16 Korrespondenzanalyse

Jörg Blasius

Universität Bonn

Zusammenfassung. In den Sozialwissenschaften, insbesondere in der Umfrageforschung, gibt es eine Vielzahl von kategorialen Daten. Diese können mit Hilfe der Korrespondenzanalyse in einen gemeinsamen Raum projiziert und die Distanzen zwischen den Merkmalsausprägungen können als Ähnlichkeiten interpretiert werden; dabei gilt: Je dichter zwei Ausprägungen beieinander liegen bzw. je ähnlicher deren Winkel zum Achsenkreuz sind, desto ähnlicher sind sie. Diese Möglichkeit der Visualisierung ist vermutlich der wichtigste Grund für die in den letzten Jahren zu verzeichnende deutliche Zunahme der Anwendungen dieses Verfahrens. Die meisten und die wohl auch bekanntesten sozialwissenschaftlichen Anwendungen der Korrespondenzanalyse kommen bis dato aus dem französischen Sprachbereich, wobei insbesondere die Arbeiten von Pierre Bourdieu genannt werden können. In diesem Artikel werden die grundlegenden Elemente der Korrespondenzanalyse vorgestellt und es werden Anwendungen auf verschiedene kategoriale Daten diskutiert, die dem ALLBUS 2002 entnommen wurden.

1 Einleitung

Seit etwa 20 Jahren wird in den Sozialwissenschaften zunehmend ein neues multivariates Auswertungsverfahren eingesetzt, die Korrespondenzanalyse – ein exploratives Verfahren zur grafischen und numerischen Darstellung von Zeilen und Spalten beliebiger Datenmatrizes mit nicht-negativen Einträgen. Analysiert werden können u. a. einfache Häufigkeitstabellen, zusammengesetzte Tabellen, ordinale und metrische Daten, Rangdaten, Multi-Response-Tabellen, mehrdimensionale Tabellen, quadratische Tabellen, Burt-Matrizen und Indikatormatrizen. In den Sozialwissenschaften dürfte dieses Verfahren insbesondere durch die Arbeiten von Bourdieu, so z. B. „Die feinen Unterschiede“ 1982 und „Homo Academicus“ 1984, bekannt geworden sein, der die Korrespondenzanalyse als statistische Grundlage zu seiner Theorie der sozialen Räume verwendet hat (Bourdieu 1991; Rouanet et al. 2000; Le Roux & Rouanet 2004; Blasius & Friedrichs 2008).

Die Korrespondenzanalyse kann als Hauptkomponentenanalyse mit nominalen Daten bezeichnet werden. Ähnlich wie bei der Hauptkomponentenanalyse werden mit Hilfe eines (verallgemeinerten) Kleinsten-Quadrate-Schätzers Achsen bestimmt, mit denen ein latenter Raum aufgespannt wird. Der wohl wichtigste Vorteil der Korrespondenzanalyse gegenüber der Hauptkomponentenanalyse ist, dass als Eingabedaten kategoriale Variablen verwendet werden können. In den Sozialwissenschaften gibt es

S. 367–389 in: Christof Wolf & Henning Best, Hg. (2010). Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften

C. Wolf , H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse,*
DOI 10.1007/978-3-531-92038-2_16,

eine Vielzahl von kategorialen Daten, insbesondere Umfragedaten werden überwiegend mit Hilfe von nominal oder ordinal skalierten Fragen erhoben.

Die Korrespondenzanalyse, wie sie im Folgenden dargestellt wird, wurde in den sechziger Jahren in Frankreich unter der Leitung von Jean-Paul Benzécri entwickelt. Sie ist dort, zusammen mit anderen Verfahren zur Visualisierung von Daten, wichtigster Bestandteil der Analyse des Donneés, der geometrischen Datenanalyse (vgl. Benzécri & collaborateurs 1973; Le Roux & Rouanet 1998, 2004).

Außerhalb Frankreichs begann die Diskussion der statistischen Grundlagen der Korrespondenzanalyse erst mit den beiden 1984 erschienenen englischsprachigen Lehrbüchern von Lebart et al. und von Greenacre. Mit der Ende der achtziger Jahre erfolgten Aufnahme von Prozeduren in die großen Statistikpakete, und hier vermutlich insbesondere in SPSS, ist auch in den angelsächsischen Ländern ein deutlicher Anstieg der Anzahl von Anwendungen zu beobachten. Aufgrund der vielfältigen Anwendungsmöglichkeiten der Korrespondenzanalyse, der Möglichkeit der grafischen Darstellung der Ergebnisse und insbesondere aufgrund der Verwendung von kategorialen Daten dürfte das Verfahren auch im deutschsprachigen Raum in den nächsten Jahren einen weiter zunehmenden Stellenwert bekommen.

Im einfachsten Beispiel wird eine einfache Kreuztabelle analysiert, in der z. B. in den Spalten die Ausprägungen der Sonntagsfrage stehen können und in den Zeilen die Ausprägungen des Materialismus-Postmaterialismus-Indexes (vgl. Tabelle 1 auf Seite 372). Mit Hilfe der Korrespondenzanalyse kann dann der Zusammenhang zwischen den Spalten und den Zeilen der Tabelle grafisch und numerisch dargestellt werden. Statt einer einzigen Tabelle kann auch eine zusammengesetzte betrachtet werden. In diesem Fall wird eine zu beschreibende Variable, z. B. die Sonntagsfrage, mit einer (beliebigen) Anzahl von beschreibenden Variablen kreuztabelliert, z. B. mit dem Materialismus-Postmaterialismus-Index, dem Alter (in Gruppen) und der Schulbildung. Die Häufigkeiten der einzelnen Kreuztabellen werden zeilenweise untereinander geschrieben, ein Beispiel wird in Tabelle 5 auf Seite 378 gegeben.

Anstelle einer zusammengesetzten Matrix mit einer zu beschreibenden und einem Satz von beschreibenden Variablen, kann auch jede Variable mit jeder anderen kreuztabelliert werden, einschließlich mit sich selbst. Anschließend werden die Tabellen mit den Häufigkeiten zeilen- und spaltenweise verkettet, das Ergebnis ist eine Burt-Matrix ($\mathbf{B}$). Anstelle der Burt-Matrix kann auch die Indikatormatrix ($\mathbf{H}$) als Eingabeinformation verwendet werden, also eine Matrix mit Q_c Spalten (= Anzahl der Variablenausprägungen) und N Zeilen (= Anzahl der Befragten), die als Elemente nur Nullen und Einsen haben (für „genannt“ und „nicht genannt“); aufgrund von $\mathbf{B} = \mathbf{H}^\mathsf{T}\mathbf{H}$ können die Ergebnisse der beiden Analysen ineinander überführt werden. Wird der Algorithmus der (einfachen) Korrespondenzanalyse auf die Burt-Matrix oder auf die Indikatormatrix angewendet, so wird von multipler Korrespondenzanalyse gesprochen, oder, in der niederländischen Tradition, von der Homogenitätsanalyse (vgl. Gifi 1990; Heiser & Meulman 1994; Michailidis & de Leeuw 1998). Bei der Anwendung der Korrespondenzanalyse auf eine Indikatormatrix handelt es sich um eine Individualdatenanalyse.

Während bei der Hauptkomponentenanalyse metrisches Datenniveau vorausgesetzt und auf der Basis der Korrelations- oder der Kovarianzmatrix eine kanonische

Zerlegung durchgeführt wird (siehe auch Kapitel 15 in diesem Handbuch), ist es bei der Korrespondenzanalyse eine verallgemeinerte kanonische Zerlegung (singuläre Wertezerlegung, singular value decomposition oder SVD) auf der Basis der Matrix der standardisierten Residuen. Diese enthält gemäß der Chi-Quadrat-Statistik die gewichteten Abweichungen von beobachteten und erwarteten Werten. Ähnlich wie bei der Hauptkomponentenanalyse gibt es bei der Korrespondenzanalyse Eigenwerte, erklärte Varianzen der Eigenwerte, Faktorladungen und Kommunalitäten, anhand derer die Ergebnisse numerisch beschrieben werden können (siehe Abschnitt 3). Während bei der Hauptkomponentenanalyse meistens auf eine Visualisierung der Ergebnisse verzichtet wird, ist diese bei der Korrespondenzanalyse zentraler Bestandteil für die Interpretation der Daten.

2 Mathematisch-statistische Grundlagen

Bei der Korrespondenzanalyse handelt es sich um ein auf der Chi-Quadrat-Statistik basierendes exploratives Verfahren, welches auf bekannten geometrischen Verfahren basiert (vgl. Greenacre 1984, 2007; Blasius 2001). Zur formalen Darstellung des Verfahrens wird die einfache Korrespondenzanalyse betrachtet, das heißt eine Kreuztabelle mit I Zeilen und J Spalten. Aus den Zellhäufigkeiten der Kontingenztabelle ($\mathbf{N}$) wird im ersten Schritt die Korrespondenzmatrix ($\mathbf{P}$) bestimmt mit den Elementen $p_{ij} = {}^{n_{ij}}/_{n}$ (n_{ij} = Häufigkeiten der Zelle (ij), n = Gesamtsumme der Eingabematrix; wenn nur zwei Variablen berücksichtigt werden ist $n = N$, mit N = Anzahl der Befragten). Des Weiteren werden die durchschnittlichen Zeilen- und Spaltenprofile aus der Division der Zeilen- bzw. der Spaltensummen durch die Gesamtsumme bestimmt, z. B. für die i-te Zeile $r_i = {}^{n_{i+}}/_{n}$. Die durchschnittlichen Profilelemente werden auch als Massen bezeichnet, sie entsprechen den „Gewichten“, die die Zeilen (r_i) und Spalten (c_j) in den Daten haben.

Unter dem Modell der Unabhängigkeit ist das Produkt aus (r_i) und (c_j) gleich dem prozentuierten Erwartungswert der Zelle (ij). Die Differenzen der Werte aus der Korrespondenzmatrix und den dazugehörigen prozentuierten Erwartungswerten $(p_{ij} - r_i c_j)$ entsprechen den Abweichungen von empirischen und den auf der Basis des Unabhängigkeitsmodells erwarteten Werten. Im nächsten Schritt erfolgt gemäß der Chi-Quadrat-Statistik die Gewichtung dieser Abweichungen mit den jeweiligen Massen von Zeilen und Spalten. Für ein beliebiges Element der resultierenden Matrix gilt $a_{ij} = {}^{(p_{ij} - r_i c_j)}/_{\sqrt{r_i c_j}}$, oder in Matrixschreibweise $\mathbf{A} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-1/2}$. Die Ähnlichkeit zur Chi-Quadrat-Statistik wird ersichtlich, wenn die Elemente von $\mathbf{A}(a_{ij})$ quadriert, über die $I \times J$ Zellen aufsummiert und mit n multipliziert werden: $\chi^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$, mit $\hat{n}_{ij} = {}^{(n_{i+} \times n_{+j})}/_{n}$.

