Einführung in die Korrespondenzanalyse

U. Mortensen

.

FB Psychologie und Sportwissenschaften, Institut III
Westfälische Wilhelms-Universität Münster

Überarbeitete Version: 27. 10. 2021

Inhaltsverzeichnis

1 Einführung

Gegeben sei eine Kontingenztabelle $K = (n_{ij})$, $i = 1, \cdots, I$, $j = 1, \cdots, J$, d.h. es gebe I Zeilenkategorien R_i, und J Spaltenkategorien S_j. Tabelle 6 ist ein Beispiel; die Daten wurden zur Illustration der kretschmerschen Persönlichkeitstheorie erhoben, derzufolge bestimmte Charakteristika des Körperbaus zu bestimmten psychologischen Eigenschaften einer Person korrespondieren. Kommt es bei einer Person zu einer psychischen Erkrankung, so entspricht sie dem deutschen Psychiater Ernst Kretschmer (1888 – 1964) zufolge mit hoher Wahrscheinlichkeit dem Körperbau des Patienten. Die Tabelle wurde 1931 erhoben: in allen deutschen Landeskrankenhäusern wurden Patienten einerseits hinsichtlich ihres Körperbaus und andererseits gemäß ihrer Erkrankung klassifiziert. Mit einem Gesamtumfang von 8099 Fällen ist die Stichprobe recht groß, so dass man auf eine gewisse Repräsentativität der Stichprobe hoffen kann. Nach einem systemati-

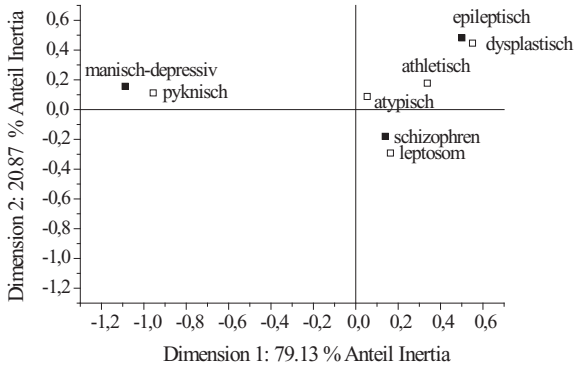
Tabelle 1: Körperbau und psychische Erkrankung: Die kretschmersche Theorie und empirische Häufigkeiten (Westphal 1931), zitiert nach Hofstätter (1971)

	Erkrankung		
Typ	man./dep.	Epilepsie	Schizophr.
pyknisch	879	83	717
athletisch	91	435	884
leptosom	261	378	2632
dysplastisch	15	444	550
atypisch	115	165	450

schen Zusammenhang (nach Kretschmer sollten die Pykniker manisch-depressiv, die Leptosomen schizophren und und die Athleten bzw Dysplastiker Epileptiker sein) sehen die Häufigkeiten zunächst einmal nicht aus, aber der χ^2-Wert der Tabelle beträgt $\chi^2 = 2641.559$ bei $(I-1)\times(j-1) = 8$ Freiheitsgraden und ist somit hochsignifikant: die Daten suggerieren demnach die Existenzt irgendwelcher Abhängigkeiten zwischen den physischen und den psychiatrischen Kategorien. Die Korrespondenzanalyse liefert dazu ein überraschendes Bild, s. Abbildung 1. Hier ist alles so, wie es nach Kretschmer sein soll: die Kategorien 'manisch-depressiv' und 'pyknisch', 'schizophren' und'leptosom' sowie 'epileptisch' und 'dysplastisch' werden durch eng beieinander liegende Punkte repräsentiert, wobei die Koordinatenachsen "latente Variablen" abbilden, deren Bedeutung man hermeneutisch aus der Abbildung herausfiltern muß. Für einen Anhänger der kretschmerschen Typenlehre wird die Abbildung wie eine wundersame Bestätigung dieser Lehre wirken, obwohl sie in der heutigen Psychiatrie und klinischen Psychologie als überholt gilt[1]. Die Kritik an der Datenerhebung ist aber für die Funktionsweise

