiell ist es möglich, die Indelkosten ebenfalls variabel zu gestalten, das heißt, dass die Indeloperationen im Laufe des Alignmentprozesses mit unterschiedlichen Kosten versehen werden. So stellt Aisenbrey (2000, S. 25) fest, dass »eine Lücke in einer Sequenz zu Beginn der Beobachtung ‚teurer‘ oder ‚billiger‘ sein müsste als am Ende [...].«. Rohwer und Pötter (2005, S. 493) schlagen zur Erzeugung von differenzierten Indelkosten eine allgemeine lineare Funktion vor:

$$c_i(t) = \alpha + \beta(t - 1) , \alpha, \beta \geq 0$$

(die Notation wurde an jene im vorliegenden Kapitel angepasst). Gegen variable Indelkosten sprechen in erster Linie (1) der damit einhergehende Komplexitätszuwachs der

Transformationskostenstruktur (die Kombination von variablen Indel- und Substitutionskosten) sowie (2) eine mehrheitlich (fehlende) theoretische Legitimation (siehe etwa Abbott und Hrycak 1990, S. 155).

19.7 Der Analyseablauf

Nach den Ausführungen zur Transformationskostenstruktur sollen in weiterer Folge die einzelnen Phasen der Analyse von sozialwissenschaftlichen Sequenzdaten auf Basis von OM dargestellt werden:

1. Aufbereitung der Daten
2. Festlegung der Analyseeinstellungen und Durchführungsoptionen
3. Durchführung der OM-Analyse
4. Clusterung
5. Formale Validierung
6. Ergebnisdarstellung als »Sequenz-Indexplot«
7. Kausalanalyse (optional)

Aufbereitung der Daten: Als Datenbasis für OM fungiert eine $(n \times T)$ -Matrix Z . Jede Zeile dieser Matrix repräsentiert die Sequenz eines Individuums i aus n in Form des Zeilenvektors z_i , das heißt, z_i bildet – bei vollständigen Sequenzen – alle Realisierungen der Zustandsvariable Z für das Individuum i von $t = 1$ bis T ab. Ferner muss Z um den Spaltenvektor i erweitert werden, der die Identifikationsnummer für alle n Individuen enthält.

Festlegung der Analyseeinstellungen und Durchführungsoptionen für die OM-Analyse: In dieser Phase wird die Transformationskostenstruktur sowie die Art der Standardisierung bei Sequenzen mit divergierenden Längen bestimmt. Weiterhin muss festgesetzt werden, ob die n Sequenzen aus Z entweder mit einer Referenzsequenz r abgeglichen werden, welche in der Regel auf theoretischen Überlegungen basiert bzw. einen für die jeweilige Fragestellung idealtypischen Verlauf repräsentiert oder ein Matching einer jeden Sequenz mit allen anderen $n - 1$ Sequenzen realisiert werden soll. Die erste Option resultiert nach Ausführung einer OM-Analyse in einem Levenshtein-Distanzvektor d , der die insgesamt n Distanzen zwischen den Sequenzen der Individuen und der Referenzsequenz beinhaltet:

$$d = \begin{pmatrix} d(z_1, r) \\ \vdots \\ d(z_i, r) \\ \vdots \\ d(z_n, r) \end{pmatrix}.$$

Die zweite Option führt zu einer $(n \times n)$ -Levenshtein-Distanzmatrix \mathbf{D} , die (aufgrund der metrischen Eigenschaften der Levenshtein-Distanz) durch eine quadratische sowie symmetrische Form gekennzeichnet ist und deren Elemente in der Hauptdiagonale alle den Wert 0 aufweisen, da die Angleichung der Sequenz eines Individuums an sich selbst keine Kosten verursacht. Für jedes Individuum i werden $n - 1$ Levenshtein-Distanzen ermittelt, was aufgrund der symmetrischen Beschaffenheit von \mathbf{D} zu einer Gesamtanzahl von $n \cdot (n - 1)/2$ unterschiedlichen Distanzen führt. Folglich enthält die untere Dreiecksmatrix von \mathbf{D} alle relevanten Informationen, die zur Weiterbearbeitung in den folgenden Analysephasen benötigt werden. Diese Option wird gewählt, wenn einerseits keine geeignete Referenzsequenz zur Verfügung steht oder andererseits die Erzeugung einer Klassifikation der vorliegenden Sequenzen das Ziel der Analysen ist. Es gilt:

$$\mathbf{D} = \begin{pmatrix} 0 & d(z_1, z_2) & \dots & d(z_1, z_i) & \dots & d(z_1, z_n) \\ d(z_2, z_1) & 0 & \dots & d(z_2, z_i) & \dots & d(z_2, z_n) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d(z_i, z_1) & d(z_i, z_2) & \dots & 0 & \dots & d(z_i, z_n) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d(z_n, z_1) & d(z_n, z_2) & \dots & d(z_n, z_i) & \dots & 0 \end{pmatrix}.$$

Durchführung der om-Analyse: Im Anschluss an die Determination der soeben angeführten Optionen kann die eigentliche om-Analyse realisiert, das heißt, die Levenshtein-Distanzmatrix \mathbf{D} erzeugt werden. Mittlerweile steht die Anwendung von om in diversen Standardstatistikpaketen, wie etwa in R (TraMineR-Paket; siehe Gabadinho, Ritschard u. a. 2009; Mills 2010) und Stata (sq-Ados; siehe Brzinsky-Fay, Kohler u. a. 2006), zur Verfügung. Darüber hinaus bieten die Freeware-Programme TDA¹⁴ (»Transition Data Analysis«; siehe Rohwer und Pötter 2005) und »Optimize«¹⁵ (entwickelt von Abbott) ebenfalls die Möglichkeit zur Durchführung von om an. Ferner sei auf die Übersichten von Abbott (1995), Abbott und Tsay (2000) sowie MacIndoe und Abbott (2009) verwiesen, die unter anderem eine Reihe von Softwarepaketen aus der Molekularbiologie beinhalten.

Clusterung: Je nach vorliegender Fragestellung können in diesem Analyseschritt die unterschiedlichen in diesem Buch beschriebenen clusteranalytischen Verfahren auf die erstellte Levenshtein-Distanzmatrix \mathbf{D} angewendet werden. Liegt das Interesse in der *räumlichen Darstellung* der Sequenzen entsprechend ihrer (Un-)Ähnlichkeit zueinander, so fällt die Wahl auf die (nichtmetrische) multidimensionale Skalierung (siehe Kapitel 4), deren Voraussetzung – die Existenz einer (Un-)Ähnlichkeitsmatrix (mit ordinaler Information) – von \mathbf{D} erfüllt wird. Zur *Erzeugung einer Klassifikation oder zur hierarchischen*

¹⁴ Verfügbar unter: <http://www.stat.ruhr-uni-bochum.de/tda.html> (Stand: 25.03.2010).

¹⁵ Verfügbar unter: <http://www.srcc.uchicago.edu/users/abbot> (Stand: 25.03.2010).

Darstellung als Baum der vorliegenden n Sequenzen wird in der Regel die Durchführung eines hierarchisch-agglomerativen clusteranalytischen Verfahrens gewählt, wobei eine Durchsicht bestehender (empirischer) Arbeiten eine heterogene Anwendungspraxis der Fusionsalgorithmen offenbart. In der Tendenz ist allerdings eine Präferenz für das Ward-Verfahren (siehe Kapitel 11) zu erkennen (Aisenbrey und Fasang 2010; Martin, Schoon u. a. 2008). Brüderl und Scherer (2004) sowie Martin, Schoon u. a. (2008) attestieren diesem clusteranalytischen Ansatz auf Basis ihrer Anwendungserfahrungen die beste Performance. Aassve, Billari u. a. (2007) führen zudem an, dass die Eigenschaft des Ward-Verfahrens, bei fortschreitender Fusionierung die Unterschiede in den Besetzungszahlen der bereits bestehenden Clustern auszugleichen, ebenfalls dessen Anwendung rechtfertigt, da alternative hierarchisch-agglomerative Algorithmen in der Regel zu einigen sehr großen und vielen kleinen Residualclustern führen. Gegen eine Anwendung des Ward-Verfahrens spricht allerdings, dass es mit der Verwendung von quadrierten euklidischen Distanzen eine Bedingung voraussetzt (siehe Abschnitt 11.1), die von der Levenshtein-Distanzmatrix \mathbf{D} nicht erfüllt werden kann. Allgemein kann festgehalten werden, dass in Bezug auf die clusteranalytische Bearbeitung der Levenshtein-Distanzmatrix noch erheblicher Forschungsbedarf besteht und folglich an dieser Stelle keine klaren Empfehlungen angeboten werden können. In Anlehnung an Martin, Schoon u. a. (2008) wird vorgeschlagen, mehrere Fusionsalgorithmen anzuwenden und die beste Lösung anhand von statistischen Entscheidungskriterien (siehe Abschnitt 9.1.5) sowie der inhaltlichen Interpretierbarkeit der Cluster zu selektieren. Im Sinne der Stabilität einer Clusterlösung sollte die ausgewählte Lösung jedoch nicht in relevanter Weise von den Ergebnissen der restlichen durchgeföhrten clusteranalytischen Verfahren abweichen.

Formale Validierung: Zur formalen Validierung der gefundenen Clusterlösung im Kontext von OM schlagen Abbott und Hrycak (1990) explizit zwei Ansätze vor, nämlich den Vergleich der Distanzmittelwerte innerhalb und zwischen den Clustern sowie die Entwicklung von idealtypischen Sequenzen (Repräsentationssequenzen) für jedes Cluster. Im ersten Fall sollte die mittlere Distanz innerhalb der Cluster deutlich unter jener zwischen den Clustern liegen. Aisenbrey und Fasang (2010) empfehlen in diesem Zusammenhang als Faustregel eine Relation kleiner als 0,5, das heißt $\bar{d}_{\text{innerhalb}}(K) < 0,5 \cdot \bar{d}_{\text{zwischen}}(K)$ für eine Clusterlösung mit K Clustern. Hinsichtlich der zweiten Validierungsstrategie gilt, dass die Distanzen der Sequenzen zur idealtypischen Sequenz des jeweiligen Clusters immer geringer sein sollten als zu den Repräsentationssequenzen der anderen Cluster (siehe auch Aisenbrey 2000). Als idealtypische Sequenzen könnten etwa die modalen Sequenzen (mit der höchsten Auftretenshäufigkeit) eines jeder Clusters oder die Sequenzen mit der geringsten Gesamtdistanz zu allen anderen Sequenzen in den jeweiligen Clustern herangezogen werden.

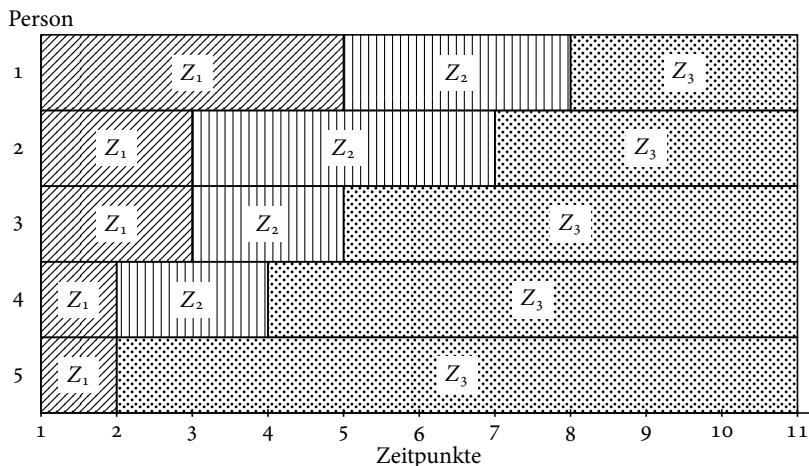


Abb. 19.1: Ein Sequenz-Indexplot

Ergebnisdarstellung: Die Ergebnisdarstellung erfolgt zumeist mit Hilfe eines sogenannten *Sequenz-Indexplots*. Jede Linie bzw. jeder Balken eines Sequenz-Indexplots repräsentiert die Sequenz eines Individuums, das heißt, die Abfolge der eingenommenen Zustände einer Variable über den Beobachtungszeitraum. Der Aufbau eines Plots ist schematisch in Abbildung 19.1 für ein Cluster dargestellt. Das Cluster besteht aus fünf Personen, untersucht wurden drei Zustände zu elf Messzeitpunkten, also zehn dazwischen liegende Perioden. Der Zustand Z₁ tritt bei Person 1 in den ersten vier Zeitperioden (zwischen den Messzeitpunkten 1 und 5) auf. In den nächsten drei Perioden folgt der Zustand Z₂ und schließlich Zustand Z₃. Bei Person 2 dauert der erste Zustand nur zwei Zeitperioden an, es folgt der zweite Zustand für vier Perioden und für die restlichen Perioden tritt der dritte Zustand auf. Die Zeitverläufe der restlichen Personen sind analog zu interpretieren. Bei einer Person müssen dabei nicht alle Zustände auftreten. In dem Beispiel fehlt bei Person 5 der Zustand Z₂.

Kausalanalyse (optional): Die im Zuge des bisherigen Analyseprozesses erzeugte Klassifikationsvariable kann – je nach Forschungsintention – abschließend in ein Kausalmodell integriert werden und als erklärende bzw. zu erklärende Variable dienen.

19.8 Fazit und Anwendungsempfehlungen

Optimal Matching stellt einen vielversprechenden Ansatz zur Klassifikation von sozialwissenschaftlichen Sequenzdaten dar. Es wurde ersichtlich, dass der gesamte Analyseprozess

von der Definition des Zustandsraumes über die Festlegung der Transformationskostenstruktur bis hin zur Auswahl des adäquaten clusteranalytischen Verfahrens in erheblichem Maße auf den zu treffenden Annahmen beruht. Allerdings wird gegenwärtig – etwa durch die Entwicklung von komplexen Verfahren zur datenbasierten Ermittlung der Transformationskosten – versucht, die Entscheidungslast der Anwender zu reduzieren und somit die Anwendbarkeit zu erleichtern. Dies erscheint von erheblicher Relevanz, denn bei adäquater Anwendung – so versichern zumindest MacIndoe und Abbott (2009) – offenbaren sich mittels OM Muster in Verlaufsdaten, die bislang mit keiner anderen Methode aufgedeckt werden konnten.

Auf der Grundlage unserer bisherigen Kenntnisse lassen sich folgende Anwendungsempfehlungen geben:

1. Der Zustandsraum sollte in Abhängigkeit von der inhaltlichen Fragestellung sorgfältig ausgewählt werden, da die Auswahl die Ergebnisse entscheidend beeinflusst.
2. Wenn möglich sollten differenzierte Substitutionskosten angewendet werden. Hierdurch können inhaltliche Unterschiede zwischen den Zuständen berücksichtigt werden.
3. Die Substitutionskosten sollten insgesamt nicht zu hoch angesetzt werden, damit sie nicht durch Indeloperationen ersetzt werden.
4. Als Clusteranalyseverfahren empfehlen wir den Weighted-Average-Linkage aus den in Abschnitt 9.6 genannten Gründen.
5. Bei dem in der Forschungspraxis häufig eingesetzten Ward-Verfahren ist Vorsicht angebracht, da die Levenshtein-Distanz keine varianzanalytische Interpretation des Verschmelzungsniveaus ermöglicht. Die Anwendungsvoraussetzungen für das Ward-Verfahren könnten durch eine vorgeschaltete (nichtmetrische) mehrdimensionale Skalierung, bei der ein euklidischer Merkmalsraum erzeugt wird, erreicht werden (siehe Kapitel 2).

20 Formale Gültigkeitsprüfung und Konsensuslösungen

Die Grundlogik der *formalen Gültigkeitsprüfung* wurde bereits in der Einleitung (Kapitel 1) skizziert. Die formale Gültigkeitsprüfung untersucht, wie gut eine oder mehrere Clusterlösungen bestimmte formale Anforderungen, die an die gesuchte Klassifikation gestellt werden, erfüllt. Dabei spielt es keine Rolle, ob diese Kriterien explizit durch das gewählte Verfahren in die Clusterbildung eingingen oder nicht. Eine Clusterlösung wird dann als formal gültig bezeichnet, wenn sie die gestellten formalen Anforderungen erfüllt. Ein absolutes Urteil ist hier oft schwer möglich. Daher ist es sinnvoll, mehrere formal geeignete Clusterlösungen zu betrachten und dann jene auszuwählen, die unter den untersuchten Lösungen am besten abschneidet (siehe Abschnitt 20.1). Ist eine Entscheidung nicht möglich, kann ein *Konsensusverfahren* eingesetzt werden (siehe Abschnitt 20.2). Erwähnt sei nochmals, dass inhaltlichen Kriterien der Vorzug zu geben ist, aber kein Ergebnis interpretiert werden sollte, welches formal nicht zulässig ist.