$\mathbf{A}$ ist die Matrix der standardisierten Residuen, die Summe ihrer quadrierten Elemente wird als Gesamtträgheitsgewicht (λ_G) bezeichnet. Dieser Wert ist gleich der Summe der Eigenwerte $(\sum \lambda_k)$, er kann zugleich als Maßzahl zur Beschreibung der Variation der Daten verwendet werden (vgl. Blasius 1994, 2001). Wird das Gesamtträgheitsgewicht mit der Gesamtsumme (n) multipliziert, so ist das Ergebnis der Chi-Quadrat-Wert der Ausgangsdaten $(\chi^2 = n\lambda_G)$.

Auf die Matrix der standardisierten Residuen ($\mathbf{A}$) wird eine verallgemeinerte Eigenwertzerlegung (Eckart & Young 1936, bezogen auf die Korrespondenzanalyse Greenacre 1984) angewendet. Die verallgemeinerte Eigenwertzerlegung der Matrix $\mathbf{A}$ mit I Zeilen und J Spalten ist definiert als das Produkt von

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^{\mathsf{T}}\,. \tag{1}$$

Dabei ist $\boldsymbol{\Gamma}$ die Diagonalmatrix mit singulären Werten in absteigender Ordnung $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_K \geq 0$, mit Rang K der Matrix $\mathbf{A}$. Die Spalten von $\mathbf{U}$, diese werden als linke singuläre Vektoren bezeichnet, und $\mathbf{V}$, diese werden als *rechte singuläre Vektoren* bezeichnet, sind orthonormal, so dass $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$. Die Verbindung der Eigenwertzerlegung und der singulären Wertezerlegung kann abgeleitet werden aus:

$$\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{U}^{\mathsf{T}}\mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^{\mathsf{T}} = \mathbf{V}\boldsymbol{\Gamma}^2\mathbf{V}^{\mathsf{T}} \tag{2}$$

und entsprechend

$$\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{U}\boldsymbol{\Gamma}^2\mathbf{U}^{\mathsf{T}}\,. \tag{3}$$

Aus den Gleichungen (1) und (2) wird ersichtlich, dass die rechten singulären Vektoren von $\mathbf{A}$ den Eigenvektoren von $\mathbf{A}^{\mathsf{T}}\mathbf{A}$ entsprechen, die linken singulären Vektoren entsprechen den Eigenvektoren von $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ und die quadrierten singulären Werte von $\boldsymbol{\Gamma}^2$ sind gleich den Eigenwerten (λ_1 bis λ_K) von $\mathbf{A}\mathbf{A}^{\mathsf{T}}$. Diese Eigenwerte werden im Kontext der Korrespondenzanalyse auch *als Trägheitsgewichte der Achsen* (Trägheitsmomente, im englischen: principal inertias) bezeichnet.

Mit Hilfe der Informationen aus der singulären Wertezerlegung können die Hauptkoordinatenwerte für die Ausprägungen der Zeilen- und Spaltenvariable(n) berechnet werden, die für die grafische Darstellung benötigt werden. Für die Lokalisationen der Zeilen ergibt sich die $(I \times K)$ Matrix $\mathbf{F}$:

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Gamma} \tag{4}$$

und für die der Spalten die $(J \times K)$ Matrix $\mathbf{G}$:

$$\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Gamma}\,. \tag{5}$$

Wie in der Hauptkomponentenanalyse und wie auch bei anderen Datenreduktionsverfahren werden so wenig Dimensionen wie möglich für die Interpretation der Ergebnisse verwendet (vgl. auch die Kapitel 14, 15 und 17 in diesem Handbuch). In der Korrespondenzanalyse sind es – u. a. aufgrund der Einschränkungen bei den grafischen Darstellung – sehr oft nur die beiden ersten Achsen. Die Bestimmung der Anzahl der zu berücksichtigenden Dimensionen kann aber auch analog zur Hauptkomponentenanalyse erfolgen: mittels des Eigenwertkriteriums – berücksichtigt werden alle Eigenwerte, deren latente Variablen mehr Varianz binden als der Durchschnitt –, mittels eines Scree-Tests oder durch die Berücksichtigung jener Dimensionen, welche inhaltlich zu interpretieren sind (vgl. ausführlich Blasius 1994, 2001). Aus pragmatischen Gründen wird in sozialwissenschaftlichen Studien meistens die zwei-dimensionale Darstellung verwendet, auch wenn diese nicht immer die den Daten angemessene ist. Wird als

Eingabeinformation die Indikatormatrix verwendet, so können die Faktorwerte der Individuen im Datensatz gespeichert werden. Die so erhaltenen latenten Variablen, die den Mittelwert Null und die Standardabweichung Eins haben, können dann in weitergehenden Analysen verwendet werden, z. B. innerhalb eines Regressionsansatzes (vgl. die Kapitel 24 und 25 in diesem Handbuch).

In der log-linearen Analyse können mit Hilfe von Interaktionseffekten unterschiedlicher Ordnung die Ausgangsdaten rekonstruiert werden (siehe Kapitel 18 in diesem Handbuch). Diese Rekonstruktion erfolgt in der Korrespondenzanalyse mit Hilfe der Lokalisationsparameter und der Eigenwerte. Wie bereits erläutert, werden bei der Korrespondenzanalyse Abweichungen vom Unabhängigkeitsmodell beschrieben (visualisiert). Was in der log-linearen Analyse die Interaktionseffekte unterschiedlicher Ordnung erklären, erklären in der Korrespondenzanalyse die latenten Variablen (die Achsen). Beiden Verfahren ist gemeinsam, dass das sparsamste Modell gewählt werden soll. In der log-linearen Analyse ist es das Modell mit den wenigsten Interaktionseffekten, in der Korrespondenzanalyse ist es das mit den wenigsten Faktoren (eine ausführliche Beschreibung des Zusammenhangs dieser Modelle geben Van der Heijden et al. 1989, 1994). Für die Rekonstruktion der Daten mit Hilfe der latenten Variablen gilt in der Korrespondenzanalyse $\mathbf{P} = \mathbf{r}\mathbf{c}^{\mathsf{T}} + \mathbf{D}_r\mathbf{F}\boldsymbol{\Gamma}^{-1}\mathbf{G}^{\mathsf{T}}\mathbf{D}_c$, wobei $\mathbf{r}\mathbf{c}^{\mathsf{T}}$ der Teil des Unabhängigkeitsmodells ist. Mit Hilfe von k Faktoren ($k = 1, \ldots, k, \ldots, K$) und den dazugehörigen Hauptkoordinaten der Variablenausprägungen auf diesen Faktoren können die Abweichungen von der statistischen Unabhängigkeit modelliert werden. Demzufolge ist die Korrespondenzanalyse nicht nur eine explorative Technik, sondern sie kann (im statistischen Sinn) als Modell bezeichnet werden (vgl. Goodman 1991; Van der Heijden et al. 1994, sowie Kapitel 22 in diesem Handbuch).

3 Ein Beispiel

3.1 Graphische Darstellung

Um einen Überblick über das Verfahren zu geben, verwenden wir eine Häufigkeitstabelle, die aus den Daten der Allgemeinen Bevölkerungsumfrage (ALLBUS) 2002 generiert wurde (Tabelle 1): In den Zeilen stehen die vier Ausprägungen des Materialismus-Postmaterialismus-Indexes, in den Spalten die Parteien, die als Antwort auf die „Sonntagsfrage“ angegeben wurden. In dem Beispiel wurden drei Gruppierungen berücksichtigt, die in vielen anderen Studien als „fehlend“ definiert werden: die Nichtwähler, die Nichtwahlberechtigten und die Verweigerer. Die Aufnahme derartiger Kategorien ist bei der Korrespondenzanalyse prinzipiell immer möglich, analysiert werden kategoriale Daten. Inhaltlich sinnvoll ist die Aufnahme derartiger Kategorien aber nur dann, wenn es eine ausreichend große Fallzahl in den jeweiligen Ausprägungen gibt. Bei großen Fallzahlen, wie sie z. B. im ALLBUS gegeben sind, könnten es 2 bis 5 Prozent sein, bei kleineren Fallzahlen, z .B. bei etwa 500 Befragten, sollten eher 10 Prozent angesetzt werden. Diese Einschränkung gilt jedoch generell und ist keine Besonderheit der Korrespondenzanalyse. Es macht inhaltlich nur selten Sinn, Kategorien zu betrachten, die nur von wenigen Personen genannt wurden, gegebenenfalls sollten die entsprechenden Kategorien zusammengefasst werden.

Tab. 1: Eingabedaten: Materialismus-Postmaterialismus Index mit „Wenn am nächsten Sonntag . . . “

	CDU	SPD	FDP	Grüne	PDS	Andere Partei	Nicht-wähler	Nicht ber.	Verwei-gert	Aktiver Rand
Post-materialisten	113	179	81	109	45	13	30	19	87	676
PM-Mischtyp	219	177	84	45	47	21	65	28	101	787
M-Mischtyp	242	178	80	34	42	22	69	44	115	826
Materialisten	147	123	37	13	27	3	48	31	84	513
Summe	721	657	282	201	161	59	212	122	387	2. 802

Tab. 2: Spaltenprofile: Materialismus-Postmaterialismus Index mit „Wenn am nächsten Sonntag. . . “

	CDU	SPD	FDP	Grüne	PDS	Andere Partei	Nicht-wähler	Nicht ber.	Verwei-gert	Durch-schnitt
Post-materialisten	0,157	0,272	0,287	0,542	0,280	0,220	0,142	0,156	0,225	0,241
PM-Mischtyp	0,304	0,269	0,298	0,224	0,292	0,356	0,307	0,230	0,261	0,281
M-Mischtyp	0,336	0,271	0,284	0,169	0,261	0,373	0,325	0,361	0,297	0,295
Materialisten	0,204	0,187	0,131	0,065	0,168	0,051	0,226	0,254	0,217	0,183
Summe	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Die Häufigkeiten, so wie sie z. B. in Tabelle 1 wiedergegeben sind (ohne die Randsummen), können als Eingabematrix für die einfache Korrespondenzanalyse verwendet werden. Da die Häufigkeiten nur wenig aussagekräftig sind, wurden die Spaltenprozente (oder Spaltenprofile, um es in der Terminologie der Korrespondenzanalyse zu formulieren) angegeben (Tabelle 2). Diese *Spaltenprofile* sind zugleich ein wichtiger Bestandteil für die Interpretation der Ergebnisse – diese erfolgt immer relativ zum Durchschnitt der Verteilung und nicht in absoluten Größen. Bezogen auf Tabelle 2 werden die Spaltenprofile, die Ausprägungen der „Sonntagsfrage“ (die Parteien), ins Verhältnis zu dem durchschnittlichen Spaltenprofil (Spalte „Durchschnitt“) gesetzt.