[1]Wenn Westphal ein Anhänger der Theorie Kretschmers war, so liegt es nahe, dass er seine Klassifizierung unter dem Eindruck der Theorie vorgenommen hat, also eine *petitio principii* begangen hat, wofür man im Deutschen die schöne Übersetzung "Erschleichung des Beweises" findet. Erkrankungen wie Schizophrenie treten eher in jüngeren Jahren auf, wenn man noch

Abbildung 1: Zuordnungen von Körperbautypen und psychischen Erkrankungen (Biplot): Kretschmertypen nach Westphal (1931). Aus Hofstätter (1971).



der Korrespondenzanalyse nicht wesentlich; die Analyse bezieht sich zunächst auf die Daten so, wie sie vorliegen, und die Kompatibiltität von Daten und Theorie ist bekanntlich noch kein Beweis für die Wahrheit der Theorie.

Die CA[2] basiert auf der Singularwertzerleung (SVD)[3] einer Matrix, hier speziell einer Datenmatrix, deren unmittelbare Anwendung die Hauptkomponentenmethode (PCA – Principal Component Anlysis) als Methode der Datenkompression und als Approximation an die Faktorenanalyse ist. Während die PCA eine Zerlegung der Gesamtvarianz der Daten impliziert, zielt die CA auf eine Zerlegung des Gesamt-χ^2 der Kontingenztabelle in voneinander unabhängige χ^2-Komponenten. Die Darstellung der CA als Methode der Datenanalyse setzt eine gewisse Bekanntschaft mit Ergebnissen der Linearen Algebra[4], insbesondere der SVD voraus, weshalb die SVD im folgenden Abschnitt kurz hergeleitet wird.

Die PCA/SVD ist eine Analyse eines in Matrixform vorliegenden Datensatzes in Bezug auf latente Dimensionen: die m Zeilen der Matrix stehen für "Beobachtungen" oder "Fälle", etwa Personen, an denen jeweils n Variablen gemessen wurden. Die Spalten der Matrix repräsentieren jeweils eine Variable. X_{ij} ist der Messwert bei der i-ten Person für die j-te Variable. Man findet gewöhnlich Korrelationen zwischen den Variablen; für die Korrelation r_{jk} zwischen der j-ten und

schlank ist, während manisch-depressive Störungen eher später auftreten, wenn sich bereits Rundungen ergeben haben, – man hat es bei den Daten also möglicherweise mit einer Konfundierung von Alterungsprozesen und Typen zu tun, etc. Außerdem ist nicht klar, ob Epilepsie nicht eher als neurologische denn als psychische Störung zu betrachten ist. Im Wikipedia-Artikel über E. Kretschmer findet sich der Link http://gigi-online.de/kretschmer20.html, in dem noch einmal die kretschmersche Theorie als elaborierter Nonsense ausgewiesen wird.

[2]Die übliche Abkürzung für *Correspondence Analysis*

[3]Abnkürzung für *Singular Value Decomposition*

[4]Man sollte mit dem Begriff des Vektors und die grundlegenden Verknüpfungen von Vektoren kennen. Die Begriffe des Skalarprodukts, der Linearkombination und der linearen Unabhängigkeit sollten hinreichend vertraut sein. Ebenso sollte der Begriff des Rangs einer Matrix und die Operationen der Verknüpfung von Matrizen bekannt sein.

der k-ten Variablen gilt

$$r_{jk} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{x}_j)(X_{ik} - \bar{x}_k)}{s_j s_k} \tag{1.1}$$

Dabei sind[5]

$$\bar{x}_j = \frac{1}{n}\sum_{i=1} X_{ij}, \quad \bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ik} \tag{1.2}$$

die arithmetischen Mittelwerte der Messwerte für die j-te und die k-te Variable und

$$cov(j,k) = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{x}_j)(X_{ik} - \bar{x}_k) \tag{1.3}$$

ist die Kovarianz zwischen den Messwerten für die j-te und k-te Variable. Für $j = k$ geht die Kovarianzformel über in die Varianzformel:

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{x}_j)^2, \quad s_k^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ik} - \bar{x}_k)^2 \tag{1.4}$$

und die Wurzeln s_j und s_k sind die Standardabweichungen ("Streuungen") der jeweilige Messwerte. Es wird angenommen, dass die Korrelationen durch die Wirkung nicht direkt gemessener, gewissermaßen unterliegender, "latenter" Variablen zustande kommen, die unabhängig voneinander sind in dem Sinne, dass die Korrelationen zwischen den latenten Variablen alle gleich Null sind. Wie im folgenden Abschnitt kurz erläutert werden wird besteht die PCA im Wesentlichen in einer Rotation der durch die gemessenen Variablen $V_1, \ldots . V_n$ definierten Koordinatenachsen in Achsen $L_1, \ldots, L_n$ derart, dass die Projektionen der Datenpunkte auf die Achse L_1 die maximal mögliche Varianz haben, die auf L_2 die zweitmaximale Varianz, etc., vergl. Abbildung 2, Seite 10. Betrachtet man die Summe dieser Varianzen als Gesamtvarianz, so kann man sagen, dass die PCA eine Zerlegung der Gesamtvarianz der Daten in voneinander unabhängige Varianzkomponenten besteht. Die CA basiert in einer noch zu klärenden Weise auf einer SVD, bei der die latenten Achsen so skaliert werden, dass sie einer Zerlegung des Gesamt-χ^2 der Tabelle in voneinander unabhängige χ^2-Komponenten entsprechen. Diese Komponenten charakterisieren dann den Beitrag, den die latenten Dimensionen zur Erklärung der Abhängigkeiten zwischen Zeilen- und Spaltenkategorien liefern.

2 Die Singularwertzerlegung einer Matrix

Gegeben sei eine (m, n)-Datenmatrix X, deren Zeilen Fälle oder "Beobachtungen" und deren Spalten gemessene Variablen repräsentieren. Die Elemente x_{ij}

[5] Im Allgemeinen wird bei der Berechnung von Stichprobenvarianzen und -kovarianzen durch $n-1$ statt durch n dividiert, womit eine *systematische* Unterschätzung ("Bias") dieser Statistiken bewirkt wird, die andernfalls insbesondere bei kleineren Stichprobenumfängen auftreten kann. Damit gelingt die Korrektur eines Bias, d.h. eines systematischen Fehlers bei der Schätzung dieser Parameter. Darauf muß hier nicht eingegangen werden, – die Division durch n entspricht der Definition von Varianzen und Kovarianzen als Mittelwerte (i) der $(X_{ij} - \bar{x}_j)^2$, (ii) der Produkte $(X_{ij} - \bar{x}_j)(X_{ik} - \bar{x}_k)$.

von X seien spaltenzentriert, d.h. $x_{ij} = X_{ij} - \bar{x}_j$, oder die X_{ij} sind standardisiert worden: $z_{ij} = (X_{ij} - \bar{x}_j)/s_j$, wobei im Folgenden x_{ij} für einen Messwert geschrieben wird, um den Eindruck zu vermeiden, nur standardisierte Werte könnten in dieser Weise analysiert werden: die SVD kann für eine beliebige Matrix berechnet werden. Dass man für eine PCA die Daten standardisiert ergibt sich aus der Notwendigkeit, den Effekt verschiedener Maßeinheiten zu vermeiden. So könnte etwa V_1 die Körpergröße sein, die man im Prinzip in Millimetern, Zentimetern, Metern oder auch Kilometern messen kann, und V_2 könnte das Körpergewicht sein, das man in Milligramm, Gramm, Kilogramm etc ausdrücken kann. Je nach Wahl der Maßeinheiten nehmen die Kovarianzen verschiedene Werte an, was ihre Interpretation erschwert. Der Übergang zu standardisierten Werten und damit von Kovarianzen zu Korrelationen vermeidet diese Schwierigkeiten.