20.1 Formale Gültigkeitsprüfung

Auf Seiten der Anwenderin ist erforderlich, dass entsprechende Kriterien T_i definiert und operationalisiert werden. Es müssen dazu entsprechende Maßzahlen $M(T_i)$ ausgewählt und berechnet werden. Wenn die Maßzahlen unterschiedlich skaliert sind, was der Regelfall ist, ist eine Transformation auf eine einheitliche Skala erforderlich, damit die Maßzahlen untereinander verglichen und aufaddiert werden können. Eine einfache Transformation ist die Vergabe von Rangplätzen bzw. Punkten. Bei C zu vergleichenden Lösungen, gibt es C Punkte bzw. Rangplätze. Die Clusterlösung K^* , die beim Kriterium T_i am besten abschneidet, erhält den höchsten Rangplatz, also die meisten Punkte (C Punkte), die Clusterlösung K^{**} mit der zweitbesten Performanz bei T_i den zweithöchsten Rangplatz ($C - 1$ Punkte) usw. Da die rangtransformierten Maßzahlen vergleichbar sind,

kann ein additiver Gesamtpunktwert für jede untersuchte Clusterlösung K^* berechnet werden mit

$$F_{K^*} = \sum_i R(M(T_{K^*i})) .$$

Der Rangplatz, den die Clusterlösung K^* beim Kriterium T_i erreicht, wird mit $R(M(T_{K^*i}))$ bezeichnet. F_{K^*} ist ein Maß für die formale Gültigkeit der Clusterlösung K^* . Sind die eingesetzten Maßzahlen alle gleich skaliert, zum Beispiel von 0 bis 1, ist eine Verwendung von Rangplätzen nicht erforderlich. F_{K^*} kann direkt über die Maßzahlen berechnet werden:

$$F_{K^*} = \sum_i M(T_{K^*i}) .$$

Selbstverständlich ist es möglich, den Kriterien T_i auch unterschiedliche Gewichte w_i zu geben. Für F_{K^*} gilt dann:

$$F_{K^*} = \sum_i w_i \cdot M(T_{K^*i})$$

bzw.

$$F_{K^*} = \sum_i w_i \cdot R(M(T_{K^*i})) .$$

Das Vorgehen soll anhand eines Beispiels verdeutlicht werden. Beim K-Means-Verfahren gab es drei formal zulässige Clusterlösungen, eine 4-, 8- und 10-Clusterlösung. Es soll nun untersucht werden, welche der drei Lösungen die höchste formale Gültigkeit aufweist. Für die formale Gültigkeitsprüfung werden folgende Kriterien und Maßzahlen ausgewählt (siehe Tabelle 20.1 auf Seite 496):

- *Der anschließende Abfall des PRE-Koeffizienten:* Dazu wird der relative Abfall berechnet mit $(\text{PRE}_{K+1}/\text{PRE}_K)$.¹
- *Der maximale F-Wert:* Ein höherer F-Wert ist ein Hinweis auf eine formal besser geeignete Lösung (siehe Tabelle 12.5 auf Seite 310).
- *Die erklärte Varianz:* Dazu wird der z-Wert aus der Zufallstestung verwendet. Ein höherer z-Wert ist ein Hinweis auf eine formal besser geeignete Lösung (siehe Tabelle 12.7 auf Seite 315).
- *Die Größe des kleinsten Clusters:* Die Cluster sollten eine bestimmte Mindestgröße haben, damit sie auch stabil und reproduzierbar sind (siehe Tabelle 12.8 auf Seite 316).

¹ Als Berechnungsgrundlage dient die Tabelle 12.5 auf Seite 310. Um der Tatsache Rechnung zu tragen, dass bei höherer Clusterzahl die PRE-Koeffizienten und damit der absolute PRE-Abfall automatisch kleiner sind, wird der relative Abfall berechnet: Der PRE-Koeffizient der 4-Clusterlösung beträgt 0,258 und derjenige der 5-Clusterlösung 0,180. Der relative Abfall ist somit $0,180/0,258 = 0,698$ bzw 69,8 Prozent. Das bedeutet, dass der PRE-Koeffizient der 5-Clusterlösung um 30 Prozent kleiner als derjenige der 4-Clusterlösung ist.

- *Die Zahl der Cluster mit weniger als 10 Prozent an Repräsentanten:* Bei einer guten Clusterlösung sollte jedes Cluster homogen sein und daher möglichst viele Repräsentanten haben. Daher sollte es keine Cluster mit wenigen Repräsentanten geben – »wenig« wird mit 10 Prozent spezifiziert (für die 8-Clusterlösung siehe Tabelle 12.12 auf Seite 324, für die anderen nicht wiedergegeben).
- *Die Zahl der Ausreißer:* Die Zahl der Ausreißer in den Clustern sollte klein sein, da die Cluster homogen sein sollten (siehe Tabelle 12.12 auf Seite 324).
- *Die Zahl der Fälle im Überlappungsbereich:* Auch diese Zahl sollte klein sein, da die Cluster gut voneinander getrennt sein sollten (siehe Tabelle 12.12 auf Seite 324).
- *Die Stabilität gegenüber der vorausgehenden und nachfolgenden Lösung:* Diese sollte hoch sein. Dafür wird der RAND-Index für die vorausgehende und nachfolgende Lösung berechnet.²
- *Die Clusterzahl:* Es gilt das Einfachheitskriterium, weshalb eine geringere Clusteranzahl einer größeren vorzuziehen ist. Im vorliegenden Fall wird daher beispielsweise der 4-Clusterlösung der Vorzug vor der 8- und 10-Clusterlösung gegeben.

Die für die Rangplätze vergebenen Punkte stehen für jede Lösung in der zweiten Spalte. Da drei Lösungen verglichen werden, beträgt die maximale Punktzahl 3. Der in einem Kriterium formal besten Lösung werden drei Punkte gegeben, der zweitbesten Lösung zwei Punkte und der drittbesten ein Punkt. Bei Gleichheit wird der Durchschnitt der zu vergebenden Punkte eingesetzt. So zum Beispiel haben die 8- und 9-Clusterlösung nur einen Ausreißer und teilen sich damit die Punkte 2 und 3 (Durchschnitt 2,5). Werden die Punkte aufaddiert, entfallen 16,5 Punkte auf die 4-Clusterlösung, 19,0 auf die 8-Clusterlösung und 18,5 auf die 10-Clusterlösung. Es gilt also: $F_4 = 16,5$, $F_8 = 19,0$ und $F_{10} = 18,5$. Die 8- und 10-Clusterlösung unterscheiden sich damit nur minimal, die formale Gültigkeit der 4-Clusterlösung fällt deutlich geringer aus. Würde man das Kriterium der niedrigen Clusterzahl höher gewichten, dann würde man eindeutig die 8-Clusterlösung bevorzugen. Die 4-Clusterlösung scheidet immer aus.

Das hier dargestellte Vorgehen lässt sich für jede Clusterlösung bzw. Klassifikation anwenden, es ist also nicht auf das K-Means-Verfahren begrenzt. Neben den hier einbezogenen Kriterien können weitere Kriterien spezifiziert werden. Bei den probabilistischen Verfahren können beispielsweise zusätzlich die Überlappungsindizes in das Bewertungsschema einbezogen werden, bei den hierarchischen Verfahren die γ -Koeffizienten usw. (siehe Abschnitt 9.1.3). Sind mehr als drei Lösungen zulässig, werden mehr als drei Punkte vergeben. Bei fünf formal zulässigen Lösungen beispielsweise fünf Punkte.

Eine Übersicht über Maßzahlen zur formalen Gültigkeitsprüfung geben Fonseca und Cardoso (2007). Hier soll nur der sogenannten *Silhouette-Koeffizient* von Kaufman und

² Der RAND-Index wurde speziell für die formale Gültigkeitsprüfung berechnet.

Tab. 20.1: Formale Gültigkeitsindizes der zulässigen Clusterlösungen

	4-Clusterlösung		8-Clusterlösung		10-Clusterlösung	
	Wert	Punkte	Wert	Punkte	Wert	Punkte
PRE-Abfall in Prozent	69,8	2	67,5	1	73,4	3
F_{\max}	167,1	1	187,3	2	190,8	3
η^2 (z-Wert)	5,528	1	11,139	3	6,894	2
Größe des kleinsten Clusters	22	3	5	2	4	1
Zahl der Cluster mit weniger als 10 Prozent Repräsentanten	1	2	1	2	1	2
Zahl der Ausreißer	4	1	1	2,5	1	2,5
Fälle im Überlappungsbereich	15	2,5	15	2,5	18	1
RAND-Index (Durchschnitt vor- und nachher)	0,533	1	0,696	2	0,710	3
Zahl der Cluster	4	3	8	2	10	1
Gesamtpunkte (form. Gültigkeit F_k)		16,5		19,0		18,5

Rousseeuw (1990, S. 84–88) dargestellt werden, da er zunehmend zur Bestimmung der Cluster und zur formalen Gültigkeitsprüfung eingesetzt wird. Zunächst wird für jedes Objekt g ein Silhouette-Koeffizient $s(g)$ berechnet mit

$$s(g) = \frac{b(g) - a(g)}{\max(b(g), a(g))}.$$

Der Term $a(g)$ enthält die durchschnittlichen Distanzen des Objekts g zu allen anderen Objekten g^* des Clusters k , dem g angehört:

$$a(g) = \frac{1}{n_k - 1} \sum_{g^* \neq g}^{n_k} d_{g^*g}.$$

Der Term $b(g)$ enthält die durchschnittlichen Distanzen des Objekts g zu dem Cluster k^* , das aus der Perspektive von g am nächsten ist, für das also gilt

$$b(g) = \min_{k^* \neq k} (\bar{d}_{gk^*}),$$

wobei \bar{d}_{gk^*} die durchschnittliche Distanz von g zu den Objekten des Clusters k^* ist. Liegt in einem Cluster vollkommene Homogenität vor, ist $a(g)$ gleich 0 und der Silhouette-Koeffizient $s(g)$ ist gleich 1,0. Unterscheiden sich dagegen zwei Cluster nur sehr gering,

ist die Differenz $b(g) - a(g)$ klein und der Koeffizient liegt nahe 0. Auch der Fall kann auftreten, dass die durchschnittlichen Distanzen $b(g)$ kleiner sind als $a(g)$. In diesem Fall ist der Silhouette-Koeffizient negativ. Ist $b(i)$ gleich 0, ist er gleich -1,0.

Aus dem Silhouette-Koeffizient für ein Objekt g lässt sich ein Silhouette-Koeffizient $S(k)$ für ein Cluster k und ein Silhouette-Koeffizient $S(K)$ für eine Clusterlösung mit K Clustern berechnen:

$$S(k) = \frac{1}{n_k} \sum_k s(g) ,$$

$$S(K) = \max_k(S(k)) .$$

Ausgewählt wird jene Lösung, für die der Silhouette-Koeffizient maximal ist. Die Berechnung des Koeffizienten ist sehr rechenintensiv. Eine schnellere Rechenmethode für die euklidischen Distanzen führen Al-Zoubi und al Rawi (2008) an. Mitunter wird auch an Stelle der Distanzen zu allen Objekten eines Clusters die Distanz zum Clusterzentrum verwendet. Die Parameter sind für Objekt g , das dem Cluster k angehört:

$$a(g) = d_{gk} ,$$

$$b(g) = \min_{k^* \neq k} (d_{gk^*}) .$$

20.2 Konsensuslösungen

Heute geht man mitunter dazu über, nicht eine Clusterlösung zu betrachten, sondern eine Art »Durchschnitt« (*Konsensus*) aus mehreren formal möglichen Lösungen zu berechnen. Erwartet wird, dass dadurch eine stabilere Lösung erreicht wird. Die »Durchschnittsbildung« kann sich beziehen auf:

- die Dendrogramme,
- die Clustermittelwerte oder
- die Clusterzuordnungen mehrerer Clusterlösungen.

Wird ein Konsensus für *Dendrogramme* gesucht, wird dieser als »consensus-tree« bezeichnet. Ein spezielles Verfahren hierzu wurde von Lapointe und Legendre (1995) entwickelt.

Ein Konsensus für die *Clustermittelwerte* kann dadurch gewonnen werden, dass die Clustermittelwerte der in Frage kommenden Lösungen (zum Beispiel Clusterlösungen aus K-Means-Verfahren für unterschiedliche Varianzkriterien, Clusterlösung aus Ward-Verfahren usw.) als neue Datenmatrix gespeichert werden. Diese wird anschließend mittels

eines hierarchischen Verfahrens analysiert. Sinnvoll ist es dabei, die Clustermittelwerte entsprechend ihrer Größe zu gewichten.

Ein Konsensus für die *Clusterzuordnungen* kann ermittelt werden, indem die Clusterzugehörigkeiten für die in Betracht kommenden Clusterlösungen abgespeichert und anschließend in eine Clusteranalyse als nominale Variablen einbezogen werden. Soll beispielsweise aus drei Clusterlösungen CL_1 , CL_2 und CL_3 ein Konsensus auf der Basis der Zuordnungen berechnet werden, können die Clusterzugehörigkeiten der drei Lösungen – beispielsweise aufgelöst in Dummies – als nominale Variablen in die Clusteranalyse einbezogen werden.

20.2.1 Konsensus für Clustermittelwerte

Das Vorgehen für die Ermittlung einer Konsensuslösung auf der Basis der Clustermittelwerte soll für das Beispiel der Wertedaten von Denz veranschaulicht werden. Einbezogen werden folgende Lösungen:

- 4-Clusterlösung des K-Means-Verfahren. Die Cluster werden abgekürzt mit $4K_{C_1}$, $4K_{C_2}$, $4K_{C_3}$ und $4K_{C_4}$. Die erste Ziffer steht für die Zahl der Cluster. Die an die erste Ziffer folgenden Buchstaben drücken das Verfahren aus, »K« steht für K-Means.
- 8-Clusterlösung des K-Means-Verfahren. Die Cluster werden abgekürzt mit $8K_{NO}$, $8K_{GPMAT}$, $8K_{R_1}$ und $8K_{R_2}$. Die erste Ziffer steht wiederum für die Zahl der Cluster. Die an die erste Ziffer folgenden Buchstaben drücken das Verfahren aus, »K« steht für K-Means. Für die Cluster wurden nicht die formalen Bezeichnungen C_1 , C_2 usw. gewählt, sondern inhaltlich Bezeichnungen. »NO« steht für Nichtorientierte, »GPMAT« für gemäßigte Postmaterialisten usw.
- 10-Clusterlösung des K-Means-Verfahren. Die Cluster werden abgekürzt mit $10K_{C_1}$, $10K_{C_2}$, $10K_{C_3}$ usw.
- 6-Clusterlösung der latenten Profilanalyse mit ALMO. Die Cluster werden abgekürzt mit $6LPC_{C_1}$, $6LPC_{C_2}$, $6LPC_{C_3}$ usw.
- 5-Clusterlösung der Latent-GOLD-Lösung. Die Cluster werden abgekürzt mit $5LC_{C_1}$, $5LC_{C_2}$, $5LC_{C_3}$ usw.
- 3-Clusterlösung der Latent-GOLD-Lösung. Die Cluster werden abgekürzt mit $3LC_{C_1}$, $3LC_{C_2}$ und $3LC_{C_3}$.

Die Struktur der Datenmatrix zeigt Tabelle 20.2. Die Anteilswerte der Cluster werden zur Gewichtung verwendet. Für die Datenmatrix wird eine hierarchisch-agglomerative Analyse gerechnet. Die Fälle (Cluster) werden entsprechend den Anteilswerten gewichtet. Als Klassifikationsmerkmale gehen die Klassenmittelwerte im Postmaterialismus und Materialismus in die Analyse ein. Als Verfahren wird der Weighted-Average-Linkage

Tab. 20.2: Datenmatrix der Konsensusanalyse für unterschiedliche Clusterlösungen der Denzdaten

Cluster	Anteilswert	gMat	gPmat
8K _{NO}	0,104	2,03	2,30
8K _{GPMAT}	0,244	1,98	1,40
8K _{R₁}	0,023	2,53	3,40
8K _{R₂}	0,023	4,30	2,00
⋮	⋮	⋮	⋮

Abkürzungen: siehe laufenden Text sowie
Tabelle 12.8 auf Seite 316

ausgewählt, als Distanzmaß die quadrierten euklidischen Distanzen. Die Ergebnisse sind in der Abbildung 20.1 auf der nächsten Seite dargestellt. In dem Dendrogramm sind sechs Hügel, also sechs Cluster erkennbar. Zwei Clustern gehören die Restcluster der 8-Clusterlösung an. Sie werden daher als Restcluster bezeichnet. Die anderen vier Cluster lassen sich als Nicht-Orientierte, als Konsenstypus, als Postmaterialisten und als extreme Postmaterialisten interpretieren. Die Clustermittelwerte zeigt die Tabelle 20.3. Das Konsensusverfahren legt eine 6-Clusterlösung mit vier inhaltlich gut interpretierbaren und zwei Restclustern mit Ausreißern nahe.