Aus Tabelle 2 wird ersichtlich, dass es bei den Grünen mehr als doppelt so viele Postmaterialisten gibt als im Durchschnitt aller Befragten, das Verhältnis ist 0,542 zu 0,241. Dem entgegen ist der entsprechende Anteil bei den Anhängern der CDU als auch bei den Nichtwählern und den Nichtwahlberechtigten unterdurchschnittlich. Diese Differenzen werden auch bei der späteren grafischen Darstellung ersichtlich: Die Anhänger der Grünen werden dann den Postmaterialisten zugeordnet, tendenziell auch die Anhänger von FDP und PDS. Anhänger der CDU hingegen werden davon relativ weit entfernt sein, da sie überdurchschnittlich oft materialistisch eingestellt sind. Des Weiteren ist insbesondere bei den Nichtwahlberechtigten der Anteil der Materialisten relativ hoch, bei den Anhängern der anderen Parteien sind es die beiden Mischtypen

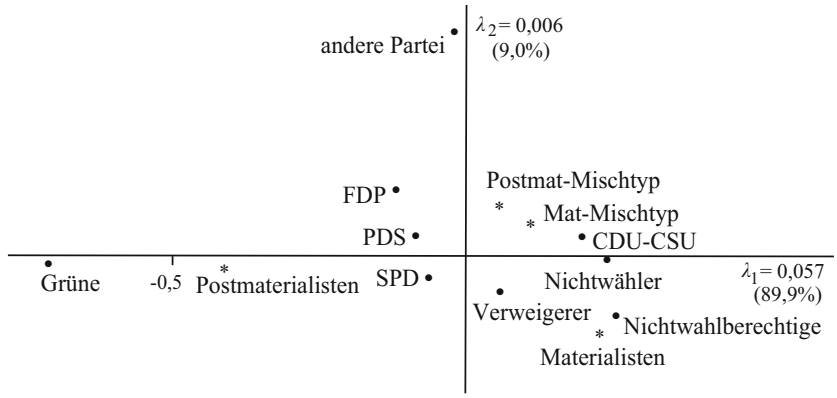

Abb. 1: Graphische Darstellung der Korrespondenzanalyse der Daten aus Tabelle 1

aus Materialisten und Postmaterialisten. Zu den „anderen Parteien“ ist einschränkend zu bemerken, dass deren Wähleranteil mit gut 2 Prozent relativ klein ist und diese Gruppe damit nur eine geringe inhaltliche Bedeutung hat, oder in der Terminologie der Korrespondenzanalyse, sie hat nur eine geringe Masse.

Zu der gleichen Interpretation wie der oben gegebenen würde man kommen, wenn die Prozentuierung der Daten nicht spaltenweise, sondern zeilenweise erfolgen würde (hier nicht gezeigt; ein Beispiel gibt Blasius 2001). Werden die Daten von Tabelle 1 als Eingabeinformation der (einfachen) Korrespondenzanalyse verwendet, so erklärt die erste Dimension 89,9 % der gesamten Variation, die zweite weitere 9,0 % und die letzte verbleibende 1,1 %. Obwohl die erste Dimension zur Interpretation der Ergebnisse ausreichen würde, zeigen wir aus didaktischen Gründen die zweidimensionale Lösung (Abbildung 1).

Die Dimensionalität der (4×9)-Tabelle ist drei (Minimum: Anzahl Zeilen, Spalten minus Eins). Mit den ersten beiden Dimensionen (Abbildung 1) werden insgesamt 98,9 % der gesamten Variation der Daten erklärt. Werden auf die erste Dimension die vier Ausprägungen der Variablen „Materialismus-Postmaterialismus“ im rechten Winkel projiziert, so wird ersichtlich, dass die Postmaterialisten im negativen Bereich liegen und die beiden Mischtypen sowie die Materialisten im positiven Bereich. Dabei bleibt die ordinale Ordnung der vier Ausprägungen im Projektionsraum erhalten – (von links nach rechts) Postmaterialisten, Postmaterialisten-Mischtyp, Materialisten-Mischtyp, Materialisten; d. h. die erste Achse spiegelt die inhaltlich erwartete Reihenfolge wider. Je weiter eine Gruppe im negativen Bereich lokalisiert ist, desto stärker ist ihre (durchschnittliche) postmaterielle Einstellung, je weiter sie im positiven Bereich lokalisiert ist, desto stärker ist ihre (durchschnittliche) materialistische Einstellung. Werden zusätzlich die neun Ausprägungen der Variablen „wenn am nächsten Sonntag Wahlen wären . . . „ auf die erste Achse projiziert, so gibt es eine deutliche Differenzierung: Ganz links im negativen Bereich sind die Anhänger der Grünen lokalisiert, mit relativ weitem Abstand, aber dann ziemlich dicht beieinander (bezogen auf die erste Achse) folgen die Anhänger von FDP, PDS, SPD und der anderen Parteien. Bereits im positiven

Bereich, aber ebenfalls wie die letztgenannten Parteien in unmittelbarer Nähe des Achsenkreuzes, sind die „Verweigerer“ zu finden. Mit etwas größerem Abstand folgen die Wähler von CDU-CSU, die bekennenden Nichtwähler und die Nichtwahlberechtigten. Auf der zweiten Achse sind lediglich die Wähler der anderen Parteien separiert, die allerdings nur von jedem fünfzigsten Befragten angegeben wurden. Ansonsten ist weder eine klare Trennung der vier Typen von Materialisten-Postmaterialisten noch eine der verbleibenden acht Wählergruppen zu verzeichnen. Da diese Trennung aufgrund der geringen Erklärungskraft dieser Achse auch nicht zu erwarten war, wird auf die inhaltliche Interpretation dieser Achse verzichtet.

Bei der *einfachen Korrespondenzanalyse* werden die Häufigkeiten von Kontingenztabellen als Eingabeinformation verwendet (z. B. Tabelle 1), es handelt sich somit um eine Aggregatdatenanalyse. Dargestellt wurden die Hauptkoordinaten von Zeilen und Spalten (Abbildung 1). Die Distanzen zwischen den Zeilen- und Spaltenmerkmalen dürfen bei dieser Art der Visualisierung nicht euklidisch interpretiert werden, ihre Zuordnung erfolgt in diesem Fall ausschließlich über gemeinsame Projektionen auf den Achsenabschnitten bzw. über die Ähnlichkeit der Winkel. Diese Art der Darstellung wird als *symmetrisch* (oder als „French Plot“) bezeichnet, sie wird in den Sozialwissenschaften mit weitem Abstand am häufigsten verwendet. Sollen die Distanzen zwischen Zeilen- und Spaltenmerkmalen interpretiert werden, so muss die „asymmetrische Darstellung“ gewählt werden, auf die hier jedoch verzichtet werden soll (zu den unterschiedlichen Möglichkeiten der grafischen Darstellung und deren Vor- und Nachteile siehe Greenacre 1984, 2007; Blasius 2001).

3.2 Numerische Darstellung

Zusätzlich zu der grafischen Darstellung der Korrespondenzanalyse gibt es auch eine numerische, die in weiten Bereichen jener der Hauptkomponentenanalyse ähnlich ist. Wie auch bei dieser gibt es Faktorwerte, Faktorladungen (hier für die einzelnen Variablenausprägungen) und Kommunalitäten, also die Anteile der erklärten Varianz für die einzelnen Variablenausprägungen, die mit den k berücksichtigten Faktoren erklärt werden können. Des Weiteren werden in der numerischen Lösung die Lokalisationen der Ausprägungen auf den berücksichtigten Achsen und die Anteile der Varianz der Achsen angegeben, die durch die einzelnen Variablen(ausprägungen) erklärt werden, diese werden als *relative Trägheitsgewichte* bezeichnet (Blasius 2001). Die Interpretation dieser relativen Trägheitsgewichte ist ergänzend zu der Interpretation der Faktorladungen. Während mit den Faktorladungen erklärt wird, wie viel Varianz der einzelnen Variablen(ausprägungen) durch die jeweilige Achse erklärt wird, wird mit den Trägheitsgewichten erklärt, wie viel Varianz der Achsen, genauer: der geometrischen Ausrichtung der Achsen im latenten Raum, durch die Variablen(ausprägungen) beschrieben wird. In den Tabellen 3 (Zeilendarstellung) und 4 (Spaltendarstellung) sind die numerischen Ergebnisse für die oben durchgeführte einfache Korrespondenzanalyse wiedergegeben. Als Ausgabeformat wurde jenes von SPSS 17 verwendet (in der deutschen Version), ergänzt durch die Abkürzungen, die nachfolgend für die exemplarischen Berechnungen verwendet wurden.