Die Komponenten des Spaltenvektors $\mathbf{x}_j$ sind die Messungen für die j-te Variable, und die Zeilenvektoren von X werden als Spaltenvektoren $\tilde{\mathbf{x}}_i$ der transponierten Matrix X' angeschrieben:

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{mj} \end{pmatrix}, \quad \tilde{\mathbf{x}}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{in} \end{pmatrix}, \tag{2.1}$$

Linearkombinationen Es seien $\mathbf{x}_1, \ldots, \mathbf{x}_p$ n-dimensionale Vektoren, d.h. alle dise Vektoren mögen n Komponenten haben. Es gelte etwa

$$\mathbf{x}_1 = a_2\mathbf{x}_1 + a_3\mathbf{x}_3 + \cdots + a_p\mathbf{x}_p \tag{2.2}$$

d.h. $\mathbf{x}_1$ sei als "gewichtete" Summe der übrigen Vektoren darstellbar; die i-te Komponente x_{i1} ist dann die gewichtete Summe der jeweils i-ten Komponenten $a_2x_{i2}, a_3x_{i3}, \ldots, a_px_{ip}$. $\mathbf{x}_1$ ist dann eine *Linearkombination* der $\mathbf{x}_2, \ldots, \mathbf{x}_p$.

Lineare Abhängigkeit und Unabhängigkeit Eine Menge $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ von Vektoren mit gleicher Anzahl von Komponenten heißt *linear abhängig*, wenn einer der $\mathbf{u}_k$ als Linearkombinationen der jeweils übrigen Vektoren $\mathbf{u}_j$, $j \neq k$ dargestellt werden kann. Kann keiner der $\mathbf{u}_k$ als Linearkombination der jeweils übrigen dargestellt werden, so heißen die $\mathbf{u}_k$, $1 \leq k \leq n$ *linear unabhängig*. Im Falle der linearen Unabhängigkeit repräsentiert jeder der Vektoren Informationen, die nicht in den übrigen enthalten sind. Dies bedeutet noch nicht, dass etwa die Korrelationen r_{jk} zwischen den Variablen, die von den $\mathbf{x}_j$ und $\mathbf{x}_k$ repräsentiert werden, notwendig gleich Null sind, sondern nur, dass die Absolutbeträge $|r_k|$ kleiner als 1 sind. Der Fall $r_{jk} = 0$ ist nur möglich, wenn die Vektoren nicht nur linear unabhängig, sondern darüber hinaus orthogonal sind; geometrisch bedeutet dies, dass sie senkrecht aufeinander stehen (*Orthogonalität*).

Rang einer Matrix Es läßt sich zeigen, dass sowohl die Zeilen- wie als auch die Spaltenvektoren als Linearkombinationen von maximal $r \leq min(m, n)$ linear unabhängigen, m-domensionalen Vektoren $\mathbf{u}_k$ bzw. n-dimensionalen $\tilde{\mathbf{u}}_i$ dargestellt

werden können, wobei r der *Rang* der Matrix ist. Diese Vektoren werden im Folgenden *Basisvektoren* genannt. Es wird zunächst $r = \min(m, n)$ angenommen ("voller Rang"), da Datenmatrizen wegen der Messfehler und anderer zufälliger Effekte in den Daten normalerweise den "vollen Rang" $r = \min(m, n)$ haben[6]. Erst in einem zweiten Stadium der PCA wird versucht, die Datenmatrix durch Annahme von $r < min(m, n)$ zu approximieren.