20.2.2 Konsensus auf der Basis der Clusterzuordnungen

Zur Ermittlung einer Konsensuslösung auf der Basis der Clusterzuordnungen müssen die Clusterzuordnungen der Objekte abgespeichert werden. Diese werden dann als Klassifikationsvariable in eine Clusteranalyse einbezogen. In Tabelle 20.4 auf Seite 501 ist auszugsweise die neue Datenmatrix abgebildet. In dem Beispiel soll aus zwei Clusterlösungen eine Konsensuslösung berechnet werden. Zwei Cluster werden hier zur besseren Veranschaulichung ausgewählt. In der Regel wird das Verfahren für eine größere Zahl von Clusterlösungen durchgeführt. Die Konsensuslösung soll in dem Beispiel mit dem

Tab. 20.3: Clustermittelwerte aus der Konsensuslösung auf der Basis der Clustermittelwerte unterschiedlicher Lösungen

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆
n	22	114	2	3	62	11
gMat	2,19	1,63	2,53	4,44	2,44	3,33
gPmat	2,32	1,51	3,43	1,94	1,40	1,20

Abkürzungen: siehe Tabelle 12.8 auf Seite 316

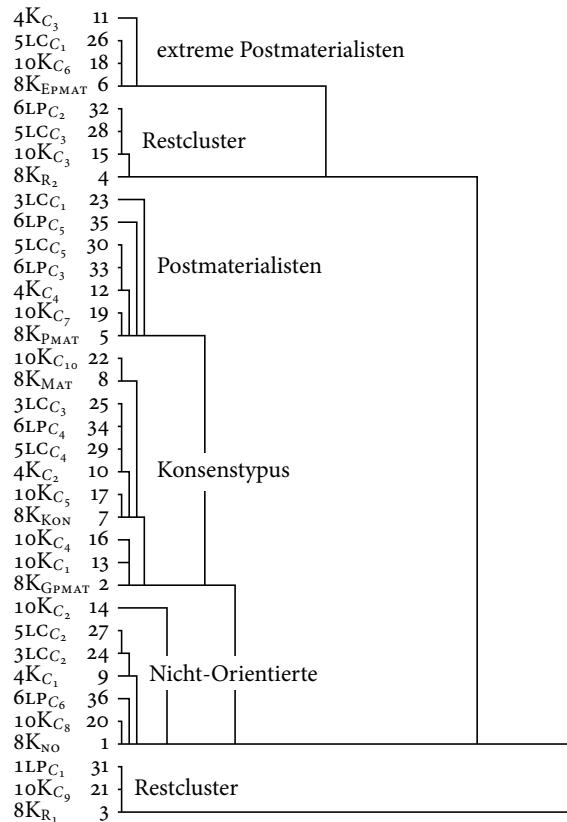


Abb. 20.1: Dendrogramm für die Konsensusrechnung

K-Means-Verfahren ermittelt werden. Da die Clusterzugehörigkeit eine nominale Variable ist, wird sie hier in Dummies aufgelöst, geclustert wird dann mit den Dummies. Die Clusterzuordnung und die Skalenwerte für den Postmaterialismus und Materialismus gehen nur als deskriptive Variablen ein, sie bleiben zur Clusterbildung inaktiv. Die Ergebnisse fasst Tabelle 20.5 zusammen. Dem Konsensuscluster C_1 gehören 21,72 Prozent der Fälle an. Diese kommen zu 90 Prozent aus dem dritten Cluster ($1C_3$) und zu 10 Prozent aus dem sechsten Cluster ($1C_6$) der ersten Clusterlösung. Es rekrutiert sich zu 100 Prozent aus dem sechsten Cluster ($2C_6$) der zweiten Clusterlösung. Der Mittelwert im Materialismus beträgt 2,36 und jener im Postmaterialismus 1,54. Es bildet den Typus der Postmaterialisten ab. Konsensuscluster C_2 wird nur von wenigen Fällen gebildet. Es handelt sich um ein Restcluster, das dadurch gekennzeichnet ist, dass beide Wertorientierungen eher unwichtig sind, wobei der Postmaterialismus stärker abgelehnt wird. Konsensuscluster C_3 lässt sich als »extreme Postmaterialisten« bezeichnen. Konsensuscluster C_5 repräsentiert den Konsenstypus, Konsensuscluster C_6 die Nicht-

Tab. 20.4: Ausschnitt aus der Datenmatrix

Obj.	CL ₁	CL ₂	gMat	gPmat	1C ₁	1C ₂	1C ₃	1C ₄	1C ₅	1C ₆	2C ₁	2C ₂	2C ₃	2C ₄	2C ₅	2C ₆
1	3	1	2,83	1,50	0	0	1	0	0	0	1	0	0	0	0	0
2	3	6	2,33	1,67	0	0	1	0	0	0	0	0	0	0	0	1
3	4	6	2,00	1,67	0	0	0	1	0	0	0	0	0	0	0	1
4	4	3	1,33	1,17	0	0	0	1	0	0	0	0	1	0	0	0
5	4	3	1,83	1,17	0	0	0	1	0	0	0	0	1	0	0	0
6	3	6	2,17	1,67	0	0	1	0	0	0	0	0	0	0	0	1
usw.																

Orientierten. Konsensuscluster C_4 wird wiederum von wenigen Fällen gebildet und lässt sich als Restcluster auffassen. Die bisherigen Ergebnisse zur Wertetypologie werden bestätigt. In dem Beispiel wurde die Konsensuslösung mit dem K-Means-Verfahren ermittelt. Eingesetzt werden können alle Verfahren. So zum Beispiel kann die Analyse latenter Klassen eingesetzt werden mit dem Vorteil, dass eine Auflösung in Dummies nicht erforderlich ist.

Tab. 20.5: Konsensuslösung aus zwei Clusterlösungen auf der Basis von Clusterzuordnungen

	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆
<i>Anteilswerte (%)</i>						
	21,72	4,07	19,46	2,26	33,48	19,01
1C ₁	0,00	0,33	0,00	0,00	0,00	0,00
1C ₂	0,00	0,00	0,00	0,80	0,00	0,00
1C ₃	0,90	0,00	0,16	0,00	0,00	0,00
1C ₄	0,00	0,00	0,00	0,00	1,00	0,57
1C ₅	0,00	0,00	0,81	0,00	0,00	0,00
1C ₆	0,10	0,67	0,02	0,20	0,00	0,43
2C ₁	0,00	0,00	0,56	0,00	0,00	0,00
2C ₂	0,00	0,00	0,00	0,00	0,00	1,00
2C ₃	0,00	0,00	0,16	0,00	0,91	0,00
2C ₄	0,00	1,00	0,00	0,00	0,00	0,00
2C ₅	0,00	0,00	0,00	1,00	0,00	0,00
2C ₆	1,00	0,00	0,28	0,00	0,09	0,00
<i>Mittelwerte</i>						
gMat	2,36	2,48	2,58	4,30	1,58	1,72
gPmat	1,54	3,00	1,15	2,00	1,45	2,05

Literatur

- Aassve, Arnstein, Francesco C. Billari und Raffaella Piccarreta (2007). »Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories«. In: *European Journal of Population* 23.4, S. 369–388.
- Abbott, Andrew (1995). »Sequence Analysis: New Methods fort Old Ideas«. In: *Annual Review of Sociology* 21, S. 93–113.
- Abbott, Andrew und John Forrest (1986). »Optimal Matching Methods for Historical Sequences«. In: *Journal of Interdisciplinary History* 16.3, S. 471–494.
- Abbott, Andrew und Alexandra Hrycak (1990). »Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers«. In: *American Journal of Sociology* 96.1, S. 144–185.
- Abbott, Andrew und Angela Tsay (2000). »Sequence Analysis and Optimal Matching Methods in Sociology. Review and Prospect«. In: *Sociological Methods & Research* 29.1, S. 3–33.
- Agresti, Alan (2002). *Categorical Data Analysis*. 2. Aufl. Hoboken, NJ: Wiley.
- Aisenbrey, Silke (2000). *Optimal Matching Analyse. Anwendungen in den Sozialwissenschaften*. Opladen: Leske + Budrich.
- Aisenbrey, Silke und Anette E. Fasang (2010). »New Life for Old Ideas: The ›Second Wave‹ of Sequence Analysis Bringing the ›Course‹ Back Into the Life Course«. In: *Sociological Methods and Research* 38.3, S. 420–462.
- Akaike, Hirotugu (1973). »Information Theory and an Extension of the Maximum Likelihood Principle«. In: *Second International Symposium on Information Theory*. Budapest: Akadémiai Kiadó, S. 267–281.
- Akaike, Hirotugu (1974). »A New Look at the Statistical Model Identification«. In: *IEEE Transactions on Automatic Control* 19.6, S. 716–723.
- Akaike, Hirotugu (1987). »Factor Analysis and AIC«. In: *Psychometrika* 52.3, S. 317–332.
- Al-Zoubi, Moh'd Belal und Mohammad al Rawi (2008). »An Efficient Approach for Computing Silhouette Coefficients«. In: *Journal of Computer Science* 4.3, S. 252–255.
- Aldenderfer, Mark S. und Robert K. Blashfield (1984). *Cluster Analysis*. Beverly Hills, London u. a.: Sage Publications.
- Andersen, Erling B. (1991). *The Statistical Analysis of Categorical Data*. 2. Aufl. Berlin, Heidelberg u. a.: Springer.

- Andrefß, Hans-Jürgen, Jacques A. Hagenaars und Steffen-M. Kühnel (1997). *Analyse von Tabellen und kategorialen Daten. Log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz*. Berlin, Heidelberg u. a.: Springer.
- Andrews, Rick L. und Imran S. Currim (2003). »A Comparison of Segment Retention Criteria for Finite Mixture Logit Models«. In: *Journal of Marketing Research* 40.2, S. 235–243.
- Arminger, Gerhard (1979). *Faktorenanalyse*. Stuttgart: Teubner.
- Arnold, Stephen John (1979). »A Test for Clusters«. In: *Journal of Marketing Research* 16.4, S. 545–551.
- Bacher, Johann (1986). *Faktorenanalyse und Modelle des Antwortverhaltens*. Wien.
- Bacher, Johann (1987). »Faktorenanalyse von Rangordnungen«. In: *Österreichische Zeitschrift für Soziologie* 12.2, S. 85–89.
- Bacher, Johann (1990). »Einführung in die Logik der Skalierungsverfahren«. In: *Historical Social Research* 15, S. 4–171.
- Bacher, Johann (1992). »Welchen zusätzlichen Erkenntnisgewinn können statistische Verfahren vermitteln, die über die üblicherweise verwendeten Verfahren hinausgehen?« In: *Österreichische Zeitschrift für Soziologie* 17, S. 52–63.
- Bacher, Johann (1994). *Clusteranalyse*. 1. Aufl. München, Wien: Oldenbourg.
- Bacher, Johann (1995a). »Goodness-of-Fit Measures for Multiple Correspondence Analysis«. In: *Quality & Quantity* 29.1, S. 1–16.
- Bacher, Johann (1995b). »Latenter Wertedissens oder Konsens in Österreich? Ergebnisse einer Sekundäranalyse des Sozialen Survey Österreichs 1993«. In: *SWS-Rundschau* 1995.2, S. 175–199.
- Bacher, Johann (1996). *Clusteranalyse*. 2. Aufl. München, Wien: Oldenbourg.
- Bacher, Johann (2000). »A probabilistic clustering model for variables of mixed type«. In: *Quality & Quantity* 34.3, S. 223–235.
- Bacher, Johann (2002). »Statistisches Matching: Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS«. In: *ZA-Information* 51, S. 38–66.
- Bacher, Johann und Jeroen K. Vermunt (2010). »Analyse latenter Klassen«. In: *Handbuch der sozialwissenschaftlichen Datenanalyse*. Hrsg. von Christof Wolf und Henning Best. im Erscheinen. Wiesbaden: vs.
- Bacher, Johann, Knut Wenzig und Melanie Vogler (2004). *SPSS TwoStep Cluster – A First Evaluation*. 2. Aufl. Arbeits- und Diskussionspapiere des Lehrstuhls für Soziologie & Empirische Sozialforschung der Friedrich-Alexander-Universität Nürnberg-Erlangen. Bd. 2004-2. Nürnberg: Selbstverlag.
- Backer, Eric (1978). *Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets*. Delft: Delft University Press.
- Bailey, Kenneth D. (1975). »Cluster Analysis«. In: *Sociological Methodology* 6, S. 59–128.

- Banfield, Jeffrey D. und Adrian E. Raftery (1993). »Model-based Gaussian and non-Gaussian clustering«. In: *Biometrics* 49.3, S. 803–821.
- Bartlett, Maurice S. (1950). »Tests of Significance in Factor Analysis«. In: *British Journal of Psychology. Statistical Section* 3, S. 77–85.
- Batagelj, Vladimir (1981). »Note on Ultrametric Hierarchical Clustering Algorithms«. In: *Psychometrika* 46.3, S. 351–353.
- Behboodian, Javad (1970). »On the Models of Mixture of Two Normal Distributions«. In: *Technometrics* 12, S. 131–139.
- Bentler, Peter M. (1985). *Theory and Implementation of EQS: A Structural Equations Program*. Los Angeles: BMDP Statistical Software.
- Bethmann, Arne und Knut Wenzig (2008). »Open Source Clustering«. Naples (Italy), Campus di Monte Sant'Angelo, »RC33 7th International Conference on Social Science Methodology«.
- Bethmann, Arne und Knut Wenzig (2010). »Lehrbuchbeispiele für SPSS TwoStep Cluster – Reanalysiert mit AutoClass und Latent GOLD«. Dortmund, Technische Universität, »Statistik unter einem Dach«, 2. gemeinsame Tagung der Deutschen Arbeitsgemeinschaft Statistik.
- Biernacki, Christophe, Gilles Celeux und Gérard Govaert (2000). »Assessing a Mixture model for Clustering with the integrated Completed Likelihood«. In: *IEEE Transactions on Pattern analysis and Machine Intelligence* 22.7, S. 719–725.
- Bijnen, Emanuel Joseph (1973). *Cluster Analysis. Survey and Evaluation of Techniques*. Groningen.
- Blasius, Jörg (2001). *Korrespondenzanalyse*. München und Wien.
- Blossfeld, Hans-Peter, Katrin Golsch und Götz Rohwer (2007). *Event History Analysis with Stata*. New York, London: Lawrence Erlbaum Associates.
- Bock, Hans Hermann (1974). *Automatische Klassifikation*. Göttingen: Vandenhoeck & Ruprecht.
- Bock, Hans Hermann (1989). »Probabilistic Aspects in Cluster Analysis«. In: *Conceptual and Numerical Analysis of Data*. Hrsg. von Otto Optiz. Berlin, Heidelberg u. a.: Springer, S. 12–44.
- Bock, R. Darrell und Murray Aitkin (1981). »Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm«. In: *Psychometrika* 46.4, S. 443–459.
- Bollen, Kenneth A. (1989). *Structural Equations with Latent Variables*. New York, Chichester u. a.: Wiley.
- Borg, Ingwer (1981). *Anwendungsorientierte Multidimensionale Skalierung*. Berlin, Heidelberg u. a.: Springer.
- Bortz, Jürgen (2005). *Statistik für Human- und Sozialwissenschaftler*. 6. Aufl. Heidelberg: Springer Medizin Verlag.

- Bortz, Jürgen, Gustav A. Lienert und Klaus Boehnke (2008). *Verteilungsfreie Methoden in der Biostatistik*. 3. Aufl. Heidelberg: Springer Medizin Verlag.
- Bozdogan, Hamparsum (1987). »Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions«. In: *Psychometrika* 52.3, S. 345–370.
- Bozdogan, Hamparsum (1993). »Choosing the Number of Component Clusters in the Mixtures-Model Using a new Informational Complexity Criterion of the Inverse-Fisher Information Matrix«. In: *Information and Classification. Concepts, Methods, and Application: Proceedings of the 16th Annual Conference of the »Gesellschaft für Klassifikation«*. Hrsg. von Otto Opitz, Berthold Lausen und Rüdiger Klar. Berlin, Heidelberg, New York: Springer, S. 40–54.
- Brosius, Felix (2008). *SPSS 16*. Heidelberg: Redline.
- Brüderl, Josef und Stefani Scherer (2004). »Methoden zur Analyse von Sequenzdaten«. In: *Methoden der Sozialforschung*. Hrsg. von Andreas Diekmann. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 44. Wiesbaden: VS, S. 330–347.
- Bryant, Peter G. (1991). »Large-Sample Results for Optimization-Based Clustering Methods«. In: *Journal of Classification* 8.1, S. 31–44.
- Brzinsky-Fay, Christian, Ulrich Kohler und Magdalena Luniak (2006). »Sequence Analysis with Stata«. In: *The Stata Journal* 4, S. 435–460.
- Buchmann, Marlis und Stefan Sacchi (1995). »Mehrdimensionale Klassifikation beruflicher Verlaufsdaten«. In: *Sozialstruktur und Lebenslauf*. Hrsg. von Peter A. Berger und Peter Sopp. Sozialstrukturanalyse 5. Opladen: Leske + Budrich, S. 49–64.
- Bunge, Mario Augusto (1967). *Scientific Research II. The Search for Truth*. Berlin, Heidelberg u. a.: Springer.
- Calinski, Tadeusz und Jerzy Harabasz (1974). »A dendrite method for cluster analysis«. In: *Communications in Statistics – Theory and Methods* 3.1, S. 1–27.
- Carroll, J. Douglas und Jih-Jie Chang (1970). »Analysis of Individual Differences in Multidimensional Scaling via an N-way Generalization of Eckart-Young Decomposition«. In: *Psychometrika* 35.3, S. 283–319.
- Carroll, J. Douglas, Paul E. Green und Catherine M. Schaffer (1986). »Interpoint Distance Comparisons in Correspondence Analysis«. In: *Journal of Marketing Research* 23.3, S. 271–280.
- Carroll, J. Douglas, Paul E. Green und Catherine M. Schaffer (1989). »Reply to Greenacre's Commentary on the Carroll-Green-Schaffer Scaling of Two-Way Correspondence Analysis Solutions«. In: *Journal of Marketing Research* 26.3, S. 366–368.
- Cattell, Raymond B. (1949). » r_p and other Coefficients of Pattern Similarity«. In: *Psychometrika* 14.4, S. 279–298.
- Cattell, Raymond B. (1966). »The scree test for the number of factors«. In: *Multivariate Behavioral Research* 1.2, S. 245–276.