Tab. 3: Zeilendarstellung

		Wert in Dimension			Beitrag				
					des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		
Inglehart-Index	Masse (r_i)	f_{1i}	f_{2i}	Trägheit (a_i)	s_{1i}	s_{2i}	l_{1i}	l_{2i}	Ges. (L_i)
Post-materialisten	0,241	−0,413	−0,028	0,041	0,717	0,033	0,995	0,005	1,000
PM-Mischtyp	0,281	0,055	0,074	0,003	0,015	0,269	0,311	0,562	0,873
M-Mischtyp	0,295	0,148	0,039	0,007	0,113	0,078	0,883	0,061	0,944
Materialisten	0,183	0,221	−0,139	0,013	0,156	0,620	0,713	0,284	0,997
Summe	1,000			0,064	1,000	1,000			

Tab. 4: Spaltendarstellung

		Wert in Dimension			Beitrag				
Wenn am nächsten Sonntag ...					des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		
	Masse (c_j)	g_{1j}	g_{2j}	Trägheit (b_j)	t_{1j}	t_{2j}	m_{1j}	m_{2j}	Ges. (M_j)
CDU-CSU	0,257	0,195	0,036	0,010	0,171	0,059	0,967	0,033	1,000
SPD	0,234	−0,067	−0,043	0,002	0,019	0,074	0,695	0,278	0,973
FDP	0,101	−0,130	0,089	0,003	0,030	0,140	0,679	0,321	1,000
Grüne	0,072	−0,719	−0,014	0,037	0,645	0,002	0,999	0,000	0,999
PDS	0,057	−0,099	0,008	0,001	0,010	0,001	0,830	0,005	0,835
Andere Partei	0,021	−0,020	0,365	0,003	0,000	0,488	0,003	0,963	0,966
Nicht-wähler	0,076	0,237	−0,002	0,004	0,074	0,000	0,983	0,000	0,983
Nicht berechtigt	0,044	0,242	−0,116	0,004	0,044	0,101	0,713	0,163	0,876
Verweigert	0,138	0,057	−0,075	0,001	0,008	0,134	0,361	0,630	0,992
Summe	1,000			0,064	1,000	1,000			

Die nachfolgenden Berechnungen beziehen sich überwiegend auf die Darstellung der Zeilen, jene der Spalten ist analog zu verstehen und wird nur in wenigen Fällen erläutert. Bei den Massen (r_i) handelt es sich um die relativen Anteile der Zeilen (vgl. auch Tabelle 2, letzte Spalte), so wurden z. B. 24,1 % aller Befragten den Postmaterialisten zugeordnet ($r_1 = 0{,}241$). Bei den Werten f_{1i} und f_{2i} handelt es sich um die *Lokalisationen* der $I = 4$ Zeilen auf den ersten beiden Achsen (in allgemeiner Schreibweise: f_{ki}), also um die Distanzen zum Schwerpunkt der Darstellung; die analogen Werte für die $J = 9$ Spalten sind g_{1j} und g_{2j}.

Aus den Massen und deren Entfernungen zum Schwerpunkt (in der grafischen Darstellung symbolisiert durch das Achsenkreuz), kann das *absolute Trägheitsgewicht* für jede Variablenausprägung auf jeder Achse bestimmt werden. Jenes ergibt sich wie in der Physik (vgl. dort das Prinzip der Balkenwaage) aus dem Quadrat der Entfernung zum Schwerpunkt multipliziert mit der Masse (hier dem Anteil, den die jeweilige Ausprägung an allen Ausprägungen hat). Für die vierte Ausprägung auf der ersten Achse ergibt sich $a_{14} = f_{14}^2 \times r_4 = 0{,}221^2 \times 0{,}183 = 0{,}0089$, und für die zweite Achse $a_{24} = f_{24}^2 \times r_4 = -0{,}139^2 \times 0{,}183 = 0{,}0035$. Die beiden hier berechneten Werte sind in Tabelle 3 nicht wiedergegeben, abgebildet ist die Summe der *absoluten Trägheitsgewichte* aller $K = 3$ Achsen (a_i). Dieser Wert ist für die vierte Ausprägung der Zeilen (Materialisten) $a_4 = 0{,}013$ (vgl. Tabelle 3, Spalte „Trägheit“). Die Summen der absoluten Trägheitsgewichte ergeben die Eigenwerte der jeweiligen Achse (= *Trägheitsgewichte der Achsen*), also $\sum_{i=1}^{I} a_{ki} = \sum_{j=1}^{J} b_{kj} = \lambda_k$ (im gegebenen Beispiel mit $I = 4$ und $J = 9$) oder für die erste Dimension in der Darstellung der Zeilen $\sum_{i=1}^{I} a_{1i} = \lambda_1 = 0{,}057$. Werden die absoluten Trägheitsgewichte auf die Trägheitsgewichte der korrespondierenden Achsen bezogen, so ergeben sich die *relativen Trägheitsgewichte* (für die Zeilen: s_{ki}, für die Spalten t_{kj}). Für die vierte Ausprägung der Zeilen auf der ersten Achse ergibt sich $s_{14} = {}^{a_{14}}/_{\lambda_1} = {}^{0{,}0089}/_{0{,}057} = 0{,}156$. Damit erklärt auf der Ebene der Zeilen die vierte Ausprägung 15,6 % der Variation der ersten Achse (oder besser deren geometrische Ausrichtung im Projektionsraum). Einen deutlich größeren Anteil an der geometrischen Ausrichtung der ersten Achse haben die Postmaterialisten mit 71,7 %.

Die gleichen Berechnungen werden für die Spaltendarstellung durchgeführt. Bei den relativen Trägheitsgewichten fällt hier insbesondere der hohe Wert für die Grünen auf ($t_{14} = 0{,}645$), d. h. auf der Ebene der Spalten determinieren die Grünen zu 64,5 % die geometrische Ausrichtung der ersten Achse. Werden die Trägheitsgewichte bei der inhaltlichen Interpretation berücksichtigt, so wird die gesamte Variation der Daten insbesondere durch die sehr hohen Werte für die Grünen und für die Postmaterialisten erklärt. Die Interpretation der relativen Trägheitsgewichte der Variablenausprägungen ist zentraler Bestandteil vieler Anwendungen der Korrespondenzanalyse im Rahmen der „französischen“ geometrischen Datenanalyse; so verwendet Bourdieu (1982) diese Koeffizienten u. a. in seinen „feinen Unterschieden“, wo sie in der deutschen Übersetzung als „Trägheiten“ bezeichnet werden.

Die absoluten Trägheitsgewichte der einzelnen Variablenausprägungen auf den einzelnen Achsen können auch auf die Summe der absoluten Trägheitsgewichte dieser Ausprägungen bezogen werden, betrachtet wird dann $l_{ki} = {}^{a_{ki}}/_{a_i}$. Diese Werte

entsprechen in der Hauptkomponentenanalyse den *quadrierten Faktorwerten*. Für die vierte Ausprägung der Zeilen auf der ersten Achse ergibt sich ein Wert von $l_{14} = {}^{a_{14}}/_{a_4} = {}^{0{,}0089}/_{0{,}013} = 0{,}713$; d. h. 71,3 % der Variation der Ausprägung „Materialisten“ werden durch den ersten Faktor erklärt, weitere 28,4 % durch den zweiten Faktor. In der Summe sind dies 99,7 %. Dieser letzte Wert wird in der Hauptkomponentenanalyse als *Kommunalität* bezeichnet, er entspricht dem Anteil der durch die ersten k Dimensionen erklärten Varianz (hier 99,7 %). Während bei der Hauptkomponentenanalyse (und ebenso bei der Faktorenanalyse) meistens die Faktorladungen angegeben werden, sind es bei der Korrespondenzanalyse fast immer die quadrierten Faktorladungen (die einfachen Faktorladungen können durch einfaches Radizieren bestimmt werden, die Vorzeichen sind den korrespondierenden Lokalisationsparametern zu entnehmen).

Die Angabe der Faktorwerte, also der Werte, welche die einzelnen Objekte (hier die Befragten) auf den einzelnen Achsen haben, ist bei der Korrespondenzanalyse genauso wie bei der Hauptkomponentenanalyse in den meisten Fällen nicht sinnvoll, aber ebenfalls möglich. Inhaltlich sinnvoll kann eine derartige Angabe sein, wenn nur wenige Objekte vorhanden sind, deren Ähnlichkeiten (Unähnlichkeiten) inhaltlich interpretiert werden sollen, z. B. jene von Politikern oder Professoren (vgl. Bourdieu 1984; Blasius 2001). Die entsprechenden Werte können aber auch gespeichert und in weiteren Analyseschritten als abhängige bzw. als unabhängige Variablen verwendet werden, z. B. in einem Regressionsmodell. Wie bei der Hauptkomponentenanalyse sind die Faktorwerte der Korrespondenzanalyse standard-normalverteilt mit Mittelwert Null und Standardabweichung Eins.

4 Erweiterungen der Korrespondenzanalyse

4.1 Zusammengesetzte Tabellen

Im vorangegangen Abschnitt wurden die Ergebnisse der Korrespondenzanalyse der Tabelle 1 „Parteipräferenz nach Materialismus/Postmaterialismus“ diskutiert. Dabei wurde gesagt, dass mit Hilfe der singulären Wertezerlegung die Variation in der Tabelle, ausgedrückt als $\lambda_G = {}^{\chi^2}/_n$, derart zerlegt wird, dass mit dem ersten Eigenwert (λ_1) ein Maximum an Variation erklärt wird, mit dem zweiten (λ_2) ein Maximum der verbleibenden Variation usw. In dem gegebenen Beispiel bildet die erste Dimension überwiegend den Gegensatz von Materialismus und Postmaterialismus sowie den korrespondierenden Parteipräferenzen ab. Der zweiten Dimension wurde aufgrund ihrer geringen Erklärungskraft keine inhaltliche Interpretation zugeschrieben.

Für eine Erweiterung der Analyse, die als *joint bivariat* bezeichnet werden kann, werden im Folgenden weitere Merkmale mit der Sonntagsfrage kreuztabelliert. Die resultierenden Häufigkeiten werden zeilenweise mit Tabelle 1 verknüpft. In der zu analysierenden Matrix stehen die neun Ausprägungen der Variablen „Parteipräferenz“ in den Spalten und die damit verbunden Variablen(ausprägungen) in den Zeilen. Zusätzlich aufgenommen wurden das Geschlecht, das Alter (fünf Gruppen) und der Schulabschluss (vier Gruppen). Nicht berücksichtigt wurden die fehlenden Werte,

Tab. 5: Eingabedaten: Zusammengesetzte Tabelle

	CDU	SPD	FDP	Grüne	PDS	Andere Partei	Nicht-wähler	Nicht Ber.	Verwei-gerer	Summe
Postmaterialisten	113	179	81	109	45	13	30	19	87	676
PM-Mischtyp	219	177	84	45	47	21	65	28	101	787
M-Mischtyp	242	178	80	34	42	22	69	44	115	826
Materialisten	147	123	37	13	27	3	48	31	84	513
Männer	400	339	148	82	76	42	95	63	152	1397
Frauen	322	321	136	119	85	18	124	59	239	1423
Bis 29 Jahre	117	111	44	45	30	9	34	36	63	489
30 bis 44 Jahre	173	198	70	87	50	28	93	51	139	889
45 bis 59 Jahre	175	156	79	51	48	15	48	26	93	691
60 bis 74 Jahre	209	151	72	11	27	8	33	7	66	584
75 Jahre plus	48	44	19	6	6	0	9	1	27	160
Hauptschule	297	271	93	37	39	22	109	49	163	1080
Realschule	225	195	95	47	60	24	77	33	123	879
Fachabitur	57	51	26	19	13	6	11	5	22	210
Abitur	133	134	69	91	46	6	20	29	72	600
Summe	2877	2628	1133	796	641	237	865	481	1546	11204

von denen es hier auch nur wenige gibt. Die betroffenen Variablen haben damit ein etwas geringeres Gewicht, welches proportional zu der Anzahl der fehlenden Werte ist (zu den Effekten Blasius 2001). Die Anzahl der Ausprägungen schwankt bei den vier beschreibenden Variablen zwischen zwei (Geschlecht) und fünf (Alter), was noch als unbedenklich angesehen werden kann. Große Unterschiede in der Anzahl der Ausprägungen sollten jedoch vermieden werden. Die Eingabedaten sind in Tabelle 5 wiedergegeben; in Abbildung 2 sind die gewichteten Abweichungen vom Unabhängigkeitsmodell der Effekte von „Parteipräferenz“ mit „Materialismus-Postmaterialismus“, „Alter“, „Geschlecht“ und „Schulabschluss“ grafisch dargestellt, die Effekte zwischen den vier beschreibenden Variablen werden bei der Analyse dieser zusammengesetzten Tabelle nicht berücksichtigt.