Im Rahmen der PCA wird überdies nicht nur lineare Unabhängigkeit, sondern darüber hinaus Orthogonalität (Rechtwinkligkeit) der Basisvektoren postuliert. Sind $\mathbf{u}_1, \ldots, \mathbf{u}_n$ die Basisvektoren, so bedeutet die (paarweise) Orthogonalität, dass die Skalarprodukte verschiedener Basisvektoren gleich Null sind, $\mathbf{u}_j'\mathbf{u}_k = 0$ für $j \neq k$[7]. So soll etwa für den j-ten Spaltenvektor $\mathbf{x}_j$ – seine Komponenten sind die Messwerte für die j-te Variable – die Gleichung

$$\mathbf{x}_j = v_{j1}\mathbf{u}_1 + v_{j1}\mathbf{u}_2 + \cdots + v_{jn}\mathbf{u}_n, \quad j = 1, \ldots, n \tag{2.3}$$

gelten. Die Koeffizienten v_{jk}, $k = 1, \ldots, n$, sind spezifisch für den j-ten Vektor $\mathbf{x}_j$. Die Datenmatrix besteht aus den nebeneinander angeschriebenen Spaltenvektoren $\mathbf{x}_j$, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. Ebenso kann man die $\mathbf{u}_k$ zu einer Matrix $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ zusammenfassen, und die Koeffizienten $v_{j1}, \ldots, v_{jn}$ zu einem Vektor $\tilde{\mathbf{v}} = (v_{j1}, \ldots, v_{jn})'$, so dass die Darstellung (2.3) von $\mathbf{x}_j$ in der Form

$$\mathbf{x}_j = U\tilde{\mathbf{v}}_j$$

geschrieben werden kann. Fasst man darüber hinaus die $\tilde{\mathbf{v}}_j$ zu einer Matrix $V' = [\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_n]$ zusammen, so ergibt sich für die Matrix X die Darstellung

$$X = UV'. \tag{2.4}$$

Daraus folgt sofort[8]

$$X' = VU', \tag{2.5}$$

Die Gleichung (2.4) bedeutet, dass sich x_{ij}, also der (zentrierte doer standardisierte) Messwert des i-ten Falls bei der j-ten Variablen als Skalarprodukt der latenten Vareiablen darstellen läßt:

$$x_{ij} = \tilde{\mathbf{u}}_i'\tilde{\mathbf{v}}_j. \tag{2.6}$$

Diese Aussage ergibt sich unmittelbar aus der Regeln für die Matrixmultiplikation. Der Vergleich der Gleichungen (2.4) und (2.5) liefert den Grund für die Schreibweise V' in (2.4), wo die Koeffizientenmatrix als eine transponierte Matrix eingeführt wird: in (2.5) tritt sie dann in nicht transponierter Form auf; die Spalten von V enthalten dann die Basisvektoren für die Darstellung der Zeilenvektoren $\tilde{\mathbf{x}}_i$ von X, und die Spaltenvektoren von U' enthalten nun die zugehörigen

[6] Ausführliche Begründung in http://www.uwe-mortensen.de/LineareAlgebraNeua.pdf

[7] Es läßt sich zeigen, dass $\cos\theta = \mathbf{x}'\mathbf{y}/(\|\mathbf{x}\|\|\mathbf{y}\|)$, $\|\mathbf{x}\|$ und $\|\mathbf{y}\|$ die Längen der Vektoren θ der Winkel zwischen ihnen. $\mathbf{x}'\mathbf{y} = 0$ genau dann, wenn $\cos\theta = 0$, und dies ist der Fall, wenn $\theta = \pi/2$, entsprechend 90^o.

[8] Sind A und B zwei Matrizen, wobei die Anzahl der Spalten von A gleich der Anzahl der Zeilen von B ist, so gilt $(AB)' = B'A'$:

Koeffizienten. Die $\mathbf{x}_j$ sind Elemente eine m-dimensionalen Vektorraums, die $\tilde{\mathbf{x}}_i$ sind Elemente eines n-dimensionalen Vektorraumes, die wegen $X = UV'$ und $X' = VU'$ in einer *dualen* Beziehung zueinander stehen, was im allgemeinsten Sinne der Bedeutung heißt, dass die beiden Räume keine voneinander unabhängigen Repräsentationen der Daten in X sind: – hat man die $\mathbf{u}_k$ auf der Basis einer Annahme festgelegt, so hat man damit auch die $\mathbf{v}_k$ festgelegt, und umgekehrt. Ausgeschrieben hat man für U und V

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{pmatrix}, \quad V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix} \tag{2.7}$$

U ist eine (m, n)-Matrix mit den Spalten $\mathbf{u}_k$ für die latenten Dimensionen und den Zeilen $\tilde{\mathbf{u}}_i$ für die Fälle, und V ist eine (n, n)-Matrix mit den Spalten $\mathbf{v}_k$ für die Dimensionen und den Zeilen $\tilde{\mathbf{v}}_j$ für die gemessenen Variablen[9].