- Charrad, Malika, Yves Lechevallier, Mohamed Ben A. Ahmed und Gilbert Saporta (2010). »On the Number of Clusters in Block Clustering Algorithms«. In: *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, S. 392–397. URL: www.aaai.org/ocs/index.php/FLAIRS/2010/paper/download/1276/1781.
- Chaturvedi, Anil, Paul E. Green und J. Douglas Carroll (2001). »K-modes Clustering«. In: *Journal of Classification* 18.1, S. 35–55.
- Cheeseman, Peter und John Stutz (1996). »Bayesian Classification (AutoClass): Theory and Results«. In: *Advances in Knowledge Discovery and Data Mining*. Hrsg. von Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth und Ramasamy Uthurusamy. Massachusetts: AAAI Press / MIT Press, S. 153–180.
- Chen, Ke (2006). »On k-Median clustering in high dimensions«. In: *SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. New York, NY: ACM, S. 1177–1185.
- Chickering, David M. und David Heckerman (1997). »Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables«. In: *Machine Learning* 29.2–3, S. 181–212.
- Chiou, Tom, DongPing Fang, John Chen, Yao Wang und Christopher Jeris (2001). »A robust and scalable clustering algorithm for mixed type attributes in large database environment«. In: *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY: ACM, S. 263–268.
- Clogg, Clifford C. (1981). »New Developments in Latent Structure Analysis«. In: *Factor Analysis and Measurement in Sociological Research: A Multi-Dimensional Perspective*. Hrsg. von David J. Jackson und Edgar F. Borgatta. Beverly Hills, CA: Sage Publications, S. 215–246.
- Clogg, Clifford C., Donald B. Rubin, Nathaniel Schenker, Bradley Schultz und Lynn Weidman (1991). »Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression«. In: *Journal of the American Statistical Association* 86.413, S. 68–78.
- Cohen, Jacob (1968). »Multiple Regression as a General Data-Analytic System«. In: *Psychological Bulletin* 70.6, S. 426–443.
- Coombs, Clyde H. (1964). *Theory of Data*. New York: Wiley.
- Coombs, Clyde H. (1975). »A Note on the Relation between the Vector and the Unfolding Model for Preferences«. In: *Psychometrika* 40.1, S. 115–166.
- Cormack, Robert M. (1971). »A Review of Classification«. In: *Journal of the Royal Statistical Society, Series A (General)* 134.3, S. 321–367.
- Coxon, Anthony P. M. (1982). *The User's Guide to Multidimensional Scaling, with Special Reference to MDS(x) Library of Computer Programs*. London: Ashgate Publishing Limited.

- Cronbach, Lee J. (1951). »Coefficient alpha and the internal structure of tests«. In: *Psychometrika* 16.3, S. 297–334.
- Cunningham, James P. (1978). »Free Trees and Bidirectional Trees as Representations of Psychological Distance«. In: *Journal of Mathematical Psychology* 17.2, S. 165–188.
- Davier, Matthias von (1997). *Methoden zur Prüfung probabilistischer Testmodelle*. Zugl. Dissertation an der Universität Kiel, 1996. Kiel: IPN.
- Day, William H. E. und Fred R. McMorris (1994). »Alignment, Comparison and Consensus of Molecular Sequences«. In: *New Approaches in Classification and Data Analysis*. Hrsg. von Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand und Bernard Burtschy. Berlin: Springer, S. 327–346.
- Dayton, C. Mitchell und George B. Macready (1988). »Concomitant-Variable Latent-Class Models«. In: *Journal of the American Statistical Association* 83.401, S. 173–178.
- De Soete, Geert und Wayne S. DeSarbo (1991). »A Latent Probit Model for Analyzing Pick Any/N Data«. In: *Journal of Classification* 8.1, S. 45–64.
- Dempster, Arthur P., Nan M. Laird und Donald B. Rubin (1977). »Maximum Likelihood from Incomplete Data via the EM Algorithm«. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, S. 1–38.
- Denz, Hermann (1977). »Regressionsanalyse mit ordinalen Variablen«. In: *Die Befragung* 5. Hrsg. von Kurt Holm. München: Francke, S. 103–122.
- Denz, Hermann (1982). *Analyse latenter Strukturen*. München: Francke.
- Denz, Hermann (1983). »Entfremdung und Wertewandel«. In: *Österreichische Zeitschrift für Soziologie* 8.4, S. 121–136.
- Denz, Hermann (1989). *Einführung in die empirische Sozialforschung*. Wien und New York: Springer.
- DeSarbo, Wayne S. (1982). »GENNCLUS: New Models for General Nonhierarchical Clustering Analysis«. In: *Psychometrika* 47.4, S. 449–475.
- DeSarbo, Wayne S. und Vijay Mahajan (1984). »Constrained Classification: The Use of A Priori Information in Cluster Analysis«. In: *Psychometrika* 49.2, S. 187–215.
- Dias, José M. G. (2004). *Finite Mixture Models: Review, Applications, and Computer-intensive Methods*. Phd. Dissertation.
- Dijkstra, Wil und Toon Taris (1995). »Measuring the Agreement Between Sequences«. In: *Sociological Methods & Research* 24.2, S. 214–232.
- Dreger, Ralph M. (1986). »Microcomputer Programs for the Rand Index of Cluster Similarity«. In: *Educational and Psychological Measurement* 46.3, S. 655–661.
- Dunn, Graham und Brian S. Everitt (1982). *An Introduction to Mathematical Taxonomy*. Cambridge, London u. a.: Cambridge University Press.
- Dunn, Joseph C. (1976). »Indices of Partition Fuzziness and the Detection of Clusters in Large Data Sets«. In: *Fuzzy Automata and Decision Processes*. Hrsg. von Madan M. Gupta. New York: North-Holland.

- Dunson, David B. (2001). »Commentary: Practical Advantages of Bayesian analysis in Epidemiologic«. In: *American Journal of Epidemiology* 153.12, S. 1222–1226.
- Dziuban, Charles D. und Edwin C. Shirkey (1974). »When is a Correlation Matrix Appropriate for Factor Analysis? Some Decision Rules«. In: *Psychological Bulletin* 81.6, S. 358–361.
- Efron, Bradley (1979). »Bootstrap Methods: Another Look at the Jackknife«. In: *The Annals of Statistics* 7.1, S. 1–26.
- Eisen, Michael und Michiel de Hoon (2002). *Cluster 3.0 Manual*. URL: <http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/cluster3.pdf> (besucht am 12. 01. 2010).
- Elzinga, Cees (2003). »Sequence Similarity: A Nonaligning Technique«. In: *Sociological Methods & Research* 32.1, S. 3–29.
- Erzberger, Christian (2001). »Sequenzmusteranalyse als fallorientierte Analysestrategie«. In: *Strukturen des Lebenslaufs. Übergang – Sequenz – Verlauf*. Hrsg. von Reinhold Sackmann und Matthias Würgens. Weinheim und München: Juventa, S. 135–162.
- Erzberger, Christian und Gerald Prein (1997). »Optimal-Matching-Technik: Ein Analyseverfahren zur Vergleichbarkeit und Ordnung individuell differenter Lebensverläufe«. In: *ZUMA-Nachrichten* 40, S. 52–80.
- Espeland, Mark A. und Stanley L. Handelman (1989). »Using Latent Class Models to Characterize and Assess Relative Error in Discrete Measurements«. In: *Biometrics* 45.2, S. 587–599.
- Esping-Andersen, Gøsta (1990). *The three worlds of welfare capitalism*. Cambridge: Princeton University Press.
- Everitt, Brian S. (1980). *Cluster Analysis*. 2. Aufl. London: John Wiley & Sons Inc.
- Everitt, Brian S. (1988). »A Monte Carlo Investigation of the Likelihood Ratio Test for Number of Classes in Latent Class Analysis«. In: *Multivariate Behavioral Research* 23.4, S. 531–538.
- Everitt, Brian S., Sabine Landau und Morven Leese (2001). *Cluster Analysis*. 4. Aufl. London: Arnold Publishers.
- Fahrmeir, Ludwig und Alfred Hamerle (1984). »Grundlegende multivariate Schätz- und Testprobleme«. In: *Multivariate statistische Verfahren*. Hrsg. von Ludwig Fahrmeir und Walter Häußler. Berlin und New York: de Gruyter, S. 49–82.
- Ferligoj, Anuška und Vladimir Batagelj (1982). »Clustering with Relational Constraint«. In: *Psychometrika* 47.4, S. 413–426.
- Ferligoj, Anuška und Vladimir Batagelj (1983). »Some Types of Clustering with Relational Constraints«. In: *Psychometrika* 48.4, S. 541–552.
- Fisher, Ronald A. (1936). »The use of multiple measurements in taxonomic problems«. In: *Annals of Eugenics* 7, S. 179–188.
- Fisz, Marek (1980). *Wahrscheinlichkeitsrechnung und mathematische Statistik*. 10. Aufl. Berlin: Deutscher Verlag der Wissenschaften.

- Fleiss, Joseph L. (1981). *Statistical Methods for Rates and Proportions*. 2. Aufl. New York, Chichester u. a.: Wiley.
- Fonseca, Jaime R. S. und Margarida G. M. S. Cardoso (2007). »Mixture-Model Cluster Analysis using Information Theoretical Criteria«. In: *Intelligent Data Analysis* 11.2, S. 155–173.
- Forgy, Edward W. (1965). »Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications«. In: *Biometrics* 21. Abstract, S. 768–769.
- Fox, John (1982). »Selective Aspects of Measuring Resemblance for Taxonomy«. In: *Classifying Social Data. New Applications of Analytic Methods for Social Science Research*. Hrsg. von Herschel C. Hudson. San Francisco, Washington u. a.: Jossey-Bass, S. 127–151.
- Fraboni, Maryann und Robert Saltstone (1992). »The WAIS-R Number-of-Factors Quandary: A Cluster Analytic Approach to Construct Validation«. In: *Educational and Psychological Measurement* 52.3, S. 603–613.
- Fraley, Chris und Adrian E. Raftery (1999). »MCLUST: Software for Model-Based Cluster Analysis«. In: *Journal of Classification* 16.2, S. 297–306.
- Fraley, Chris und Adrian E. Raftery (2006). *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*. Technical Report 504. Seattle: Department of Statistics, University of Washington.
- Friedman, Herman P. und Jerrold Rubin (1967). »On Some Invariant Criteria for Grouping Data«. In: *Journal of the American Statistical Association* 62.320, S. 1159–1178.
- Frühwirth-Schnatter, Sylvia (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.
- Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer und Nicolas S. Müller (2009). *Mining Sequence Data in R with the TraMineR Package: A User's Guide*. URL: <http://mephisto.unige.ch/traminer> (besucht am 25. 03. 2010).
- Gauthier, Jacques-Antoine, Eric D. Widmer, Philipp Bucher und Cédric Notredame (2009). »How Much Does It Cost?: Optimization of Costs in Sequence Analysis of Social Science Data«. In: *Sociological Methods & Research* 38.1, S. 197–231.
- Gelman, Andrew, John B. Carlin, Hal S. Stern und Donald B. Rubin (1995). *Bayesian Data Analysis*. London: Chapman und Hall.
- Gerich, Joachim (2001). *Nichtparametrische Skalierung nach Mokken*. Linz: Trauner Verlag.
- Gibson, Wilfred A. (1959). »Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis«. In: *Psychometrika* 24.3, S. 229–252.
- Giegler, Helmut (1985). »Gewinnung einer Typologie zur Klassifikation von Freizeitaktivitäten – Eine explorative Studie«. In: *Angewandte Sozialforschung* 13.4, S. 339–364.

- Goodman, Leo A. (1974). »The Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable. Part I—a Modified Latent Structure Approach«. In: *American Journal of Sociology* 79.5, S. 1179–1259.
- Goodman, Leo A. und William H. Kruskal (1954). »Measures of Association for Cross Classifications«. In: *Journal of the American Statistical Association* 49.268, S. 732–764.
- Goodman, Leo A. und William H. Kruskal (1959). »Measures of Association for Cross Classifications. II: Further Discussion and References«. In: *Journal of the American Statistical Association* 54.285, S. 123–163.
- Goodman, Leo A. und William H. Kruskal (1963). »Measures of Association for Cross Classifications III: Approximate Sampling Theory«. In: *Journal of the American Statistical Association* 58.302, S. 310–364.
- Goodman, Leo A. und William H. Kruskal (1972). »Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances«. In: *Journal of the American Statistical Association* 67.338, S. 415–421.
- Gordon, Allen D. (1981). *Classification*. London und New York: Chapman & Hall.
- Gordon, Allen D. (1999). *Classification*. 2. Aufl. London und New York: Chapman & Hall.
- Gowda, K. Chidananda und G. Krishna (1978). »Agglomerative Clustering Using the Concept of Mutual Nearest Neighborhood«. In: *Pattern Recognition* 10.2, S. 105–112.
- Gower, John C. (1971). »A General Coefficient of Similarity and Some of Its Properties«. In: *Biometrics* 27.4, S. 857–871.
- Green, Paul E., Frank J. Carmone und Jonathan Kim (1990). »A Preliminary Study of Optimal Variable Weighting in K-Means-Clustering«. In: *Journal of Classification* 7.2, S. 271–285.
- Green, Paul Edgar und Donald F. Tull (1982). *Methoden und Techniken der Marktforschung*. 4. Aufl. Stuttgart: Schäffer-Poeschel.
- Greenacre, Michael J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.
- Greenacre, Michael J. (1989). *Theory and Applications of Correspondence Analysis*. London.
- Gusfield, Dan (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.
- Guttman, Louis (1941). »The Quantification of a Class of Attributes: A Theory and Method of Scale Construction«. In: *The Prediction of Personal Adjustment*. Hrsg. von Paul Horst, Paul Wallin und Louis Guttman. New York: Social Science Research Council.
- Guttman, Louis (1950a). »The Basis for Scalogram Analysis«. In: *Studies in Social Psychology in World War II*. Bd. IV: *Measurement and Prediction*. Hrsg. von Samuel A. Stouffer, Louis Guttman u. a. Princeton, NJ: Princeton University Press, S. 60–90.
- Guttman, Louis (1950b). »The Principal Components of Scale Analysis«. In: *Measurement and Prediction*. Hrsg. von Samuel A. Stouffer, Louis Guttman u. a. Bd. IV. Studies

- in social psychology in World War II. Princeton, NJ: Princeton University Press, S. 312–361.
- Guttman, Louis (1954a). »Some necessary conditions for common-factor analysis«. In: *Psychometrika* 19.2, S. 149–161.
- Guttman, Louis (1954b). »The Principal Components of Scalable Attitudes«. In: *Mathematical Thinking in the Social Sciences*. Hrsg. von Paul F. Lazarsfeld. Glencoe, IL: The Free Press.
- Hagenaars, Jacques A. (1993). *Loglinear Models with Latent Variables*. Cambridge: Cambridge University Press.
- Halpin, Brendan (2010). »Optimal Matching Analysis and Life-Course Data: The Importance of Duration«. In: *Sociological Methods & Research* 38.3, S. 365–388.
- Hamerle, Alfred und Heinz Pape (1984). »Grundlagen der mehrdimensionalen Skalierung«. In: *Multivariate statistische Verfahren*. Hrsg. von Ludwig Fahrmeir und Walter Häußler. Berlin und New York: de Gruyter, S. 663–688.
- Hamming, Richard W. (1950). »Error Detecting and Error Correcting Codes«. In: *The Bell System Technical Journal* 26.2, S. 147–160.
- Hanson, Robin, John Stutz und Peter Cheeseman (1991). *Bayesian Classification Theory*. Technical Report FIA-90-12-7-01. NASA Ames Research Center, Artificial Intelligence Branch.
- Hartigan, John A. (1975). *Clustering Algorithms*. New York, Chichester u. a.: Wiley.
- Hartigan, John A. und Surya Mohanty (1992). »The RUNT Test for Multimodality«. In: *Journal of Classification* 9.1, S. 63–70.
- Hartung, Joachim und Bärbel Elpelt (1984). *Multivariate Statistik. Lehr- und Handbuch der angewandten Statistik*. München und Wien: Oldenbourg.
- Hayduk, Leslie Alec (1987). *Structural Equation Modeling with LISREL*. Baltimore und London: Johns Hopkins University Press.
- Heckman, James und Burton Singer (1984). »A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data«. In: *Econometrica* 52.2, S. 271–320.
- Hilbe, Joseph M. (2007). *Negative Binomial Regression*. Cambridge: Cambridge University Press.
- Hojtink, Herbert (2001). »Confirmatory Latent Class Analysis: Model Selection Using Bayes Factors and (Pseudo) Likelihood Ratio Statistics«. In: *Multivariate Behavioral Research* 36.4, S. 563–588.
- Holm, Kurt (1975). »Die Erstellung einer Korrelationsmatrix«. In: *Die Befragung* 2. Hrsg. von Kurt Holm. München: Francke, S. 155–218.
- Holm, Kurt (1976). »Die Faktorenanalyse – ihre Anwendung auf Fragebatterien«. In: *Die Befragung* 3. Hrsg. von Kurt Holm. München: Francke, S. 11–268.