Die Dimensionalität der Lösung wird bei zusammengesetzten Tabellen aus dem Minimum von Zeilen und Spalten minus der jeweiligen Anzahl von Variablen bestimmt, also min $(I - Q_r, J - Q_c) = min(9 - 1{,}15 - 4) = 8$. Das Gesamtträgheitsgewicht ($\lambda_G = 0{,}047$) entspricht dem durchschnittlichen Trägheitsgewicht der vier Tabellen, wobei die vier Variablen die folgenden Anteile haben: Materialismus-Postmaterialismus: $\lambda_{G.M} = 0{,}064$, Geschlecht: $\lambda_{G.G} = 0{,}018$, Alter: $\lambda_{G.A} = 0{,}055$ und Bildung: $\lambda_{G.B} = 0{,}053$; damit sind die Effekte von „Materialismus-Postmaterialismus“, „Alter“ und „Bildung“ auf die Sonntagsfrage wesentlich stärker als jener des Geschlechts. In der grafischen Darstellung wird dies dadurch sichtbar, dass die beiden Ausprägungen des Geschlechts relativ dicht am Achsenkreuz liegen, während die der anderen drei Variablen relativ stark im Projektionsraum streuen (vgl. Abbildung 2).

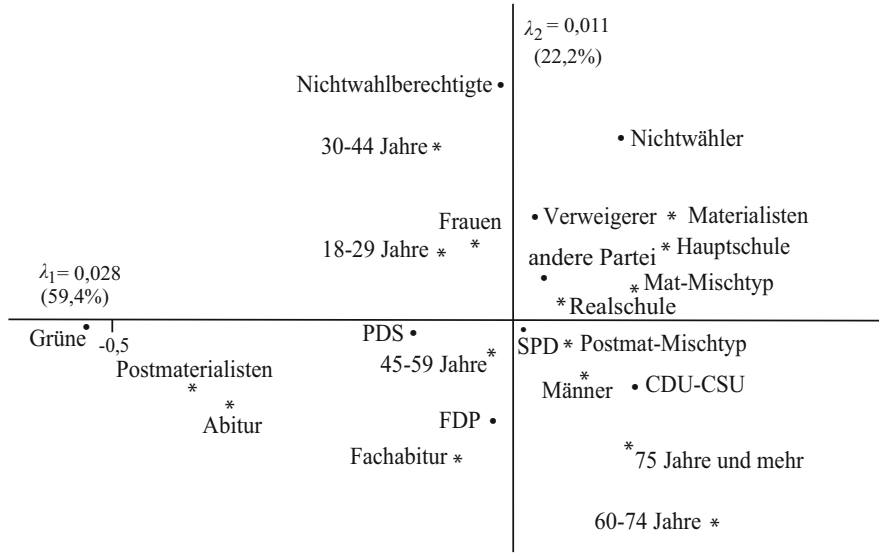

Abb. 2: Graphische Darstellung der Korrespondenzanalyse der Daten aus Tabelle 5

Im Gegensatz zum ersten Beispiel ist die Lösung der Korrespondenzanalyse auf die zusammengesetzte Tabelle zweidimensional mit $\lambda_1 = 0{,}028$ (59,4 % erklärte Varianz) und $\lambda_2 = 0{,}011$ (22,2 %). Dabei spiegelt die erste Dimension auf der Ebene der präferierten Parteien insbesondere den Gegensatz von Grünen (und tendenziell auch PDS) vs. den Anhängern von CDU-CSU und den bekennenden Nichtwählern wider. Die zweite Achse reflektiert insbesondere die Nichtwähler und die Nichtwahlberechtigten vs. Anhänger der FDP und tendenziell jene der CDU-CSU. Auf der Ebene der beschreibenden Variablen korrespondieren die Anhänger der Grünen (und tendenziell jene der PDS) mit den Postmaterialisten und den Abiturienten, die Wähler von CDU-CSU mit den Materialisten, den Älteren (60 bis 74 Jahre sowie 75 Jahre und älter) und den Hauptschülern. Bei der Interpretation von Abbildung 2 ist zu beachten, dass die Effekte zwischen „Materialismus-Postmaterialismus“, „Alter“, „Geschlecht“ und „Bildung“ in der Analyse nicht berücksichtigt wurden. Sollen diese Effekte in die Analyse eingehen, dann muss die multiple Korrespondenzanalyse verwendet werden.

4.2 Multiple Korrespondenzanalyse

Im Folgenden sollen die Zusammenhänge innerhalb eines Sets von Variablen betrachtet werden, also alle Interaktionseffekte erster Ordnung. Bei dieser Art der Analyse, die der Hauptkomponentenanalyse am ähnlichsten ist, wird nach Strukturen (nach latenten Variablen) gesucht, mit denen die Zusammenhänge zwischen den Variablen beschrieben werden können; z. B. haben die Personen, die zu den Materialisten zählen, überdurchschnittlich oft einen Hauptschulabschluss und sind Abiturienten überdurchschnittlich oft Postmaterialisten?

Als Eingabeinformation für die multiple Korrespondenzanalyse kann sowohl die Burt-Matrix als auch die Indikatormatrix verwendet werden, die Lösungen sind ineinander überführbar. So entsprechen die Eigenwerte der Burt-Matrix dem Quadrat der Eigenwerte der Indikatormatrix ($\lambda_{B.k} = \lambda^2_{H.k}$). Das Verhältnis der Lokalisationen der Variablenausprägungen der Burt-Matrix ($\lambda_{B.k}$) zu denen der Indikatormatrix ($\lambda_{H.k}$) kann wie folgt angegeben werden (vgl. Blasius 2001, S. 186): $y^2_{B.jk}/\lambda_{B.k} = y^2_{H.jk}/\lambda_{H.k}$. Das Gesamtträgheitsgewicht errechnet sich im Fall der Burt-Matrix aus der Summe der Trägheitsgewichte der einzelnen bivariaten Tabellen, dividiert durch deren Anzahl (Q^2 mit Q = Anzahl der Variablen). Wie Benzécri (1979) zeigt, sind bei der Zerlegung der Burt- bzw. der Indikatormatrix jene Eigenwerte irrelevant, die kleiner $1/Q$ (bei der Indikatormatrix) bzw. als kleiner $1/Q^2$ (bei der Burt-Matrix) sind. Die verbleibenden Eigenwerte können ebenso wie die Koordinatenwerte reskaliert werden. Die Reskalierung der Eigenwerte erfolgt im Fall der Indikatormatrix durch:

$$\tilde{\lambda}_k = \left(\frac{Q}{Q-1}\left(\lambda_{H.k} - \frac{1}{Q}\right)\right)^2 \text{ mit } \sum_{k=1}^{K} \tilde{\lambda}_k = \tilde{\lambda}_G\,. \tag{6}$$

Werden die reskalierten Eigenwerte auf die Summe der reskalierten Eigenwerte bezogen, so ist der Anteil der erklärten Varianz des ersten Faktors (bzw. in Abhängigkeit von der Anzahl der relevanten Dimensionen, der ersten Faktoren) in der Regel deutlich größer als jene(r) ohne Reskalierung. Greenacre (1988) zeigt, dass die erklärte Varianz des ersten Faktors (der ersten Faktoren) des auf der Basis der singulären Wertezerlegung hervorgehenden Eigenwertes unterschätzt und dass die erklärte Varianz, die auf der Basis der von Benzécri reskalierten Eigenwerte berechnet wurde, überschätzt ist. Durch die Reskalierung verschieben sich die Koordinaten der Variablenausprägungen in Abhängigkeit der Trägheitsgewichte der Achsen, d. h. die Distanzen im latenten Raum werden um achsenspezifische Faktoren verändert. Da aber die relativen Abstände auf den einzelnen Achsen erhalten bleiben und da die Interpretationen über die Projektionen auf den Achsen erfolgen sollte, bleibt die Interpretation der Ergebnisse unverändert (ausführlich dazu Blasius 2001).

Als Beispiel für eine multiple Korrespondenzanalyse verwenden wir die im gleichen Datensatz vorhandenen Beurteilungen von neun abweichenden Verhaltensweisen, die jeweils auf einer vierstufigen Skala beantwortet werden sollten. Die Ausprägungen reichen von „halte ich für sehr schlimm“ bis „halte ich für überhaupt nicht schlimm“. Der Wortlaut der Fragen und deren univariate Verteilungen sind in Tabelle 6 wiedergegeben.