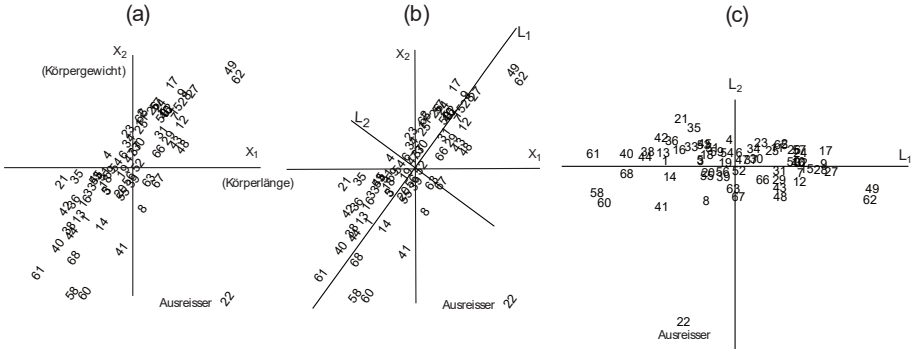
Nach (2.5) hat man für den i-ten Spaltenvektor $\tilde{\mathbf{x}}_i$ – seine Komponenten sind die Messungen der Variablen für den i-ten Fall – die Gleichung

$$\tilde{\mathbf{x}}_i = V\tilde{\mathbf{u}}_i, \tag{2.8}$$

d.h. der Koeffizientenvektor ist der i-te Zeilenvektor von U (= Spaltenvektor von U'). Wie in der Vektoralgebra gezeigt wird kann diese Gleichung als Transformation des Vektors $\tilde{\mathbf{u}}_i$ in den Vektor $\tilde{\mathbf{x}}_i$ aufgefasst werden. V heißt desbalb auch *Transformationsmatrix*. Für eine beliebige (n, n)-Matrix V wird sich $\tilde{\mathbf{x}}_i$ hinsichtlich seiner Orientierung und seiner Länge von $\tilde{\mathbf{u}}_i$ unterscheiden. Die Inspektion von Abbildung 2 zeigt, dass *eine* Möglichkeit, latente Dimensionen zu bestimmen, darin besteht, zB die X_1-Achse des Koordinatensystems so zu rotieren, dass sie mit der Geraden L_1 zusammenfällt, und die X_2-Achse ebenfalls zu rotieren, so dass sie mit der zu L_1 orthogonalen Achse L_2 zusammenfällt. Die Projektionen der Punkte (sie repräsentieren die Fälle!) auf L_1 sind die Koordinaten der Fälle in dem durch L_1 und L_2 gegebenen Koordinatensystem. Diese Projektionen ergeben sich durch eine Rotation der Vektoren $\tilde{\mathbf{x}}_i$, deren Endpunkte die in der Abbildung gezeigten Punkte (hier durch Zahlen i ersetzt) sind. Umgekehrt ergeben sich die $\tilde{\mathbf{x}}_i$ dann durch inverse Rotation aus den $\tilde{\mathbf{u}}_i$. Rotationen sind eine spezielle Form der Transformation, da sie die Längen der transformierten Vektoren invariant lassen. Wie sich zeigen läßt muß die Transformationsmatrix *orthonormal* sein, damit sie eine Rotation bewirkt, d.h. sowohl verschiedene Spalten- wie die Zeilenvektoren sind paarweise *ortho*gonal, und alle Vektoren sind auf die Länge 1 *normiert*. Wenn die Matrix V in (2.8) orthonormal ist, so gilt $V'V = VV' = I_n$, I_n die (n, n)-Identitäts- oder Einheitsmatrix, deren Zeilen und Spalten aus den Einheitsvektoren $\mathbf{e}_j$ bestehen, deren Komponenten alle gleich Null sind mit Ausnahme des j-ten Elements, das gleich 1 ist. Multipliziert man die Gleichung (2.8)