- Holm, Kurt (1979). »Das allgemeine lineare Modell«. In: *Die Befragung* 6. Hrsg. von Kurt Holm. München: Francke, S. 11–213.
- Holm, Kurt (1993). *ALMO-Statistiksystem*. Linz: Eigenverlag.
- Holman, Eric W. (1972). »The Relation between Hierarchical and Euclidean Models for Psychological Distances«. In: *Psychometrika* 37.4, S. 417–423.
- Holtmann, Dieter (1975). »Metrische multidimensionale Skalierung und ein ›inhaltliches‹ Verfahren zur Bestimmung der Achsen«. In: *Zeitschrift für Soziologie* 4.3, S. 248–253.
- Horn, John L. (1965). »A rationale and test for the number of factors in factor analysis«. In: *Psychometrika* 30.2, S. 179–185.
- Horst, Paul (1935). »Measuring Complex Attitudes«. In: *Journal of Social Psychology* 6.3, S. 369–374.
- Horst, Paul (1965). *Factor Analysis of Data Matrices*. New York, Chicago u. a.
- Hubert, Lawrence J. und Joel R. Levin (1977). »Inference Models for Categorical Clustering«. In: *Psychological Bulletin* 84.5, S. 878–887.
- Hubert, Lawrence (1974). »Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarchical Clustering Procedures«. In: *Journal of the American Statistical Association* 69.347, S. 698–704.
- Hubert, Lawrence und Phipps Arabie (1985). »Comparing Partitions«. In: *Journal of Classification* 2.1, S. 193–218.
- Inglehart, Ronald (1979). »Wertwandel in den westlichen Gesellschaften«. In: *Wertwandel und gesellschaftlicher Wandel*. Hrsg. von Helmut Klages und Peter Kmiciak. Frankfurt und New York: Campus-Verlag, S. 279–316.
- Jahnke, Hermann (1988). *Clusteranalyse als Verfahren der schließenden Statistik*. Göttingen: Vandenhoeck & Ruprecht.
- Jain, Anil K. und Richard C. Dubes (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.
- Jeffries, Neal O. (2003). »A note on ›Testing the number of components in a normal mixture‹«. In: *Biometrika* 90.4, S. 991–994.
- Jöreskog, Karl G. und Dag Sörbom (1984). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods*. 3. Aufl. Mooresville: Scientific Software.
- Kaiser, Henry F. (1970). »A Second Generation Little Jiffy«. In: *Psychometrika* 35.4, S. 401–415.
- Kaiser, Henry F. und Kern W. Dickman (1959). »Analytic determination of common factors«. In: *American Psychologist* 14.7. Abstract, S. 425.
- Kaiser, Henry F. und John Rice (1974). »Little Jiffy, Mark IV«. In: *Educational and Psychological Measurement* 34.1, S. 111–117.
- Kaufman, Leonard und Peter J. Rousseeuw (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. New York, Chichester u. a.: Wiley.

- Kaufman, Robert L. (1985). »Issues in Multivariate Cluster Analysis. Some Simulations Results«. In: *Sociological Methods & Research* 13.4, S. 467–486.
- Kaufmann, Heinz und Heinz Pape (1984). »Clusteranalyse«. In: *Multivariate statistische Verfahren*. Hrsg. von Ludwig Fahrmeir und Alfred Hamerle. Berlin und New York: de Gruyter, S. 371–472.
- Kaufmann, Heinz und Heinz Pape (1996). »Clusteranalyse«. In: *Multivariate statistische Verfahren*. Hrsg. von Ludwig Fahrmeir, Wolfgang Brachinger, Alfred Hamerle und Gerhard Tutz. Berlin und New York: de Gruyter, S. 437–536.
- Kendall, Maurice G. (1962). *Rank Correlation Methods*. 3. Aufl. London: Griffin.
- Kendall, Maurice G. (1980). *Multivariate Analysis*. 2. Aufl. London: Griffin.
- Kettenring, Jon R. (2006). »The Practice of Cluster Analysis«. In: *Journal of Classification* 23.1, S. 3–30.
- Kirchler, Erich und Renate Nagl (1993). »Der Freundeskreis und das Freizeitverhalten«. In: *Kindsein in Österreich*. Hrsg. von Liselotte Wilk und Johann Bacher. Unveröffentlichter Forschungsbericht, Institut für Soziologie an der Universität Linz, Kapitel 8.
- Klastorin, Theodore D. (1983). »Assessing Cluster Analysis Results«. In: *Journal of Marketing Research* 20.1, S. 92–98.
- Kluge, Susann und Udo Kelle, Hrsg. (2001). *Methodeninnovation in der Lebenslauforschung. Integration qualitativer und quantitativer Verfahren in der Lebenslauf- und Biographieforschung*. Weinheim und München: Juventa.
- Kozelka, Robert M. (1982). »How to Work Through a Clustering Problem«. In: *Classifying Social Data. New Applications of Analytic Methods for Social Science Research*. Hrsg. von Herschel C. Hudson. San Francisco, Washington u. a.: Jossey-Bass, S. 1–12.
- Kreutz, Henrik und Johann Bacher (1991). »Analysemethoden und Mikrosimulation: der Königsweg der pragmatischen Soziologie«. In: *Disziplin und Kreativität. Sozialwissenschaftliche Computersimulation: theoretische Experimente und praktische Anwendung*. Hrsg. von Henrik Kreutz und Johann Bacher. Opladen: Leske + Budrich, IX–XXIX.
- Kriz, Jürgen (1978). *Statistik in den Sozialwissenschaften*. Reinbek bei Hamburg: Rowohlt.
- Kruskal, Joseph B. (1964a). »Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis«. In: *Psychometrika* 29.1, S. 1–27.
- Kruskal, Joseph B. (1964b). »Nonmetric Multidimensional Scaling: A Numerical Method«. In: *Psychometrika* 29.2, S. 115–129.
- Kruskal, Joseph B. (1983). »An Overview of Sequence Comparison«. In: *Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison*. Hrsg. von David Sankoff und Joseph B. Kruskal. Reading: Addison-Wesley.

- Langeheine, Rolf und Frank Van de Pol (1990). »A Unifying Framework for Markov Modeling in Discrete Space and Discrete Time«. In: *Sociological Methods & Research* 18.4, S. 416–441.
- Lapointe, François-Joseph und Pierre Legendre (1995). »Comparison Tests for Dendograms: A Comparative Evaluation«. In: *Journal of Classification* 12.2, S. 265–282.
- Lathrop, Richard G. und Janice E. Williams (1987). »The Reliability of the Inverse Scree Test for Cluster Analysis«. In: *Educational and Psychological Measurement* 47.4, S. 953–959.
- Lathrop, Richard G. und Janice E. Williams (1989). »The Shape of the Inverse Scree Test for Cluster Analysis«. In: *Educational and Psychological Measurement* 49.4, S. 827–834.
- Lathrop, Richard G. und Janice E. Williams (1990). »The Validity of the Inverse Scree Test for Cluster Analysis«. In: *Educational and Psychological Measurement* 50.2, S. 325–330.
- Lazarsfeld, Paul F. (1966). »Latent Structure Analysis and Test Theory«. In: *Readings in Mathematical Social Science*. Hrsg. von Paul F. Lazarsfeld und Neil W. Henry. Chicago: Science Resaerch Associates, S. 78–88.
- Lazarsfeld, Paul F. und Neil W. Henry (1968). *Latent structure analysis*. Boston, New York u. a.: Houghton Mifflin.
- Lebart, Ludovic, Alain K. Morineau und Kenneth M. Warwick (1984). *Multivariate Descriptive Statistical Analysis*. New York, Chichester u. a.: Wiley.
- Lesnard, Laurent (2006). *Optimal Matching and Social Sciences*. Working Paper. URL: <http://www.crest.fr/doctravail/document/2006-01.pdf> (besucht am 25. 03. 2010).
- Levenshtein, Vladimir I. (1966). »Binary Codes Capable of Correcting Deletions, Insertions ans Reversals«. In: *Cybernetics and Control Theory* 10.8, S. 707–710.
- Levine, Joel H. (2000). »But What Have You Done for Us Lately? Commentary on Abbott and Tsay«. In: *Sociological Methods & Research* 29.1, S. 34–40.
- Lienert, Gustav A. (1973). *Verteilungsfreie Methoden in der Biostatistik*. 2. Aufl. Meisenheim am Glan: Hain.
- Lienert, Gustav A. (1975). *Verteilungsfreie Methoden in der Biostatistik. Tafelband*. Meisenheim am Glan: Hain.
- Little, Roderick J. A. und Donald B. Rubin (2002). *Statistical Analysis with Missing Data*. 2. Aufl. Hoboken, nj: Wiley-Interscience.
- Lo, Yungtai, Nancy R. Mendell und Donald B. Rubin (2001). »Testing the Number of Components in a Normal Mixture«. In: *Biometrika* 88.3, S. 767–778.
- Lohse, Heinz, Rolf Ludwig und Michael Röhr (1982). *Statistische Verfahren für Psychologen, Pädagogen und Soziologen*. Berlin: Verlag Volk und Wissen.
- Lorr, Maurice und Belur K. Radhakrishnan (1967). »A Comparision of two Methods of Cluster Analysis«. In: *Educational and Psychological Measurement* 27.1, S. 47–53.

- Lüdtke, Hartmut (1989). *Expressive Ungleichheit. Zur Soziologie der Lebensstile*. Opladen: Leske + Budrich.
- Lund, Thorleif (1974). »Multidimensional Scaling of Political Parties«. In: *Scandinavian Journal of Psychology* 15.1, S. 108–118.
- MacIndoe, Heather und Andrew Abbott (2009). »Sequence Analysis and Optimal Matching. Techniques for Social Science Data«. In: *The Handbook of Data Analysis*. Hrsg. von Melissa A. Hardy und Alan Bryman. London: Sage Publications Ltd., S. 387–406.
- MacQueen, James B. (1967). »Some Methods for Classification and Analysis of Multivariate Observations«. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Hrsg. von Lucien M. Le Cam und Jerzy Neyman. Bd. 1: Statistics. Berkeley, CA: University of California Press, S. 281–297.
- Maechler, Martin, Peter Rousseeuw, Anja Struyf und Mia Hubert (2005). »Cluster Analysis Basics and Extensions«.
- Magidson, Jay und Jeroen Vermunt (2004). »Latent Class Models«. In: *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, London u. a.: SAGE Publications, S. 175–198.
- Mahalanobis, Prasanta C. (1936). »On the generalised distance in statistics«. In: *Proceedings of the National Institute of Science of India*. Hrsg. von National Institute of Science of India. Bd. 12, S. 49–55.
- Mann, Henry B. und Donald R. Whitney (1947). »On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other«. In: *The Annals of Mathematical Statistics* 18.1, S. 50–60.
- Martin, Peter, Ingrid Schoon und Andy Ross (2008). »Beyond Transitions: Applying Optimal Matching Analysis to Life Course Research«. In: *International Journal of Social Research Methodology* 11.3, S. 179–199.
- Massey Jr., Frank J. (1951). »The Kolmogorov-Smirnov Test for Goodness of Fit«. In: *Journal of the American Statistical Association* 46.253, S. 68–78.
- McKeown, Bruce und Dan Thomas (1988). *Q Methodology*. London: Sage Publications.
- McLachlan, Geoffrey J. und Kaye E. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.
- McLachlan, Geoffrey J. und David Peel (2000). *Finite mixture models*. Bd. 299. Probability and Statistics – Applied Probability and Statistics Section. New York: Wiley.
- McNemar, Quinn (1947). »Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages«. In: *Psychometrika* 12.2, S. 153–157.
- Miller, Rupert G. (1981). *Simultaneous Statistical Inference*. 2. Aufl. New York: Springer.
- Milligan, Glenn W. (1979). »Ultrametric Hierarchical Clustering Algorithms«. In: *Psychometrika* 44.3, S. 343–346.

- Milligan, Glenn W. (1980). »An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms«. In: *Psychometrika* 45.3, S. 325–342.
- Milligan, Glenn W. (1981). »A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis«. In: *Psychometrika* 46.2, S. 187–199.
- Milligan, Glenn W. und Martha C. Cooper (1987). »Methodology Review: Clustering Methods«. In: *Applied Psychological Measurement* 11.4, S. 329–354.
- Mills, Melinda (2010). *Introducing Survival and Event History Analysis*. im Erscheinen. London: Sage.
- Mirkin, Boris (1996). *Mathematical Classification and Clustering*. Dordrecht: Kluwer Academic Publishers.
- Mojena, Richard (1977). »Hierarchical Grouping Methods and Stopping Rules: An Evaluation«. In: *Computer Journal* 20.4, S. 359–363.
- Mokken, Robert J. (1971). *A theory and procedure of scale analysis. With applications in political research*. The Hague: Mouton.
- Morey, Leslie C. und Alan Agresti (1984). »The Measurement of Classification Agreement: An Adjustment to the Rand Statistics for Chance Agreement«. In: *Educational and Psychological Measurement* 44.1, S. 33–37.
- Münchmeier, Richard (1997). »Jung – und ansonsten ganz verschieden«. In: *Jugend '97 Zukunftsperspektiven. Gesellschaftliches Engagement. Politische Orientierungen*. Hrsg. von Jugendwerk der Deutschen Shell. Opladen: Leske + Budrich, S. 379–390.
- Muthén, Bengt O. (2004). »Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data«. In: *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, London u. a.: SAGE Publications, S. 345–370.
- Needleman, Saul B. und Christian D. Wunsch (1970). »A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins«. In: *Journal of Molecular Biology* 48.3, S. 443–453.
- Neumann, Jörg, Wolfgang Frindte, Friedrich Funke und Susanne Jacob (1999). »Sozial-psychologische Hintergründe von Fremdenfeindlichkeit und Rechtsextremismus«. In: *Rechtsextremismus und Fremdenfeindlichkeit. Bestandsaufnahme und Interventionsstrategien*. Hrsg. von Frieder Dünkel und Bernd Geng. Mönchengladbach: Forum-Verlag Godesberg, S. 111–138.
- Nohlen, Dieter, Hrsg. (1984). *Lexikon Dritte Welt*. Reinbek bei Hamburg: Rowohlt.
- Nylund, Karen L., Tihomir Asparouhov und Bengt O. Muthén (2007). »Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study«. In: *Structural Equation Modeling* 14.4, S. 535–569.
- Opitz, Otto (1980). *Numerische Taxonomie*. Stuttgart und New York: Fischer.

- Opitz, Otto und Raimund Wiedemann (1989). »An Agglomerative Algorithm of Overlapping Clustering«. In: *Conceptual and Numerical Analysis of Data*. Hrsg. von Otto Opitz. Berlin, Heidelberg u. a.: Springer, S. 201–211.
- Ost, Friedmann (1984). »Faktorenanalyse«. In: *Multivariate statistische Verfahren*. Hrsg. von Ludwig Fahrmeir und Alfred Hamerle. Berlin und New York: de Gruyter, S. 575–632.
- Overall, John E. und Douglas K. Spiegel (1969). »Concerning Least Squares Analysis of Experimental Data«. In: *Psychological Bulletin* 72.5, S. 311–322.
- Pearson, Karl (1900). »On the Criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling«. In: *Philosophical Magazine Series 5* 50.302, S. 157–175.
- Pearson, Karl (1904). *On the Theory of Contingency and its Relation to Association and Normal Correlation*. Bd. 1K. Biometric Series. London: Drapers' Co. Memoirs.
- Peck, Laura R. (2005). »Using Cluster Analysis in Program Evaluation«. In: *Evaluation Review* 29.2, S. 178–196.
- Pöge, Andreas (2007). *Soziale Milieus und Kriminalität im Jugendalter. Eine Untersuchung von Werte- und Musiktypologien in Münster und Duisburg*. Zugl. Dissertation an der Universität Trier, 2007. Münster, New York u. a.: Waxmann.
- Pollard, David (1981). »Strong Consistency of k -Means Clustering«. In: *Annals of Statistics* 9.9, S. 135–140.
- Pollard, David (1982). »A Central Limit Theorem for k -Means Clustering«. In: *Annals of Probability* 10.4, S. 919–926.
- Pruzansky, Sandra, Amos Tversky und J. Douglas Carroll (1982). »Spatial versus Tree Representations of Proximity Data«. In: *Psychometrika* 47.1, S. 3–24.
- Punj, Girish und David W. Stewart (1983). »Cluster Analysis in Marketing Research: Review and Suggestions for Applications«. In: *Journal of Marketing Research* 20.2, S. 134–148.
- Puri, Madan L. und Pranab K. Sen (1971). *Nonparametric Methods in Multivariate Analysis*. New York, London u. a.: Wiley.
- Rand, William M. (1971). »Objective Criteria for the Evaluation of Clustering Methods«. In: *Journal of the American Statistical Association* 66.336, S. 846–850.
- Rasch, Georg (1960). *Probabilistic models for some intelligence and attainment tests*. København: Danmarks pædagogiske Institut.
- Reinecke, Jost (2005). *Strukturgleichungsmodelle in den Sozialwissenschaften*. München und Wien: Oldenbourg.
- Rigdon, Steven E. und Robert K. Tsutakawa (1983). »Parameter Estimation in Latent Trait Models«. In: *Psychometrika* 48.4, S. 567–574.

- Rinn, John L. (1961). »Q Methodology: An Application to Group Phenomena«. In: *Educational and Psychological Measurement* 22.2, S. 315–330.
- Rohwer, Götz und Ulrich Pötter (2005). *TDA User's Manual*. Unveröffentlichtes Manuskript. URL: <http://www.stat.ruhr-uni-bochum.de/tman.html>.
- Rost, Jürgen (2004). *Lehrbuch Testtheorie – Testkonstruktion*. 2. Aufl. Bern, Göttingen u. a.: Hans Huber.
- Sackmann, Reinhold und Matthias Wingens (2001). »Theoretische Konzepte des Lebenslaufs: Übergang, Sequenz und Verlauf«. In: *Strukturen des Lebenslaufs. Übergang – Sequenz – Verlauf*. Hrsg. von Reinhold Sackmann und Matthias Wingens. Weinheim und München: Juventa, S. 17–48.
- Saint-Arnaud, Sébastien und Paul Bernard (2003). »Convergence or Resilience? A Hierarchical Cluster Analysis of the Welfare Regimes in Advanced Countries«. In: *Current Sociology* 51.5, S. 499–527.
- Sankoff, David und Joseph B. Kruskal, Hrsg. (1983). *Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison*. Reading: Addison-Wesley.
- SAS Institute (1991). *SAS/STAT User's Guide*. Cary: SAS Institute Inc.
- Sattath, Shmuel und Amos Tversky (1977). »Additive Similarity Trees«. In: *Psychometrika* 42.3, S. 319–345.
- Sawrey, William L., Leo Keller und John J. Conger (1960). »An Objective Method of Grouping Profiles by Distance Functions and its Relations to Factor Analysis«. In: *Educational and Psychological Measurement* 20.4, S. 651–673.
- Schafer, Joseph L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman und Hall.
- Schendera, Christian F. G. (2010). *Clusteranalyse mit SPSS*. München: Oldenbourg.
- Schlosser, Otto (1976). *Einführung in die sozialwissenschaftliche Zusammenhangsanalyse*. Reinbek bei Hamburg: Rowohlt.
- Schmierer, Christian (1975). »Tabellenanalyse«. In: *Die Befragung* 2. Hrsg. von Kurt Holm. München: Francke, S. 86–137.
- Schmitt, Neal (1996). »Uses and Abuses of Coefficient Alpha«. In: *Psychological Assessment* 8.4, S. 350–353.
- Schwarz, Gideon (1978). »Estimating the Dimension of a Model«. In: *Annals of Statistics* 6.2, S. 461–464.
- Shepard, Roger N. und Phipps Arabie (1979). »Additive Clustering: Representation of Similarities as Combination of Discrete Overlapping Properties«. In: *Psychological Review* 86.2, S. 87–123.
- Siegel, Sidney (1976). *Nichtparametrische statistische Methoden*. Frankfurt a. M.: Fachbuchhandlung für Psychologie, Verl.-Abt.