Anhand von Tabelle 6 wird ersichtlich, dass alle Variablen relativ viel Varianz haben und dass sie unterschiedlich verteilt sind. Während fast alle Befragten es als zumindest ziemlich schlimm beurteilen, wenn der Mann seine Ehefrau zum Geschlechtsverkehr zwingt, finden dies nur etwas weniger als 25 % der Befragten hinsichtlich homosexueller Beziehungen, nahezu jede(r) zweite beurteilt diese Verhaltensweise als „überhaupt nicht schlimm“. Da es in allen Variablen nur relativ wenige fehlende Werte gibt und da diese zudem noch hoch miteinander korreliert sind, lassen wir sie aus den nachfolgenden Berechnungen heraus – diese Vorgehensweise entspricht dem des „listwise deletion“. Die Fallzahl reduziert sich damit von $N = 2.802$ (vgl. Tabelle 1) auf $N = 2.673$. Eine

Tab. 6: Beurteilungen von Verhaltensweisen, Angaben in Prozent (nur gültige Fälle)

		N	Sehr schlimm	Ziemlich schlimm	Weniger schlimm	Nicht schlimm
A	Ein Mann schlägt sein 10-jähriges Kind, weil es ungehorsam war.	2799	49,6	32,7	16,6	1,1
B	Eine Frau lässt einen Schwangerschaftsabbruch vornehmen, weil sie keine Kinder haben möchte.	2772	21,8	26,6	33,5	18,1
C	Ein Arzt gibt einem unheilbar kranken Patienten auf dessen Verlangen hin ein tödliches Gift.	2775	15,2	13,9	40,7	30,1
D	Ein Arbeitnehmer macht absichtlich beim Lohnsteuerjahresausgleich falsche Angaben und erhält dadurch 500 Euro zuviel Lohnsteuerrückerstattung.	2798	19,4	36,6	35,2	8,8
E	Jemand fährt mit öffentlichen Verkehrsmitteln, ohne einen gültigen Fahrausweis zu besitzen.	2814	16,3	30,5	45,0	8,2
F	Ein Mann zwingt seine Ehefrau zum Geschlechtsverkehr.	2801	79,5	17,3	2,5	0,7
G	Jemand raucht mehrmals in der Woche Haschisch.	2802	45,3	25,2	21,8	7,7
H	Ein Mann hat homosexuelle Beziehungen zu einem anderen Mann.	2800	14,2	10,4	27,0	48,4
I	Ein verheirateter Mann hat mit einer anderen Frau ein Verhältnis.	2786	31,4	40,7	22,0	5,9

elegantere Möglichkeit als den fallweisen Ausschluss der Werte diskutieren Greenacre & Pardo (2006) im Rahmen ihrer *Subset Korrespondenzanalyse*.

Werden die Interaktionseffekte der neun Variablen zu den Beurteilungen von abweichenden Verhaltensweisen mit Hilfe der multiplen Korrespondenzanalyse beschrieben, so ist der erste Eigenwert $\lambda_1 = 0{,}293$ und der zweite $\lambda_2 = 0{,}198$. Die dazugehörigen erklärten Varianzen sind 9,8 % und 6,6 %, diese sind allerdings stark unterschätzt (siehe oben). Die grafische Darstellung der Ergebnisse (die ersten beiden Dimensionen) ist in Abbildung 3 wiedergegeben. Um die Abbildung übersichtlich zu gestalten, wurden die Variablen mit einzelnen Buchstaben gekennzeichnet (zum Wortlaut der Fragen und der Abkürzungen vgl. Tabelle 6). Die Zahlen stehen für die Ausprägungen (1 = sehr schlimm, 2 = ziemlich schlimm, 3 = weniger schlimm, 4 = überhaupt nicht schlimm). Von den neun Variablen sind sieben stark mit der ersten Dimension korreliert, was in der grafischen Darstellung u. a. daran zu erkennen ist, dass in diesen Fällen die ordinale Reihenfolge der jeweils vier Ausprägungen erhalten bleibt (vgl. die rechtwinkligen Projektionen auf der ersten Achse). Die sukzessiven Ausprägungen dieser sieben Variablen wurden zur besseren Lesbarkeit durch gestrichelte Linien verbunden. Damit misst die erste Dimension eine generelle Einstellung zu abweichenden Verhaltenswei-

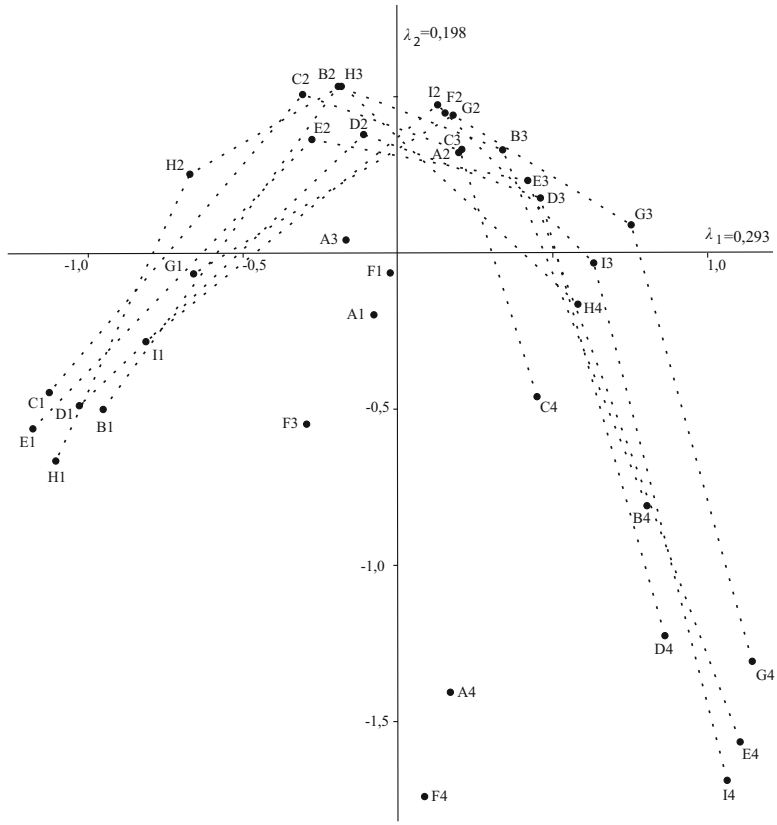

Abb. 3: Graphische Darstellung der multiplen Korrespondenzanalyse

sen, wobei die Faktorwerte umso niedriger sind, je stärker die Ablehnung derartiger Verhaltensweisen ist (negativer Bereich der ersten Dimension). Die beiden Ausnahmen sind die Einstellungen zur Gewalt gegen Kinder (Variable A) und zu Vergewaltigungen in der Ehe (F). Mit beiden Variablen wird anscheinend etwas anderes gemessen als die generelle Einstellung zu abweichenden Verhaltensweisen. Da bei beiden Variablen die Ablehnung überdurchschnittlich hoch ist (vgl. Tabelle 6), kann vermutet werden, dass diese Verhaltensweisen auch von Personen als (sehr) schlimm bewertet werden, die gegenüber anderen abweichenden Verhaltensweisen eine weniger ablehnende Einstellung haben. Die zweite Dimension reflektiert bei allen neun Variablen einen sogenannten „Horseshoe-“ oder Guttman-Effekt, mit negativen Werten bei den Randkategorien und positiven Werten bei den mittleren Kategorien. Dieser Horseshoe-Effekt ist methodisch bedingt (vgl. Greenacre 1984; Van Rijckevorsel 1987; Blasius & Thiessen 2006, 2009), er kann in dem gegebenen Beispiel als zusätzliches Kriterium für die Bedeutung der ersten Dimension angesehen werden. Soll weitere inhaltlich bedingte Variation visualisiert werden, z. B. jene, die durch die beiden Variablen „Gewalt gegen Kinder“

und „Vergewaltigung in der Ehe“ verursacht wird, so kann auch die erste gegen die dritte oder die dritte gegen die vierte Dimension grafisch dargestellt werden.

4.3 Aktive und passive Merkmale

Sowohl bei der einfachen als auch bei der multiplen Korrespondenzanalyse ist es möglich, Variablen(ausprägungen) in einen bereits bestehenden Raum zu projizieren. Diese zusätzlich berücksichtigten *passiven Variablen(ausprägungen)* haben keinen Einfluss auf die geometrische Ausrichtung der Achsen, sie können aber gemeinsam mit den aktiven Variablen(ausprägungen) interpretiert werden. Passive Variablen(ausprägungen) müssen im Fall der einfachen Korrespondenzanalyse lediglich die gleiche Spalten- bzw. die gleiche Zeilenstruktur wie die Ausgangsdaten haben. Auch in der multiplen Korrespondenzanalyse können passive (ergänzende, illustrierende) Variablen oder auch nur einzelne Variablenausprägungen verwendet werden. So können Lebensstilmerkmale, wie sie von Bourdieu (1982) verwendet wurden, auch multipel verknüpft werden. Mit einem derartigen Set von Variablen kann der Projektionsraum aufgespannt werden, und dort können dann sozio-demografische Merkmale als passive Ausprägungen berücksichtigt werden (z. B. Blasius & Friedrichs 2008). Es kann aber auch mit Hilfe von Indikatoren, wie bevorzugten TV-Sendungen, Lebensstilmerkmalen, Schauspielern, Künstlern und Sportlern, ein „sozialer Raum“ aufgespannt werden, in den nachträglich (passiv) die Präferenzen für Produkte projiziert werden. Damit wäre eine Zuordnung von Merkmalen, die u. a. für die Werbung relevant sind (welche Schauspieler bzw. welche Sportler sind die geeigneten Sympathieträger für ein bestimmtes Produkt, welche Sendungen sind ideal für die Platzierung von Werbung), und Produkten möglich (Blasius & Mühlichen 2010). Passive Merkmale können auch dann verwendet werden, wenn es bei einer Variablen viele fehlende Werte oder gar strukturelle Nullen gibt, und wenn nur die inhaltlich relevanten Ausprägungen in die Interpretation einbezogen werden sollen. Da passive Merkmale auch als Ausprägungen mit einem Gewicht von „Null“ betrachtet werden können, und da sie keinen Einfluss auf die geometrische Ausrichtung der Achsen haben, ist die Verteilung der Merkmale relativ beliebig. Die Fallzahlen der einzelnen Kategorien und deren Anzahl pro Variable können sehr unterschiedlich sein.

