

# Inhalt

<b>1</b>	<b>Einleitung</b>	<b>15</b>
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
<b>I</b>	<b>Unvollständige Clusteranalyseverfahren</b>	<b>35</b>
<b>2</b>	<b>Einleitende Übersicht</b>	<b>37</b>
<b>3</b>	<b>Multiple Korrespondenzanalyse</b>	<b>43</b>
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 . . . . .	14
<b>4</b>	<b>Nichtmetrische mehrdimensionale Skalierung</b>	<b>77</b>
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
<b>5</b>	<b>Weitere räumliche Darstellungsverfahren</b>	<b>109</b>
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
<b>II</b>	<b>Deterministische Clusteranalyseverfahren</b>	<b>145</b>
<b>6</b>	<b>Einleitende Übersicht</b>	<b>147</b>
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
<b>7</b>	<b>Gewichtung und Transformation von Variablen</b>	<b>175</b>
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
<b>8</b>	<b>Unähnlichkeits- und Ähnlichkeitsmaße</b>	<b>195</b>
8.1	Auswahl eines (Un-)Ähnlichkeitsmaßes . . . . .	196
8.2	Dichotome Variablen . . . . .	197
8.3	Nominale 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
<b>9</b>	<b>Nächste-Nachbarn- und Mittelwertverfahren</b>	<b>233</b>
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
<b>10</b>	<b>Repräsentanten-Verfahren</b>	<b>277</b>
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
<b>11</b>	<b>Hierarchische Verfahren zur Konstruktion von Clusterzentren</b>	<b>285</b>
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

<b>12</b>	<b>K-Means-Verfahren</b>	<b>299</b>
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
<b>III</b>	<b>Probabilistische Clusteranalyseverfahren</b>	<b>349</b>
<b>13</b>	<b>Einleitende Übersicht</b>	<b>351</b>
<b>14</b>	<b>Latente Profilanalyse</b>	<b>355</b>
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
<b>15</b>	<b>Analyse latenter Klassen für nominale, ordinale und gemischtskalierte Variablen</b>	<b>377</b>
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

<b>16</b>	<b>Latent-GOLD-Ansatz</b>	<b>395</b>
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	$\chi^2$ -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
<b>17</b>	<b>Weiterentwicklungen und Modifikationen</b>	<b>439</b>
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
<b>IV</b>	<b>Spezielle Anwendungsfragen</b>	<b>457</b>
<b>18</b>	<b>Häufig gestellte Anwendungsfragen</b>	<b>459</b>
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
<b>19</b>	<b>Klassifikation von Verläufen mittels Optimal Matching</b>	
	<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
<b>20</b>	<b>Formale Gültigkeitsprüfung und Konsensuslösungen</b>	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
<b>Literatur</b>		503
<b>Register</b>		523