- Sixtl, Friedrich (1982). *Meßmethoden der Psychologie. Theoretische Grundlagen und Probleme*. 2. Aufl. Weinheim und Basel: Beltz.
- Sneath, Peter H. A. und Robert R. Sokal (1973). *Numerical Taxonomy*. San Francisco: Freeman.
- Sodeur, Wolfgang (1974). *Empirische Verfahren zur Klassifikation*. Stuttgart: Teubner.
- Spence, Ian und Dennis W. Domoney (1974). »Single Subject Incomplete Design for Nonmetric Multidimensional Scaling«. In: *Psychometrika* 39.4, S. 469–490.
- SPSS Inc. (1990a). *SPSS/PC+ Advanced Statistics*. Chicago und Gornichen: SPSS Inc.
- SPSS Inc. (1990b). *SPSS/PC+ Categories*. Chicago und Gornichen: SPSS Inc.
- SPSS Inc. (2000). *The SPSS TwoStep cluster component. A scalable component to segment your customers more effectively*. Chicago und Gornichen: SPSS Inc. URL: <ftp://ftp.spss.com/pub/web/wp/TSCWP-1100.pdf>.
- SPSS Inc. (2008). *SPSS Statistics 17.0 Algorithms*. Chicago und Gornichen: SPSS Inc.
- Steinhausen, Detlef und Klaus Langer (1977). *Clusteranalyse. Einführung in Methoden und Verfahren der automatischen Klassifikation*. Berlin und New York: de Gruyter.
- Steinley, Douglas (2004). »Properties of the Hubert–Arabie Adjusted Rand Index«. In: *Psychological Methods* 9.3, S. 386–396.
- Steinley, Douglas und Michael J. Brusco (2007). »Initializing K-Means Batch clustering: A Critical Evaluation of Several Techniques«. In: *Journal of Classification* 24.1, S. 99–121.
- Stephenson, William (1959). *The Study of Behavior: Q-Technique and Its Methodology*. Chicago: University of Chicago Press.
- Student (1908). »The Probable Error of a Mean«. In: *Biometrika* 1, S. 1–25.
- Tenenhaus, Michel und Forrest W. Young (1985). »An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and other Methods for Quantifying Categorical Multivariate Data«. In: *Psychometrika* 50.1, S. 91–119.
- Trauwaert, Etienne (1988). »On the meaning of Dunn's partition coefficient for fuzzy clusters«. In: *Fuzzy Sets and Systems* 25.2, S. 217–242.
- Tryon, Robert C. (1958). »Cumulative Communality Cluster Analysis«. In: *Educational and Psychological Measurement* 18.1, S. 3–35.
- Tucker, Ledyard R. und Samuel J. Messick (1963). »Individual Difference Model for Multidimensional Scaling«. In: *Psychometrika* 28.4, S. 333–367.
- Van de Pol, Frank und Jan de Leeuw (1986). »A Latent Markov Model to Correct Measurement Error in Categorical Data«. In: *Sociological Methods & Research* 15.1–2, S. 118–141.
- Van de Pol, Frank, Rolf Langeheine und Wil de Jong (1989). *Panmark User Manual*. Voorburg: Netherlands Central Bureau of Statistics.

- Vermunt, Jeroen K. und Jay Magidson (2000). *Latent GOLD User's Guide*. Boston: Statistical Innovations Inc.
- Vermunt, Jeroen K. und Jay Magidson (2002). »Latent Class Cluster Analysis«. In: *Applied Latent Class Analysis*. Hrsg. von Jacques A. Hagenaars und Allan L. McCutcheon. Cambridge: Cambridge University Press, S. 89–106.
- Vermunt, Jeroen K. und Jay Magidson (2005a). *Latent GOLD 4.0 User's Guide*. Belmont: Statistical Innovations Inc.
- Vermunt, Jeroen K. und Jay Magidson (2005b). *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Belmont, MA: Statistical Innovations Inc.
- Vogel, Friedrich (1975). *Probleme und Verfahren der numerischen Klassifikation*. Göttingen: Vandenhoeck & Ruprecht.
- Weller, Susan C. und Antone Kimball Romney (1990). *Metric Scaling. Correspondence Analysis*. Newbury Park, London u. a.: Sage Publications.
- Wilcoxon, Frank (1945). »Individual Comparisons by Ranking Methods«. In: *Biometrics Bulletin* 1.6, S. 80–83.
- Wilcoxon, Frank (1947). »Probability Tables for Individual Comparisons by Ranking Methods«. In: *Biometrics* 3.3, S. 119–122.
- Wilk, Liselotte und Johann Bacher, Hrsg. (1993). *Kindsein in Österreich*. Unveröffentlichter Forschungsbericht, Institut für Soziologie an der Universität Linz. Linz.
- Windham, Michael P. (1986). »A Unification of Optimization-Based Numerical Classification Algorithms«. In: *Classification as a Tool of Research*. Hrsg. von Wolfgang Gaul und Martin Schader. Amsterdam: North-Holland, S. 447–452.
- Wishart, David (2003). »k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values«. In: *Exploratory Data Analysis in Empirical Research. Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation e. V., University of Munich, March 14–16, 2001*. Berlin, Heidelberg u. a.: Springer, S. 212–226.
- Wolfe, John H. (1970). »Pattern Clustering by Multivariate Mixture Analysis«. In: *Multivariate Behavioral Research* 5.3, S. 329–350.
- Wolff, Hans-Georg und Johann Bacher (2008). »Dimensionale Analyse multidimensionaler Items«. In: *Klassifikationsanalysen in Theorie und Praxis*. Hrsg. von Jost Reinecke und Christian Tarnai. Münster, New York u. a.: Waxmann, S. 19–41.
- Wolff, Hans-Georg und Johann Bacher (2010). »Hauptkomponentenanalyse und explorative Faktorenanalyse«. In: *Handbuch sozialwissenschaftliche Datenanalyse*. Hrsg. von Christoph Wolff und Hening Best. im Druck. Wiesbaden: vs.
- Wu, C. F. Jeff (1983). »On the Convergence of the EM Algorithm«. In: *The Annals of Statistics* 11.1, S. 95–103.
- Wu, Lawrence L. (2000). »Some Comments on Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect«. In: *Sociological Methods & Research* 29.1, S. 41–64.

- Young, Forrest W. (1970). »Nonmetric Multidimensional Scaling: Recovery of Metric Information«. In: *Psychometrika* 35.4, S. 455–473.
- Zegers, Frits E. und Jos M. F. ten Berge (1985). »A Family of Association Coefficients for Metric Scales«. In: *Psychometrika* 50.1, S. 17–24.
- Zhang, Tian, Raghu Ramakrishnan und Miron Livny (1996). »BIRCH: an efficient data clustering method for very large databases«. In: *SIGMOD Record* 25.2, S. 103–114.
- Zhang, Tian, Raghu Ramakrishnan und Miron Livny (1997). »BIRCH: A New Data Clustering Algorithm and Its Applications«. In: *Data Mining and Knowledge Discovery* 1.2, S. 141–182.

Register

A

- Abfall PRE-Koeffizient, 310
- abgeleitete
 - ~ Variablen, 459
- abgeleitetes
 - ~ (Un-)Ähnlichkeitsmaß, 195
- abhängige Stichproben
 - t-Test ~, 318
- additatives Clustermodell, 142
- Additivität
 - Koeffizient der ~, 227
- adjustierter RAND-Index, 274
- Ä, Ähnlichkeitsmatrix ~, 80, 142
- Ähnlichkeitsmaß
 - ~ Korrelationskoeffizient, 195
- Ähnlichkeitsmaße
 - ~ dichotome Variablen, 200
- Ähnlichkeitsmatrix, 41
 - ~ Ä, 80, 142
 - symmetrische ~, 143
- Ähnlichkeitsstrukturanalyse, 37
- Ähnlichkeitsurteilen
 - Befragung von ~, 90
- AIC, 364, 366, 384, 419, 420, 425, 448, 456
- AIC₃, 384, 419, 420, 425
- Akaike Information Criterion
 - Consistent ~, 364
- Akaike Informationsmaß, 364
- Algorithmus
 - EM ~, 352
 - Konvergenzverhalten K-Means ~, 301
 - Needleman-Wunsch ~, 482
 - TwoStep-Cluster ~, 446
 - Veranschaulichung K-Means ~, 304
- Alienation-Koeffizient, 93

allgemeine

- ~ latente Klassenanalyse, 392
- alternative Startwertverfahren, 335
- Analyse
 - ~ Ausreißer, 325
 - ~ latenter Klassen, 351, 352
- analysis
 - nonmetrical multidimensional ~, 37
- anderes
 - ~ (Un-)Ähnlichkeitsmaß, 195
- Annahme
 - ~ lokale Unabhängigkeit, 372
- Anpassungsindex
 - ~ der mittleren Residuenabweichungen, 46
 - ~ für den STRESS, 93
- Ansatz
 - Latent-GOLD ~, 21
- Anteil
 - ~ Klassifikationsfehler, 421
- Anteilswerten
 - lineare Restriktion von ~, 373
- Anzahl
 - ~ bedeutsame Dimensionen, 68, 128
 - Dimensionen, Überschätzung ~, 129
 - ~ maximale Dimensionen, 87, 113
 - ~ mögliche Dimensionen, 68
- A-posteriori-Wahrscheinlichkeit, 439
- Appropriate Scaling Method, 37
- Approximation
 - CS ~, 444
- A-priori-Wahrscheinlichkeit, 439
- Assoziationsmaße
 - nominale ~, 207
 - ordinale ~, 211, 212

- asymptotisches Verhalten
 - ~ K-Means-Verfahren, 301
 - Auftrittswahrscheinlichkeiten
 - bedingte ~, 378
 - Ausprägungen
 - Gewichtung ~, 227
 - Ausreißer, 168, 323
 - Analyse ~, 325
 - ~ Behandlung, 327
 - Ausscheiden
 - fallweises ~, 228
 - paarweises ~, 228
 - Auswahl
 - ~ bedeutsame Dimensionen, 41
 - ~ erforderliche Dimensionen, 41
 - ~ Faktanzahl, 123
 - Präferenzdaten fester ~, 134
 - ~ (Un-)Ähnlichkeitsmaß, 41, 195
 - AutoClass
 - Modellansatz ~, 441
 - automatische
 - ~ Orthogonalisierung, 193
 - Average-Linkage, 150, 264
 - AWE, 422
- B**
- BACKER, 369
 - Bartlett
 - ~ Test, 122
 - Bayes
 - ~ Modellierung, Vorteile, 440
 - ~ Schätzer, 352
 - ~ Theorem, 439
 - Bayesian Information Criterion, 364
 - Bealsche
 - ~ F-Werte, 308, 311
 - Bealsche F-Werte
 - Signifikanztest ~, 308
 - bedingte
 - ~ Auftrittswahrscheinlichkeiten, 378
 - ~ Variablen, 175
 - Befragte
 - durchschnittliche ~, 291
 - Befragung
 - ~ von Ähnlichkeitsurteilen, 90
- C**
- Behandlung
 - Ausreißer ~, 327
 - Berechnung
 - ~ Distanzmatrix, 42
 - ~ Eigenvektoren, 57
 - ~ Eigenwerte, 57, 61
 - ~ empirische Zusammenhangsmatrix, 57, 58
 - ~ Faktorladungen, 61
 - ~ Faktorwerte, 125, 141
 - ~ Freiheitsgrade, 68
 - ~ Koordinatenwerte, 57, 61, 110
 - ~ Skalenwerte, 63, 64
 - ~ (Un-)Ähnlichkeitsmatrix, 41
 - Bestimmung
 - ~ Clusterzahl, 305
 - ~ Klassenzahl, 362
 - ~ Koordinatenwerte, 80
 - ~ maximale Dimensionszahl, 41
 - Beurteilung
 - ~ STRESS, 92
 - Bevorzugung
 - ~ euklidische Distanz, 92
 - BIC, 364, 384, 419, 420, 422, 425, 448, 449, 456
 - Bildung
 - ~ Cluster, 140
 - Bindungen, 287, 290
 - Binomial-Verteilung, 409
 - verallgemeinerte ~, 210
 - bivariate
 - ~ Korrespondenzanalyse, 38, 109
 - ~ Residuen, 424
 - Block-Clustering-Methode, 371
 - Bootstrap
 - ~ LQ-Test, 423
 - parametrischer ~, 422

- Chebychev
 ~ Distanz, 220
 ~ Metrik, 470
- Chebychevsche
 ~ Ungleichung, 223
- χ^2
 ~ Beitrag, 113
 ~ Prüfgröße, 67
 ~ Test, 67, 68, 224
 ~ für Likelihood-Quotienten-Teststatistik, 364
 ~ Unabhängigkeitstest, 318
 ~ Wert, 67
- City-Block-Metrik, 82, 197, 209, 214, 220
 Signifikanztest ~, 222
- Cliquen, 149
- Cluster
 Bildung ~, 140
 Homogenität innerhalb ~, 16
 inhaltliche Interpretation ~, 41
 natürliche ~, 16
 Validitätsprüfung ~, 41
- Cluster Analysis
 Cumulative Communalität ~, 282
- Clusteranalyse
 Durchführung hierarchische ~, 42
 explorative ~, 22
 konfirmatorische ~, 22, 341
 objektorientierte ~, 21
 variablenorientierte ~, 22
- Clusteranalyseverfahren
 deterministische ~, 19, 39, 147
 disjunktive ~, 147
 hierarchische ~, 19
 probabilistische ~, 18, 19, 351
 überlappende ~, 20
 unvollständige ~, 18, 37, 38
- clusteranalytische
 ~ Fragestellung, 44
 ~ Interpretation, 41, 72, 111
- Clusterkern, 277
- Clusterlösung
 Homogenitätsindex ~, 248
 Korrelationsmaße ~, 248
 Robustheit einer ~, 170, 196
- Zufallstestung ~, 313
- Clustermittelwerte
 Konsensus für ~, 497
- Clustermodell
 additatives ~, 142
- Clustern
 Heterogenität zwischen ~, 16
- Clusterzahl
 Bestimmung ~, 305
 Stabilitätsprüfung ~, 328
- Clusterzentren, 285
 Konstruktion ~, 285
 Stabilitätsprüfung ~, 328
- Clusterzuordnungen
 Konsensus für ~, 498
- Complete-Linkage, 149, 233
 ~ für überlappende Cluster, 255
- Consensus-Tree, 497
- Consistent Akaike Information Criterion, 364
- Cressie-Reed- χ^2
 ~ Statistik, 417
- Criterion
 Bayesian Information ~, 364
 Consistent Akaike Information ~, 364
- cs
 ~ Approximation, 444
 ~ Kriterium, 444
- Cumulative Communalität Cluster Analysis, 282
- D**
- Darstellung
 ~ Idealpunkt, 105
 ~ kanonische Skalierung, 113
 ~ Spaltenskalierung, 115
 ~ (Un-)Ähnlichkeitsmatrix, 77
- Daten
 ipsative ~, 134
- Datenanalyse
 objektorientierte ~, 16, 153
 variablenorientierte ~, 16, 153
 Vorgehen ~, 40
- Datenmatrix Z , standardisierte ~, 137
- Datenmatrix ZQ , transponierte ~, 137