In Abbildung 4 wurden die Merkmale Parteipräferenz (neun Ausprägungen), Alter (fünf Ausprägungen), Schulabschluss (vier Ausprägungen) und Materialisten-Postmaterialisten (vier Ausprägungen) in den Raum projeziert, der bereits auf der Basis der neun Verhaltensweisen bestimmt wurde (Abbildung 3). Zur besseren Lesbarkeit wurden dies Merkmale kursiv gesetzt. Durch die Berücksichtigung dieser zusätzlichen Merkmale kommt es zu keiner Änderung der geometrischen Ausrichtung der Achsen, der Zusammenhang zwischen den neun abweichenden Verhaltensweisen mit ihren insgesamt 36 Ausprägungen bleibt also unverändert. Die neuen Ausprägungen können aber in die Interpretation eingebunden werden, so dass jetzt u. a. gesagt werden kann, dass je jünger die Befragten sind, desto weniger schlimm finden sie abweichende Verhaltensweisen im Allgemeinen (Abbildung 4). Werden die fünf Ausprägungen des Alters („18-29“ bis „75+“) auf die erste Achse projiziert, dann wird ersichtlich, dass die ordinale Reihenfolge der Altersausprägungen fehlerfrei wiedergegeben wird – je älter die Befragten sind, desto weiter links sind sie auf der ersten Achse lokalisiert,

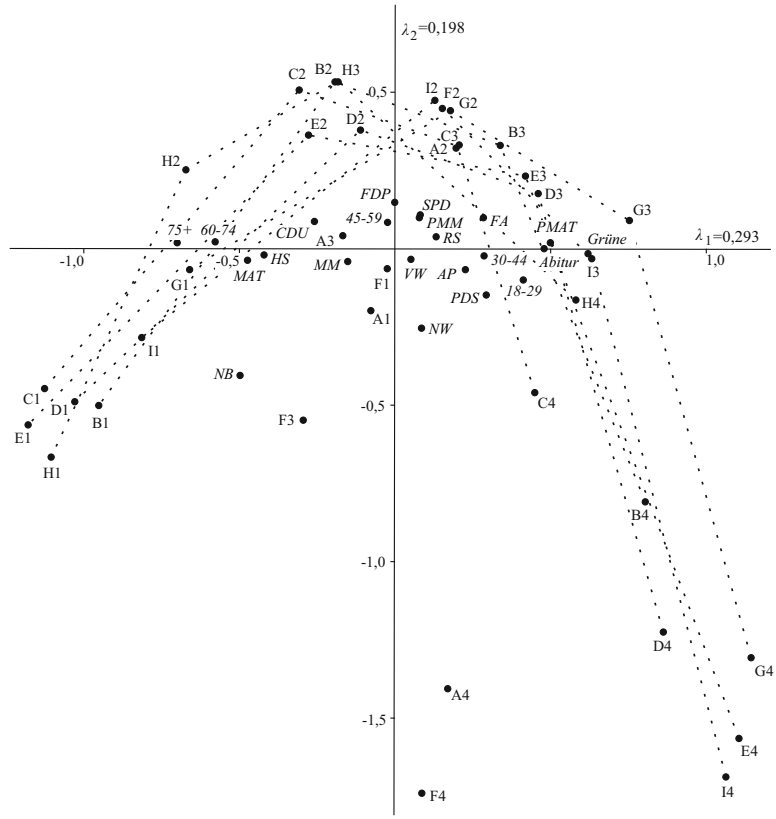

Abb. 4: Graphische Darstellung der Korrespondenzanalyse mit passiven Merkmalen

desto schlimmer beurteilen sie abweichende Verhaltensweisen. Einen ähnlich deutlichen Zusammenhang mit der ersten Dimension gibt es bei der Schulbildung (HS-RS-FA-Abitur): Je niedriger der formale Bildungsabschluss ist, desto stärker die Ablehnung abweichender Verhaltensweisen. Mit der Einbeziehung dieser beiden Indikatoren kann bereits ein plausibles Argument gegeben werden, warum die beiden Merkmale „Gewalt gegen Kinder“ und „Vergewaltigung in der Ehe“ nicht mit der ersten Dimension korrelieren, sondern etwas anderes als eine generelle Ablehnung abweichender Verhaltensweisen messen: Die Ablehnung dieser beiden Verhaltensweisen ist – im Gegensatz z. B. zur Beurteilung des Konsums von Haschisch, der insgesamt ähnlich häufig als sehr schlimm beurteilt wurde wie die Gewalt gegen Kinder – unabhängig von Alter und Bildung der Befragten.

Sehr deutliche Unterschiede in der Beurteilung abweichender Verhaltensweisen bestehen auch bezüglich der Parteipräferenz (NB = Nichtwahlberechtigte, NW = Nichtwähler, VW = Verweigerer der Angabe, AP = andere Partei, CDU = CDU-CSU). Insbesondere die Anhänger der Grünen, aber auch jene der PDS und der

anderen Parteien, beurteilen die vorgegebenen abweichenden Verhaltensweisen als weniger schlimm als der Durchschnitt der Befragten (Ausnahmen: Gewalt gegen Kinder und Vergewaltigung in der Ehe). Dem entgegen beurteilen die Anhänger von CDU und CSU, insbesondere jedoch die Nichtwahlberechtigten, abweichende Verhaltensweisen deutlich kritischer als der Durchschnitt der Bevölkerung. Da es sich bei den Nichtwahlberechtigten vorwiegend um Ausländer handeln dürfte, wird hier das Ergebnis von Friedrichs & Blasius (2000) bestätigt, demzufolge die türkischen Bewohner von benachteiligten Wohngebieten in Köln deutlich normenbewusster als ihre deutschen Nachbarn sind. Bezogen auf die materialistische-postmaterialistische Einstellung gilt, dass je mehr postmaterialistische Werte vorhanden sind, desto weniger schlimm werden abweichende Verhaltensweisen beurteilt (MAT = Materialisten, MM = Materialisten-Mischtyp, PMM = Postmaterialisten-Mischtyp, $PMAT$ = Postmaterialisten).

4.4 Andere Datenformate

Wie im vorangegangenen Abschnitt gezeigt, werden bei der multiplen Korrespondenzanalyse alle Interaktionseffekte erster Ordnung berücksichtigt, was dem Vorgehen bei der Hauptkomponentenanalyse entspricht. Für einige Fragestellungen müssen jedoch die Effekte zweiter bzw. höherer Ordnung berücksichtigt werden. Soll zur Lösung eines derartigen Problems die Korrespondenzanalyse verwendet werden, so muss als Eingabeinformation eine höherdimensionale Kontingenztabelle gewählt werden. Beispiele für derartige Anwendungen geben Greenacre (2007), der die Selbsteinschätzung der eigenen Gesundheit (in fünf Stufen, von „sehr gut“ bis „sehr schlecht“) mit den interaktiv verknüpften Variablen Alter, Geschlecht und Nationalität analysiert. Blasius (2001) untersucht den Zusammenhang der interaktiv verknüpften Variablen „Alter“ und „Geschlecht“ mit den ebenfalls interaktiv verknüpften kulturellen Kompetenzen „Hosen umzunähen“ und „nach Popmusik zu tanzen“. Dabei findet er deutliche Unterschiede bei den Fähigkeiten hinsichtlich von Geschlecht und Alter – Alter ist eng mit Tanzen zu Popmusik verbunden, Geschlecht mit der Fähigkeit Hosen umzunähen –, aber nur einen marginalen Effekt zwischen den beiden kulturellen Kompetenzen. Die Ergebnisse der grafischen Darstellung können hier mit Hilfe des log-linearen Modells auf statistische Signifikanz überprüft werden.

In der Wirtschafts- und Sozialforschung liegen sehr oft Tabellen vor, bei denen in den Zeilen und Spalten die gleichen Ausprägungen stehen. Ein Beispiel für derartige quadratische Tabellen sind bibliometrische Daten: In den Spalten stehen die zitierenden, in den Zeilen die zitierten Autoren und in den Zellen die Häufigkeiten des Zitierens, wobei die Hauptdiagonale die Anzahl der Selbstzitationen enthält. Bei der Auswertung von Paneldaten kann die Forschungsfrage auf die Veränderung des Berufsstatus oder der sozialen Schicht bezogen sein, also auf die vertikale, horizontale bzw. soziale Mobilität. In der Wahlforschung kann gefragt werden, wie groß der Anteil der Stammwähler der Parteien ist, von welcher Partei zu welcher Partei gewechselt wurde, für welche Parteien sich die Nichtwähler der vorangegangenen Wahl entschieden haben und welche Parteien Stimmen an die Nichtwähler verloren haben.

Datengrundlage sind in den gegebenen Beispielen quadratische Tabellen, bei denen in der Regel die Hauptdiagonalen überdurchschnittlich stark besetzt sind: Es bleiben

z. B. mehr Personen ihrer Partei treu als dass Personen zu einer bestimmten anderen Partei wechseln. Mit den ersten Dimensionen der Korrespondenzanalyse von derartigen quadratischen Tabellen werden daher insbesondere die gewichteten Abweichungen der Hauptdiagonalelemente von ihren Erwartungswerten beschrieben. An diesen Stellen wird die meiste Variation verursacht. Von besonderem Interesse sind aber oft die Personen, die von Partei A zu Partei B wechseln – und gerade diese Wechselwähler sollen angemessen beschrieben und grafisch dargestellt werden. Um diese Daten angemessen auswerten zu können, wird sich eines Tricks bedient: Die Daten werden in einen symmetrischen und schief-symmetrischen Teil überführt. Mit dem symmetrischen Teil kann dann die Stabilität grafisch dargestellt werden, mit dem schief-symmetrischen der Wechsel (Greenacre 2000; Blasius 2001). Des Weiteren können Ranking und Rating Daten ebenso analysiert werden wie metrische Daten und Multi-Response-Fragen. In diesen Fällen muss das Eingabeformat der Daten nur so gestaltet werden, dass die Gewichtung der Variablen keinen bzw. nur den gewünschten Effekt auf das Ergebnis hat.

5 Häufige Fehler

Der Vorteil der Korrespondenzanalyse ist, dass man bei der Anwendung so gut wie keine Fehler machen kann und die richtige Interpretation ist meistens nur eine Frage des sorgfältigen Lesens der grafischen Darstellung. Dazu gehört allerdings etwas Übung, um z. B. sofort zu erkennen, wie eine Konfiguration „gelesen“ werden muss. Fehler in der Interpretation können zudem vermieden werden, wenn die Randauszählungen der Variablen standardmäßig einbezogen werden. Zu beachten ist, dass die Interpretation der Ergebnisse relativ zu allen Werten erfolgt, also relativ zum Durchschnitt und nicht in absoluten Zahlen. Schlussfolgerungen wie „sehr hoch“ oder „sehr niedrig“ können allenfalls auf der Basis der Randsummen erfolgen, sie sind kein Ergebnis der Korrespondenzanalyse, stattdessen sollte von „überdurchschnittlich hoch“ oder „relativ niedrig“ gesprochen werden.