- Definition ~ mittlere Gesamtpunktwerte, 126
- Dendogramm, 497
- Deskriptionsvariablen inaktive ~, 332
- Determinante Minimierung ~, 338
- Determinantenkriterium, 338
- deterministische ~ Clusteranalyseverfahren, 19, 39, 147
- df, 67
- dichotome ~ Variablen, 122, 127, 197
~ Ähnlichkeitsmaße, 200
- quadrierte euklidische Distanz für ~, 198
- Dichtefunktion φ , 357
- Dilatationseffekt, 152
- Dimensionen Anzahl bedeutsame ~, 68, 128
Anzahl maximale ~, 87, 113
Anzahl mögliche ~, 68
- Auswahl bedeutsame ~, 41
- Auswahl erforderliche ~, 41
- inhaltliche Interpretation ~, 41, 118
- latente ~, 118, 120
- Überschätzung Anzahl ~, 129
- Dimensionszahl, 67
- Bestimmung maximale ~, 41
- Richtwert maximale ~, 87
- disjunktive ~ Clusteranalyseverfahren, 147
- Distanz Bevorzugung euklidische ~, 92
Chebychev ~, 220
euklidische ~, 82, 220
Signifikanztest ~, 221
- Hamming ~, 479, 480
- Levenshtein ~, 479, 481
- Mahalanobis ~, 339
- quadrierte euklidische ~, 73, 197, 220, 289
- Signifikanztest ~, 221
- standardisierte ~, 323
- Distanzen
- gleichmäßige Gewichtung ~, 226
- Distanzmaß falsches ~, 168, 170
Unähnlichkeitmaß ~, 195
- Distanzmaße Wahrscheinlichkeitsverteilungen ~, 224
- Distanzmatrix Berechnung ~, 42
- divisive ~ Verfahren, 19
- Dual Scaling, 37
- Dummies, 185, 208
- DUNN_K, 369
- Durchführung ~ hierarchische Clusteranalyse, 42
- durchschnittliche ~ Befragte, 291
~ Unähnlichkeit, 249
- E**
- Eigenvektoren Berechnung ~, 57
Reskalierung ~, 57, 61, 110
- Eigenwertabfall, 120, 123, 129
- Eigenwerte Berechnung ~, 57, 61
- Eigenwertzerlegung, 61, 110, 118, 123
- Einfachstruktur, 120, 124, 125, 127
- Prüfung ~, 97
- Einfluss ~ Kategorisierung, 231
~ Schwierigkeitsgrad Faktorenanalyse, 129
- Eingabe ~ Startwerte, 97
- EM-Algorithmus, 352
- Grundprinzip ~, 359
- empirische ~ Mittelwertzentrierung, 188
~ Skalenwerte, 178
- empirische Zusammenhangsmatrix G , 58
- erklärte ~ Streuung η^2 , 306, 321
~ Streuung, modifizierte, $\tilde{\eta}^2$, 340

- ~ Varianz, 417
- Erwartungswert
 - ~ η^2 , 313
- E-Schritt, 359
- $\tilde{\eta}^2$
 - erklärte Streuung ~, 306, 321
 - Erwartungswert ~, 313
 - modifizierte erklärte Streuung ~, 340
 - Schwellenwerte ~, 310
- euklidische Distanz, 82, 220
 - quadrierte ~, 73, 197, 220, 289
 - Signifikanztest ~, 221
 - Signifikanztest ~, 221
- Expectation-Maximization
 - ~ Verfahren, 414
- explorative
 - ~ Clusteranalyse, 22
- externe
 - ~ Isolierung, 16
- Extremwertnormalisierung, 177
- F**
- Faktoranzahl
 - Auswahl ~, 123
- Faktoren
 - Rotation ~, 66, 118, 124, 138
- Faktorenanalyse, 37–39, 122
 - ~ dichotome Variablen, 127
 - Einfluss Schwierigkeitsgrad ~, 129
 - nominale ~, 38, 117
 - ~ ordinale Variablen, 127
- faktorenanalytische
 - ~ Fragestellung, 44
 - ~ Interpretation, 41, 46
 - Interpretation, Überprüfung ~, 69
- Faktorisierung, 122
- Faktorladungen, 123
 - Berechnung ~, 61
 - Nachteil Reskalierung ~, 65
 - Reskalierung ~, 61
- Faktorladungsmatrix L , 124
- Faktorwerte, 127, 138, 140
 - Berechnung ~, 125, 141
- fallweises
 - ~ Ausscheiden, 228
- falsches
 - ~ Distanzmaß, 168, 170
- fehlende
 - ~ Werte, 228
 - Schätzwerte für ~, 228
- Fehleranalyse, 28
- Fehlerstreuung, 299
- Festsetzung
 - ~ Transformationskosten, 485
- F_{\max}
 - ~ Statistik, 307
 - ~ Wert, maximaler, 311
- Forgys Methode, 300
- formale
 - ~ Gültigkeitsprüfung, 493
 - ~ Validitätsprüfung, 27
- Fragestellung
 - clusteranalytische ~, 44
 - faktorenanalytische ~, 44
 - ~ Spaltenskalierung, 113
- Freiheitsgrade, 67
 - Berechnung ~, 68
 - Verhältnis ~, 87
- F-Statistik
 - maximale ~, 307
- Funktion
 - Likelihood ~, 356
 - Log-Likelihood ~, 356, 358
 - marginale Likelihood ~, 444
- Fuzzy-Clustering, 369
- F-Wert, 307, 321
 - Signifikanztest ~, 321
- F-Werte
 - Bealsche ~, 308, 311
 - Signifikanztest ~, 308
- G**
- G , empirische Zusammenhangsmatrix ~, 58
- G , Homogenitätsindex ~, 249
 - Signifikanztest ~, 249
- G , Teilmatrix Zusammenhangsmatrix ~, 109
- γ -Koeffizient, 162
- Schwellenwerte ~, 238

- Signifikanztest ~, 238
- gemischtes
 - ~ Messniveau, 336
- Gesamtpunktwerte
 - Definition mittlere ~, 126
 - mittlere ~, 125, 127
- Gesamtstreuungssquadratsumme
 - mittlere ~, 323
- Gewichtung
 - ~ Ausprägungen, 227
 - ~ durch Metrikparameter, 227
 - ~ durch Standardisierung, 226
- Gewinnen
 - ~ (Un-)Ähnlichkeitsmatrix, 78
- g-Faktorenmodell, 133
- GFID*-Index, 73
 - Schwellenwerte ~, 74
- GFIR*-Index, 46, 69
 - Schwellenwerte ~, 74
- GFIS-Index, 93
- Gleichheit
 - Koeffizient der ~, 227
- gleichmäßige Gewichtung
 - ~ Distanzen, 226
- Gradientenmethode, 85
- Gradientenverfahren, 37
- graphische Darstellung Ergebnisse, 41
- Grundprinzip
 - ~ EM-Algorithmus, 359
- Gruppe
 - homogene ~, 16
- Gültigkeitsprüfung
 - formale ~, 493
- Guttmanskala
 - klassische ~, 135
- H**
- Hamming-Distanz, 479, 480
- Hauptkomponenten
 - unrotierte ~, 193
- Hauptkomponentenanalyse, 38, 61, 122
- Hauptkomponentenskalierung, 111
- HETERO
 - Heterogenitätsindex ~, 322
- Heterogenität
- Schwellenwerte ~, 280
 - ~ zwischen Clustern, 16
- Heterogenitätsindex
 - ~ HETERO, 322
- heuristische
 - ~ Verfahren, 20
- hierarchische
 - ~ Clusteranalyseverfahren, 19
 - ~ Variablen, 183
- Hill-Climbing-Methode, 338
- HOMO
 - Homogenitätsindex ~, 318, 322
- homogene
 - ~ Gruppe, 16
- Homogenität, 16
 - ~ innerhalb Cluster, 16
 - Schwellenwerte ~, 280
- Homogenitätsindex
 - ~ Clusterlösung, 248
 - ~ G, 249
 - Signifikanztest ~, 249
 - ~ HOMO, 318, 322
- Homogeneity Analysis, 37
- I**
- ICL-BIC, 384, 425
- Idealpunkt
 - Darstellung ~, 105
 - Interpretation ~, 106
 - Koordinatenwert ~, 103
- Idealpunktrepräsentation, 102, 103
- Identifikation
 - ~ zu schätzendes Modell, 354
- imputierte
 - ~ Werte, 228
- inaktive
 - ~ Deskriptionsvariablen, 332
 - ~ Kovariaten, 396
- Indikatoren, 395
- inertia, 69
- Information Criterion
 - Bayesian ~, 364
 - Consistent Akaike ~, 364
- Informationsmaß
 - Akaike ~, 364

- inhaltliche
 ~ Interpretation Cluster, 41
 ~ Interpretation Dimensionen, 41, 118
 ~ Interpretierbarkeit, 27
 ~ Validitätsprüfung, 332
- Inkommensurabilität, 175
- interne
 ~ Kohäsion, 16
- Interpretation
 clusteranalytische ~, 41, 72, 111
 faktorenanalytische ~, 41, 46
 ~ Idealpunkt, 106
 ~ Koordinatenwerte, 63, 66
 Modellprüfgrößen clusteranalytische
 ~, 72
 ~ Verschmelzungsniveau, 236
- Interpretationsziel
 ~ K-Means-Verfahren, 314
- Interpretierbarkeit
 inhaltliche ~, 27
- inverser
 ~ Scree-Test, 241, 290
- ipsative
 ~ Daten, 134
- irrelevante
 ~ Variablen, 168, 170
- Isolierung
 externe ~, 16
- Item-Response-Modell, 135
- iteratives reallokatives Sum-of-Squares-Verfahren, 301
- J**
- JACCARD-II-Koeffizient, 215
- JACCARD-I-Koeffizient, 201
 Signifikanztest ~, 205
- K**
- K, Kovarianzmatrix ~, 118
- Kaiserkriterium, 68, 123
- Kaiser-Meyer-Olkin
 ~ Kriterium, 122
- kanonische
 ~ Skalierung, 111–113
- κ , Übereinstimmungskoeffizient ~, 201
- Signifikanztest ~, 206
- κ^{nom} , Übereinstimmungskoeffizient ~, 209
- κ^{ord} , Übereinstimmungskoeffizient ~, 216
- Kategorisierung
 Einfluss ~, 231
- K-Cluster-Verfahren, 346
- Klassen
 Analyse latenter ~, 351, 352
 latente ~, 352, 395
- Klassenanalyse
 allgemeine latente ~, 392
 Verallgemeinerung latente ~, 352
- Klassenzahl
 Bestimmung ~, 362
- Klassenzentren
 lineare Restriktion ~, 374
- Klassenzugehörigkeitswahrscheinlichkeiten, 420
- Klassifikations
 ~ Log-Likelihood, 422
 ~ Statistik, 420
- Klassifikationsfehler
 Anteil ~, 421
- Klassifikationsmatrix C , 143
- Klassifizierung
 ~ von Verläufen, 475
- klassische
 ~ Guttmanskala, 135
- Klumpen, 149
- K-Means ~
 ~ Algorithmus
 ~ Konvergenzverhalten ~, 301
 ~ Veranschaulichung ~, 304
 ~ Verfahren, 150, 299
 asymptotisches Verhalten ~, 301
 Interpretationsziel ~, 314
 Verallgemeinerung ~, 351, 352
 z-Werte ~, 317
- K-Median-Verfahren, 345
- K-Medoids-Verfahren, 347
- K-Modus-Verfahren, 345, 346
- Koeffizient
 Alienation ~, 93
 ~ der Additivität, 227
 ~ der Gleichheit, 227

- ~ der Linearität, 227
- ~ der Profilähnlichkeit r_p , 223
- ~ der Proportionalität, 227
- $\gamma \sim$, 162
- JACCARD-I ~, 201
Signifikanztest ~, 205
- JACCARD-II ~, 215
- $\Phi \sim$, 197
Signifikanztest ~, 205
- PRE ~, 307
- Silhouette ~, 495
- Simple-Matching ~, 201
- Kohäsion
 - interne ~, 16
- Kolmogorov-Smirnov-Test, 224
- Kommensurabilität, 175
- Kommunalitäten, 123
 - ~ Schätzung, 118, 123
- konfirmatorische
 - ~ Clusteranalyse, 22, 341
- Konsensus, 497
 - ~ für Clustermittelwerte, 497
 - ~ für Clusterzuordnungen, 498
- Konstruktion
 - ~ Clusterzentren, 285
 - ~ von Clustern, Verfahren zur ~, 150
- Kontraktionseffekt, 152
- Konvergenzkriterium, 415
- Konvergenzverhalten
 - ~ K-Means Algorithmus, 301
- Koordinatenwert
 - ~ Idealpunkt, 103
- Koordinatenwerte
 - Berechnung ~, 57, 61, 110
 - Bestimmung ~, 80
 - Interpretation ~, 63, 66
- kophenetische
 - ~ Korrelation, 240
 - ~ Matrix, 238
- Korrelation
 - ~, 197
 - kophenetische ~, 240
 - $\Phi \sim$, 197
 - polychorische ~, 129
 - Produkt-Moment ~, 197, 219
- Signifikanztest ~, 221
- tetrachorische ~, 129
- Korrelationskoeffizient
 - Ähnlichkeitsmaß ~, 195
- Korrelationskoeffizienten
 - ordinale ~
 - Signifikanztest ~, 218
- Korrelationsmaße
 - ~ Clusterlösung, 248
- Korrelationsmatrix
 - ~ Q , 138
 - ~ R , 122
- Korrespondenzanalyse
 - bivariate ~, 38, 109
 - Modellprüfgrößen multiple ~, 67
 - multiple ~, 37, 43, 57, 77, 118
 - Trägheit bivariate ~, 69
 - Trägheit multiple ~, 69
 - Vorgehen bivariate ~, 109
- Korrespondenzanalyseskalierung
 - multiple ~, 111
- Kovarianzmatrix K , 118
- Kovariaten, 395
 - inaktive ~, 396
- Kriterium
 - CS ~, 444
 - Kaiser-Meyer-Olkin ~, 122
- Kruskalverfahren, 37
- L**
- L , Faktorladungsmatrix ~, 124
- Ladungsmatrix, 124
- latente
 - ~ Dimensionen, 118, 120
 - ~ Klassen, 352, 395
 - ~ Profilanalyse, 351, 355
- Latent-GOLD-Ansatz, 21
- LCA, 351
- Leading
 - ~ Case, 277
- Levenshtein-Distanz, 479, 481
- Likelihood
 - ~ Funktion, 356
 - ~ Funktion, marginale, 444
- Likelihood-Quotienten-Teststatistik

- χ^2 Test für ~, 364
Likelihood-Ratio- χ^2
~ Statistik, 417
lineare Restriktion
~ Klassenzentren, 374
~ Variablenausprägung, 387
~ von Anteilswerten, 373
Linearität
Koeffizient der ~, 227
Linkage
Average ~, 150, 264
Complete ~, 149, 233
~ für überlappende Cluster, 255
Single ~, 149, 251, 326
Verkettungseffekt Single ~, 252
Weighted-Average ~, 150, 264
Within-Average ~, 150, 265
Log-Likelihood, 420
~ Funktion, 356, 358
Klassifikations ~, 422
~ Statistik, 420
Log-Posterior, 420
Log-Prior, 420
lokale Unabhängigkeit, 352
Lo-Mendell-Rubin-Likelihood-Ratio-Test,
422
LQ-Test
Bootstrap ~, 423
 LQ_K (Wolfe), 365
- M**
Mahalanobis-Distanz, 339
Mann-Whitney-Test, 318
marginale
Likelihood-Funktion, 444
Maßzahlen
 R^2 ~, 421
varianzanalytische ~, 366
Matrix
kophenetische ~, 238
maximale
~ F-Statistik, 307
maximaler F_{\max} Wert, 311
Maximierung
~ Spur, 338
- Maximum-a-posteriori-Schätzer, 443
Maximum-Likelihood
~ Methode, 411
~ Schätzer, 352
Maximum-Methode, 149
Maximum-posteriori-Schätzer, 443
McNemar-Test, 318
Median
~ Test, 318
~ Verfahren, 150, 285, 286, 289
Medoid, 277
~ Verfahren, 283
mehrdimensionale
~ Skalierungsverfahren, 135
mehrdimensionale Skalierung
nichtmetrische ~, 37, 41, 77
ordinale ~, 37
Messfehler
Modellierung zufälliger ~, 353
Messniveau
gemischtes ~, 336
Methode
Block-Clustering ~, 371
Forgys ~, 300
Hill-Climbing ~, 338
Maximum-Likelihood ~, 411
Minimum ~, 149
Posterior-Mode ~, 411
Split-Half ~, 329
Square-Error-Clustering ~, 300
Metrik
~, 195, 219
Canberra ~, 214, 215
Chebychev ~, 470
City-Block ~, 82, 197, 209, 214, 220
Signifikanztest ~, 222
Minkowski ~, 81, 197
Metrikparameter
Gewichtung durch ~, 227
~ p , 81, 90
~ q , 220, 469
~ r , 220, 469
Minimaldistanzverfahren, 300
Minimierung
~ Determinante, 338

- ~ Spur, 338
 - ~ STRESS, 103
 - Minimum-Methode**, 149
 - Minkowski-Metrik, 81, 195, 197, 219
 - Mischverteilungsverfahren, 352
 - Submodelle von ~, 352
 - Mittelwertprofil, 315
 - Mittelwertsubstitution, 228
 - Mittelwertverfahren, 149, 150, 264
 - Mittelwertzentrierung, 188, 190
 - empirische ~, 188
 - mittlere
 - ~ Gesamtpunktwerte, 125, 127
 - ~ Gesamtstreuungsquadratsumme, 323
 - ~ Unähnlichkeit, 249
 - Modell**
 - Identifikation zu schätzendes ~, 354
 - Item-Response ~, 135
 - Modellanpassung**
 - Prüfung der ~, 27
 - Modellansatz**
 - ~ AutoClass, 441
 - ~ Ward-Verfahren, 299
 - modellbasierte**
 - ~ Verfahren, 20
 - Modellierung**
 - Vorteile Bayes ~, 440
 - ~ zufälliger Messfehler, 353
 - Modellprüfgrößen**, 366
 - ~ clusteranalytische Interpretation, 72
 - ~ multiple Korrespondenzanalyse, 67
 - Modellspezifikationen**
 - Stabilitätsprüfung ~, 330
 - modifizierte erklärte Streuung $\tilde{\eta}^2$, 340
 - MOJENA-I**
 - ~ Teststatistik, 243
 - MOJENA-II**
 - ~ Teststatistik, 244
 - Mokken-Skalierung, 136
 - Monotoniebedingung**
 - schwache ~, 81
 - starke ~, 81
 - M-Schritt, 359
 - multidimensional analysis**
 - nonmetrical ~, 37
 - multidimensionale Skalierung**
 - nichtmetrische ~, 37
 - multiple**
 - ~ Korrespondenzanalyse, 37, 43, 57, 77, 118
 - ~ Korrespondenzanalyseskalierung, 111
- N**
- Nachteil Reskalierung Faktorladungen, 65
 - Nächste-Nachbarn-Verfahren, 148, 233
 - verallgemeinerte ~, 259
 - natürliche**
 - ~ Cluster, 16
 - Needleman-Wunsch-Algorithmus, 482
 - Newton-Raphson**
 - ~ Verfahren, 414
 - nicht-hierarchische Variablen, 183
 - nichtlineares Projektionsverfahren, 37
 - nichtmetrische**
 - ~ mehrdimensionale Skalierung, 37, 41, 77
 - ~ multidimensionale Skalierung, 37
 - Nichtvergleichbarkeit, 176
 - Problem der ~, 353
 - nominale**
 - ~ Assoziationsmaße, 207
 - ~ Faktorenanalyse, 38, 117
 - nonmetrical multidimensional analysis, 37
- O**
- Objektauswahl**
 - Stabilitätsprüfung ~, 329
 - Objekte**
 - Standardisierung ~, 188
 - objektorientierte**
 - ~ Clusteranalyse, 21
 - ~ Datenanalyse, 16, 153
 - Objektzuordnung**
 - Stabilitätsprüfung ~, 328
 - Optimal Scaling, 37
 - optimales**
 - ~ Partitionsverfahren, 300
 - ordinale**

- ~ Assoziationsmaße, 211, 212
- ~ Korrelationskoeffizienten
 - Signifikanztest ≈, 218
- ~ mehrdimensionale Skalierung, 37
- Orthogonalisierung
 - automatische ~, 193
- P**
- p , Metrikparameter ~, 81, 90
- Paarvergleich, 78
- paarweises
 - ~ Ausscheiden, 228
- Parameter
 - Priori ~, 414
- parametrischer
 - ~ Bootstrap, 422
- partitionierende
 - ~ Verfahren, 19
- Partitions-Index, 371
 - Schwellenwerte ~, 371
- Partitionsverfahren
 - optimales ~, 300
- Pearson- χ^2
 - ~ Statistik, 417
- Φ
 - ~ Koeffizient, 197
 - Signifikanztest ≈, 205
 - ~ Korrelation, 197
- φ , Dichtefunktion ~, 357
- Pivotelement, 277
- Poissonrate, 410
- Poisson-Verteilung, 409
- polychorische Korrelation, 129
- Posterior-Mode
 - ~ Methode, 411
- Präferenzdaten
 - ~ mit Auswahl, 134
- PRE-Koeffizient, 307
 - Abfall ~, 310
- Priori
 - ~ Parameter, 414
 - ~ Wahrscheinlichkeit, 439
- probabilistische
 - ~ Clusteranalyseverfahren, 18, 19, 351
- Problem der Nichtvergleichbarkeit, 353
- Produkt-Moment-Korrelation, 197, 219
 - Signifikanztest ≈, 221
- Profilähnlichkeit
- Koeffizient r_p der ~, 223
- Profilanalyse
 - latente ~, 351, 355
- Profile, 403
- Projektionsverfahren
 - nichtlineares ~, 37
- Proportionalität
 - Koeffizient der ~, 227
- prozentuale Verbesserung
 - ~ gegenüber Nullmodell, 363
 - ~ gegenüber vorausgehende Klassenlösung, 364
- Prüfgröße
 - χ^2 ~, 67
- Prüfung
 - ~ der Modellanpassung, 27
 - ~ Einfachstruktur, 97
- PV_K , 364, 366, 425, 428
- PVO_K , 363, 428
- Q**
- Q , Korrelationsmatrix ~, 138
- q , Metrikparameter ~, 220, 469
- Q-Analyse, 136
 - Vorgehen ~, 137
- Q-Faktorenanalyse, 122
- Q-Korrelation, 219
- quadrierte euklidische Distanz, 73, 197, 220, 289
 - ~ für dichotome Variablen, 198
- Signifikanztest ~, 221
- Quantification Method, 37
- Quartimin
 - ~ Rotation, 124
- Quick-Clustering
 - ~ Verfahren, 336
- R**
- R , Korrelationsmatrix ~, 122
- r , Metrikparameter ~, 220, 469
- R^2
 - ~ Maßzahlen, 421

- R-Analyse, 122, 127
- RAND-Index, 272
 - adjustierter ~, 274
- Randlösungen, 413
- Rasch-Skalierung, 136
- rechtwinklige
 - ~ Rotation, 92
- Repräsentant, 277, 323
- Repräsentanten-Verfahren, 150, 277
 - Weiterentwicklungen ~, 282
- Residuen
 - bivariate ~, 424
 - standardisierte ~, 417
- Residuenabweichungen
 - Anpassungsindex der mittleren ~, 46
- Residuenquadratsumme
 - Wurzel mittlere ~, 69
- Reskalierung
 - ~ Eigenvektoren, 57, 61, 110
 - ~ Faktorladungen, 61
 - ~ Faktorladungen, Nachteil, 65
- Responseskalierung, 78, 90
- R-Faktorenanalyse, 122
- Richtwert
 - ~ maximale Dimensionszahl, 87
- Richtwerte
 - ~ STRESS, 85
- R-Korrelation, 219
- Robustheit
 - ~ einer Clusterlösung, 170, 196
- Rotation, 120
 - ~ Faktoren, 66, 118, 124, 138
 - Quartimin ~, 124
 - rechtwinklige ~, 92
 - schiefwinklige ~, 118, 124
 - Varimax ~, 124, 138
 - Verfahren ~, 118
- r_p , Koeffizient der Profilähnlichkeit ~, 223
- RUNT-Teststatistik, 253
- S
 - Scalogram Analysis, 37
 - Schätzer
 - Bayes ~, 352
 - Maximum-a-posteriori ~, 443
- Maximum-Likelihood ~, 352
- Maximum-posteriori ~, 443
- Schätzung
 - Kommunalitäten ~, 118, 123
- Schätzwerte
 - ~ für fehlende Werte, 228
- schiefwinklige Rotation, 118, 124
- schwache Monotoniebedingung, 81
- Schwellenwerte ~
 - ~ η^2 , 310
 - γ -Koeffizient, 238
 - ~ GFID*-Index, 74
 - ~ GFIR*-Index, 74
 - ~ Heterogenität, 280
 - ~ Homogenität, 280
 - ~ Partitions-Index, 371
- Schwierigkeitsgrad, 127
- Scree-Diagramm, 123
- Scree-Test, 90, 123
 - inverser ~, 241, 290
- Sequenzen
 - ~ Statuskonfigurationen, 475
- Signifikanz Zusammenhangsstruktur, 67
- Signifikanztest ~
 - ~ Bealsche F-Werte, 308
 - ~ City-Block-Metrik, 222
 - ~ euklidische Distanz, 221
 - ~ F-Wert, 321
 - γ -Koeffizient, 238
 - ~ Homogenitätsindex G, 249
 - ~ JACCARD-I-Koeffizient, 205
 - ~ ordinale Korrelationskoeffizienten, 218
 - ~ Φ -Koeffizient, 205
 - ~ Produkt-Moment-Korrelation, 221
 - ~ quadrierte euklidische Distanz, 221
 - ~ SMK, 205
 - ~ Übereinstimmungskoeffizient κ , 206
 - ~ (Un-)Ähnlichkeitsmaße, 204
- Silhouette-Koeffizient, 495
- Simple-Matching-Koeffizient, 201
- Single-Linkage, 149, 251, 326
 - Verkettungseffekt ~, 252
- Skalenkennwerte

- theoretische ~, 178
- Skalenwerte
- Berechnung ~, 63, 64
 - empirische ~, 178
- Skalierung
- Darstellung kanonische ~, 113
 - kanonische ~, 111–113
 - Mokken ~, 136
 - nichtmetrische mehrdimensionale ~, 37, 41, 77
 - nichtmetrische multidimensionale ~, 37
 - ordinale mehrdimensionale ~, 37
 - Rasch ~, 136
- Skalierungsverfahren
- mehrdimensionale ~, 135
- SMK, 201
- Signifikanztest ~, 205
- SMK^{nom}, 209, 210
- SMK^{ord}, 216
- Spaltenskalierung, 110–113
- Darstellung ~, 115
 - Fragestellung ~, 113
- Split-Half-Methode, 329
- Spur
- Maximierung ~, 338
 - Minimierung ~, 338
- Square-Error-Clustering
- ~ Methode, 300
- Stabilitätsprüfung, 27, 328
- ~ Clusterzahl, 328
 - ~ Clusterzentren, 328
 - ~ Modellspezifikationen, 330
 - ~ Objektauswahl, 329
 - ~ Objektzuordnung, 328
 - ~ Variablenauswahl, 328
- standardisierte
- ~ Datenmatrix Z , 137
 - ~ Distanz, 323
 - ~ Residuen, 417
- Standardisierung, 191
- Gewichtung durch ~, 226
 - ~ Objekte, 188
 - theoretische ~, 177
- starke Monotoniebedingung, 81
- Startwerte, 415
- Eingabe ~, 97
- Startwertverfahren, 335
- alternative ~, 335
- Statistik
- Cressie-Reed- χ^2 ~, 417
 - Klassifikations ~, 420
 - Likelihood-Ratio- χ^2 ~, 417
 - Log-Likelihood ~, 420
 - Pearson- χ^2 ~, 417
- Statuskonfigurationen
- Sequenzen ~, 475
- Stimulusskalierung, 78, 90, 295
- STRESS, 81
- Anpassungsindex für den ~, 93
 - Beurteilung ~, 92
 - Minimierung ~, 103
 - Richtwerte ~, 85
- STRESS-2-Koeffizient, 93
- Streuungsquadratsumme
- ~ in den Clustern, 299, 305
- Strukturheterogenität, 120, 125, 127
- Strukturmatrix, 124
- Submodelle
- ~ von Mischverteilungsverfahren, 352
- symmetrische
- ~ Ähnlichkeitsmatrix, 143
- T
- Teilmatrix Zusammenhangsmatrix G , 109
- Test
- Bartlett ~, 122
 - Bootstrap LQ ~, 423
 - χ^2 ~, 67, 68, 224
 - χ^2 -Unabhängigkeits ~, 318
 - Kolmogorov-Smirnov ~, 224
 - Lo-Mendell-Rubin-Likelihood-Ratio ~, 422
 - Mann-Whitney ~, 318
 - McNemar ~, 318
 - Median ~, 318
 - t ~, abhängige Stichproben, 318
 - t ~, unabhängige Stichproben, 318
 - U ~, 318
 - Wald ~, 417

- Wilcoxon ~, 318
- Teststatistik
 - MOJENA-I ~, 243
 - MOJENA-II ~, 244
 - RUNT ~, 253
- tetrachorische
 - ~ Korrelation, 129
- Tetradenvergleich, 78
- Theorem
 - Bayes ~, 439
- theoretische
 - ~ Skalenkennwerte, 178
 - ~ Standardisierung, 177
- Trägheit, 69, 113
 - ~ bivariate Korrespondenzanalyse, 69
 - ~ multiple Korrespondenzanalyse, 69
- Transformationskosten
 - Festsetzung ~, 485
- transponierte Datenmatrix ZQ , 137
- Triadenvergleich, 78
- t-Test
 - ~ abhängige Stichproben, 318
 - ~ unabhängige Stichproben, 318
- TwoStep-Cluster
 - ~ Algorithmus, 446
- TwoStep-Clusteranalyse, 20

- U**
- U , Unähnlichkeitsmatrix ~, 80
- Übereinstimmungskoeffizient
 - ~ κ , 201
 - Signifikanztest ~, 206
 - ~ κ^{nom} , 209
 - ~ κ^{ord} , 216
- überlappende
 - ~ Cluster, Complete-Linkage für ~, 255
 - ~ Clusteranalyseverfahren, 20
- Überlappung, 323
- Überlappungsanteil, 354, 368
- Überprüfung faktorenanalytische Interpretation, 69
- Überschätzung
 - ~ Anzahl Dimensionen, 129
- unabhängige Stichproben

- t-Test ~, 318
- Unabhängigkeit
 - Annahme lokale ~, 372
 - lokale ~, 352
- Unabhängigkeits-Test
 - χ^2 ~, 318
- Unähnlichkeit
 - durchschnittliche ~, 249
 - mittlere ~, 249
- (Un-)Ähnlichkeitsmaß
 - abgeleitetes ~, 195
 - anderes ~, 195
 - Auswahl ~, 41, 195
 - ~ Distanzmaß, 195
- (Un-)Ähnlichkeitsmaße
 - Signifikanztest ~, 204
 - Wahrscheinlichkeitsverteilungen ~, 204
- Unähnlichkeitsmatrix, 41
 - Berechnung ~, 41
 - Darstellung ~, 77
 - Gewinnen ~, 78
- Unähnlichkeitsmatrix U , 80
- Unfolding, 135
- Ungleichung
 - Chebychevsche ~, 223
- unrotierte Hauptkomponenten, 193
- unvollständige Clusteranalyseverfahren, 18, 37, 38
- U-Test, 318

- V**
- Validitätsprüfung, 27, 41
 - ~ Cluster, 41
 - formale ~, 27
 - inhaltliche ~, 332
- Variablen
 - abgeleitete ~, 459
 - Ähnlichkeitsmaße dichotome ~, 200
 - bedingte ~, 175
 - dichotome ~, 122, 127, 197
 - quadrierte euklidische Distanz für ~, 198
- Faktorenanalyse dichotome ~, 127
- Faktorenanalyse ordinale ~, 127

- hierarchische ~, 183
irrelevante ~, 168, 170
nicht-hierarchische ~, 183
Vergleichbarkeit ~, 175
Variablenausprägung
 lineare Restriktion ~, 387
Variablenauswahl
 Stabilitätsprüfung ~, 328
variablenorientierte
 ~ Clusteranalyse, 22
 ~ Datenanalyse, 16, 153
Varianz
 erklärte ~, 417
varianzanalytische Maßzahlen, 366
Varimax-Rotation, 124, 138
Vektorrepräsentation, 102, 116
verallgemeinerte
 ~ Nächste-Nachbarn-Verfahren, 259
verallgemeinerte Binomial-Verteilung, 210
Verallgemeinerung
 ~ K-Means-Verfahren, 351, 352
 ~ latente Klassenanalyse, 352
Veranschaulichung
 ~ K-Means-Algorithmus, 304
Verbesserung einer Ausgangspartition
 Verfahren zur ~, 300
Verfahren
 asymptotisches Verhalten K-Means ~,
 301
 divisive ~, 19
 Expectation-Maximization ~, 414
 heuristische ~, 20
 Interpretationsziel K-Means ~, 314
 iteratives reallokatives Sum-of-Squares ~, 301
K-Cluster ~, 346
K-Means ~, 150, 299
K-Median ~, 345
K-Medoids ~, 347
K-Modus ~, 345, 346
Median ~, 150, 285, 286, 289
Medoid ~, 283
modellbasierte ~, 20
Nächste-Nachbarn ~, 148, 233
Newton-Raphson ~, 414
partitionierende ~, 19
Quick-Clustering ~, 336
Repräsentanten ~, 150, 277
 ~ Rotation, 118
verallgemeinerte Nächste-Nachbarn ~, 259
Verallgemeinerung K-Means ~, 351, 352
Ward ~, 285, 286
Weiterentwicklungen Repräsentanten ~, 282
Zentroid ~, 150, 285, 286, 289
 ~ zur Konstruktion von Clustern, 150
 ~ zur Verbesserung einer Ausgangspartition, 300
z-Werte K-Means ~, 317
Vergleichbarkeit
 ~ Variablen, 175
Verhältnis
 ~ Freiheitsgrade, 87
Verkettungseffekt, 152
 ~ Single-Linkage, 252
Verläufen
 Klassifizierung von ~, 475
Verschmelzungsniveau
 Interpretation ~, 236
Verschmelzungsschema
 Zufallstestung ~, 245
Verteilung
 Binomial ~, 409
 Poisson ~, 409
 verallgemeinerte Binomial ~, 210
Vierfeldertafel, 199
Vorgehen
 ~ bivariate Korrespondenzanalyse, 109
 ~ Datenanalyse, 40
 ~ Q-Analyse, 137
Vorteile
 ~ Vorteile-Bayes, 440
W
Wahrscheinlichkeit
 A-priori ~, 439
 Priori ~, 439

- Wahrscheinlichkeit ~, 439
 - Wahrscheinlichkeitsverteilungen
 - ~ Distanzmaße, 224
 - ~ (Un-)Ähnlichkeitsmaße, 204
 - Wald-Test, 417
 - Ward-Verfahren, 150, 285, 286
 - Modellansatz ~, 299
 - Weighted-Average-Linkage, 150, 264
 - Weiterentwicklungen
 - ~ Repräsentanten-Verfahren, 282
 - Werte
 - fehlende ~, 228
 - imputierte ~, 228
 - Wilcoxon-Test, 318
 - Within-Average-Linkage, 150, 265
 - Wurzel
 - ~ mittlere Residuenquadratsumme, 69
- Z**
- Z**, standardisierte Datenmatrix ~, 137
 - Zählvariablen, 395, 398, 399, 409
 - Zeilenskalierung, 110–112
 - Zentroid-Verfahren, 150, 285, 286, 289
 - ZQ**, transponierte Datenmatrix ~, 137
 - zu schätzendes Modell
 - Identifikation ~, 354
 - Zufallstestung
 - ~ Clusterlösung, 313
 - ~ Verschmelzungsschema, 245
 - Zuordnungswahrscheinlichkeit, 351
 - Zusammenhangsmatrix
 - Berechnung empirische ~, 57, 58
 - ~ **G**, Teilmatrix, 109
 - Zusammenhangsstruktur
 - Signifikanz ~, 67
 - z-Werte
 - ~ K-Means-Verfahren, 317