Ein zwar einfach zu vermeidender, aber immer wieder vorkommender Fehler ist eine grafische Darstellung, bei der die x-Achse anders als die y-Achse skaliert ist, d. h. die Distanz auf der x-Achse, z. B. 1,0 Skalenpunkte (gemessen in cm), ist ungleich der gleichen Distanz auf der y-Achse, d. h., für 1,0 Skalenpunkte werden in einer Dimension mehr Zentimeter als in der anderen Dimension verwendet. Die Ursache dieses Fehlers ist in der Regel die Verwendungen von Grafikprogrammen wie Powerpoint und die Übernahme von deren Voreinstellungen. Diese liefern zwar in der Regel ein schöneres und seitenoptimiertes Bild, aber eben leider ein fehlerbehaftetes.

Bei der einfachen Korrespondenzanalyse, also bei der Eingabe von einzelnen oder zusammengesetzten Tabellen (ohne Burt-Matrizen), wird zwar in der Regel die symmetrische Darstellung verwendet, aber die Grafik wird ab und zu euklidisch interpretiert, was nicht möglich ist (SPSS 17 erlaubt eine derartige grafische Darstellung daher nicht, bei einer entsprechenden Einstellung werden nur die numerischen Koordinaten gegeben). Bei der multiplen Korrespondenzanalyse kann dieser Fehler nur begangen werden, wenn bei Verwendung der Indikatormatrix die Zeilen, also in der Regel die

Individuen, und die Spalten, also die Variablenausprägungen, in einer symmetrischen Darstellung gemeinsam visualisiert werden. Dieser Fall ist sehr theoretisch, da den einzelnen Personen meistens keine Bedeutung zukommt und sie grafisch nicht dargestellt werden – und wenn, so können deren Lagen im Projektionsraum in einer getrennten grafischen Darstellung wiedergegeben werden (z. B. Bourdieu 1984; Blasius & Mühlichen 2010).

6 Diskussion

Anhand verschiedener Beispiele wurden die wichtigsten Einsatzmöglichkeiten der Korrespondenzanalyse zur Beschreibung von kategorialen Daten diskutiert. Das Verfahren kann auf nahezu beliebige Arten von Daten angewendet werden. In vielen Fällen müssen diese jedoch zuvor in eine geeignete Form gebracht werden – diese Kodierung ist häufig der schwierigste Teil bei der Anwendung der Korrespondenzanalyse.

Der wohl größte Vorteil der Korrespondenzanalyse ist die Visualisierung der Ergebnisse. Komplexe Zusammenhänge zwischen einer Vielzahl von Merkmalen bzw. Merkmalsausprägungen können (meistens) in einer einzigen Abbildung dargestellt werden. Statt einer Vielzahl von Koeffizienten wird die Information konzentriert vermittelt. Eine Eigenschaft, der auch in der Marktforschung eine große Bedeutung zukommen sollte (vgl. Blasius & Mühlichen 2010).

Wie bei allen Datenreduktionsverfahren kann es zu Fehlinterpretationen kommen, wenn Merkmale durch eine höhere Dimension erklärt werden. Um eine derartige Fehlinterpretation zu vermeiden, können entweder die grafischen Darstellungen der höheren Dimensionen gezeigt werden oder es kann auf die numerische Ausgabe der Korrespondenzanalyse zurückgegriffen werden. Mit den numerischen Informationen ist nicht nur eine exakte Zuordnung der Merkmale zu den Achsen möglich, sondern es kann auch angegeben werden, welche Merkmale wie wichtig zur Beschreibung der geometrischen Ausrichtung der Achsen im Projektionsraum sind (vgl. Greenacre 1984, 2007; Blasius 1994, 2001; Le Roux & Rouanet 2004).

Die Korrespondenzanalyse ist ein exploratives Verfahren, statistische Tests sind – in der französischen Tradition – nicht intendiert. Dennoch kann die Korrespondenzanalyse auch als Modell im statistischen Sinn interpretiert werden; ähnlich wie bei der loglinearen Analyse können mit Hilfe der berechneten Parameter die Ausgangsdaten im K-dimensionalen Raum vollständig rekonstruiert werden (vgl. Greenacre 1984; Van der Heijden et al. 1989, 1994).

7 Literaturempfehlungen

Inzwischen gibt es eine ganze Reihe guter Einführungen in die Korrespondenzanalyse, allerdings überwiegend in Englisch oder Französisch. Immer noch aktuell und statistisch relativ anspruchsvoll sind die Einführungen von Greenacre (1984) und Lebart et al. (1984), aber auch jene von Benzécri & collaborateurs (1973) ist durchaus noch lesenswert. Einen sehr guten Überblick über das Verfahren und viele Anwendungen aus

unterschiedlichen inhaltlichen Gebieten gibt Greenacre (2007). Eine deutschsprachige Einführung mit sozialwissenschaftlichen Beispielen gibt Blasius (2001), eine statistisch anspruchvolle, z. T. in der formalen Darstellung leider auch unnötig komplizierte, aber ebenfalls mit vielen sozialwissenschaftlichen Beispielen versehene Einführung geben Le Roux & Rouanet (2004). Eine Vielzahl von Anwendungen aus unterschiedlichen thematischen Gebieten und einige theoretische Grundlagen der Korrespondenzanalyse und benachbarter Verfahren sind in den Readern von Greenacre & Blasius (1994, 2006), sowie von Blasius & Greenacre (1998) enthalten.

Literaturverzeichnis

Benzécri, J.-P. (1979). Sur le calcul des taux d'inertia dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, 4, 377–378.

Benzécri, J.-P. & collaborateurs (1973). *L'analyse des données. L'analyse des correspondances.* Paris: Dunod.

Blasius, J. (1994). Correspondence Analysis in Social Science Research. In M. Greenacre & J. Blasius (Hg.), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications* (S. 23–52). London: Academic Press.

Blasius, J. (2001). *Korrespondenzanalyse.* München: Oldenbourg.

Blasius, J. & Friedrichs, J. (2008). Lifestyles in Distressed Neighborhoods. A Test of Bourdieu's "Taste of Necessity" Hypothesis. *Poetics*, 36, 24–44.

Blasius, J. & Greenacre, M., Hg. (1998). *Visualization of Categorical Data.* London: Academic Press.

Blasius, J. & Mühlichen, M. (2010). Identifying Audience Segments Applying the "Social Space" Approach. *Poetics*, 38, 69–89.

Blasius, J. & Thiessen, V. (2006). Assessing Data Quality and Construct Comparability in Cross-National Surveys. *European Sociological Review*, 22, 229–242.

Blasius, J. & Thiessen, V. (2009). Facts and Artifacts in Cross-National Research: The Case of Political Efficacy and Trust. In M. Haller, R. Jowell, & T. W. Smith (Hg.), *Charting the Globe. The International Social Survey Programme, 1985-2005* (S. 147–169). London: Routledge.

Bourdieu, P. (1982). *Die feinen Unterschiede. Kritik der gesellschaftlichen Urteilskraft.* Frankfurt/M.: Suhrkamp.

Bourdieu, P. (1984). *Homo Academicus.* Frankfurt/M.: Suhrkamp.

Bourdieu, P. (1991). Inzwischen kenne ich alle Krankheiten der soziologischen Vernunft. Pierre Bourdieu im Gespräch mit Beate Krais. In P. Bourdieu, J.-C. Chamboreden, J.-C. Passeron, B. Krais, & H. Beister (Hg.), *Soziologie als Beruf* (S. 269–284). Berlin: Walter de Gruyter.

Eckart, C. & Young, G. (1936). The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1, 211–218.

Friedrichs, J. & Blasius, J. (2000). *Leben in benachteiligten Wohngebieten.* Opladen: Leske + Budrich.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* Chichester: Wiley.

Goodman, L. A. (1991). Measures, Models, and Graphical Display in the Analysis of Cross-Classified Data (with Discussion). *Journal of the American Statistical Association*, 86, 1085–1138.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. J. (1988). Correspondence Analysis of Multivariate Categorical Data by Weighted Least Squares. *Biometrika*, 75, 457–467.

Greenacre, M. J. (2000). Correspondence Analysis of Square Asymmetric Matrices. *Applied Statistics*, 49, 297–310.

Greenacre, M. J. (2007). *Correspondence Analysis in Practice*. Boca Raton: Chapman & Hall.

Greenacre, M. J. & Blasius, J., Hg. (1994). *Correspondence Analysis in the Social Sciences. Recent Developments and Applications*. London: Academic Press.

Greenacre, M. J. & Blasius, J., Hg. (2006). *Multiple Correspondence Analysis and Related Methods*. Boca Raton: Chapman & Hall.

Greenacre, M. J. & Pardo, R. (2006). Multiple Correspondence Analysis of Subsets of Response Categories. In M. J. Greenacre & J. Blasius (Hg.), *Multiple Correspondence Analysis and Related Methods* (S. 197–217). Boca Raton: Chapman & Hall.

Heiser, W. J. & Meulman, J. J. (1994). Homogeneity Analysis: Exploring the Distribution of Variables and their Nonlinear Relationship. In M. Greenacre & J. Blasius (Hg.), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications* (S. 179–209). London: Academic Press.

Le Roux, B. & Rouanet, H. (1998). Interpreting Axes in Multiple Correspondence Analysis: Method of the Contributions of Points and Deviations. In J. Blasius & M. Greenacre (Hg.), *Visualization of Categorical Data* (S. 197–220). San Diego: Academic Press.

Le Roux, B. & Rouanet, H. (2004). *Geometric Data Analysis*. Amsterdam: North Holland.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: Wiley.

Michailidis, G. & de Leeuw, J. (1998). The Gifi System for Descriptive Multivariate Analysis. *Statistical Science*, 13, 307–336.

Rouanet, H., Ackermann, W., & Le Roux, B. (2000). The Geometric Analysis of Questionnaires: The Lesson of Bourdieu's 'La Distinction'. *Bulletin de Méthodologie*, 65, 5–18.

Van der Heijden, P. G. M., de Falguerolles, A., & de Leeuw, J. (1989). A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Loglinear Analysis. *Applied Statistics*, 38, 249–292.

Van der Heijden, P. G. M., Mooijaart, A., & Takane, Y. (1994). Correspondence Analysis and Contingency Table Models. In M. Greenacre & J. Blasius (Hg.), *Correspondence Analysis in the Social Sciences. Recent Developments and Applications* (S. 79–111). London: Academic Press.

Van Rijckevorsel, J. (1987). *The Application of Fuzzy Coding and Horseshoes in Multiple Corresondence Analysis*. Leiden: DSWO Press.