[9] Vergl. den Abschnitt über Lineare Unabhängigkeit, Satz über die Eindeutigkeit der Koeffizienten bei Linearkombinationen lin. unabhängiger Vektoren, in http://www.uwe-mortensen.de/LineareAlgebraNeua.pdf.

Abbildung 2: Konfiguration von Fällen im ursprünglichen X_1, X_2-Koordinatensystem. In (b) sind mögliche latente Variable eingezeichnet worden: L_1 hat die Orientierung der maximalen Ausdehnung der Konfiguration, L_2 ist orthogonal zu L_1 und repräsentiert die Orientierung mit im allgemeinen zweitgrößter Ausdehnung der Konfiguration. (c) zeigt die Konfiguration im Koordinatensystem (L_1, L_2); die Koordinaten in diesem System sind die Projektionen der Punkte im ursprünglichen System auf die Achsen L_1 und L_2. Die Punkte werden durch Zahlen repräsentiert, um die Identifikation der Punkte im rotierten System zu erleichtern. Der Ausreisser 22 wurde bei der Bestimmung von L_1 und L_2 *nicht* berücksichtigt, weil er wegen seiner *Hebelwirkung* (leverage) die optimale Bestimmung dieser Achsen verhindert hätte.



von links mit V', so führt dies wegen der Orthonormalität zu der Beziehung

$$V'\tilde{\mathbf{x}}_i = V'V\tilde{\mathbf{u}}_i = I\tilde{\mathbf{u}}_i = \tilde{\mathbf{u}}_i. \tag{2.9}$$

Hat man also die Matrix V auf irgendeine Weise bestimmt, so können nach (2.9) die Koordinaten, d.h. die Komponenten von $\tilde{\mathbf{u}}_i$ des i-ten Falles im Koordinatensystem (L_1, L_2) bestimmt werden. Die Komponenten des ersten Spaltenvektors $\mathbf{u}_1$ von U sind die Koordinaten u_{i1}, $i = 1, \ldots, m$ der Fälle auf der ersten "latenten' Achse, und die Komponenten des zweiten Spaltenvektors $\mathbf{u}_2$ von U sind die Koordinaten der Fälle u_{i2} auf der zweiten latenten Achse, etc., und die Orthogonalität der Achsen L_1 und L_2 bedeutet $\mathbf{u}_1'\mathbf{u}_2 = 0$; man hat also mit der Wahl der Achsen L_k, $k = 1, 2$, implizit die Orthogonalität der Vektoren $\mathbf{u}_k$ gefordert. Zusammenfassend hat man also die

Annahmen:

A1: V repräsentiert eine Rotation, d.h. $V'V = I$,

A2: Die Vektoren $\mathbf{u}_k$ sind paarweise orthogonal, d.h. $U'U = \Lambda$ ist eine Diagonalmatrix, d.h. im allgemeinen Fall $(n > 2)$ ist $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

Die λ_k sind die Skalarprodukte der $\mathbf{u}_k$ mit sich selbst:

$$\lambda_k = \mathbf{u}_k'\mathbf{u}_k = \|\mathbf{u}_k\|^2 \geq 0, \quad k = 1, \ldots, n \tag{2.10}$$

λ_k ist also das Quadrat der Länge von $\mathbf{u}_k$, und wenn die Daten spaltenzentriert sind, so sind auch die $\mathbf{u}_k$ zentriert, so dass $\|\mathbf{u}_k\|^2$ proportional zur Varianz der Koordinaten u_{ik} der Fälle auf der k-ten latenten Dimension ist. Aus (2.4), also

der Beziehung $X = UV'$, ergibt sich wegen A1 nach Multiplikation von rechts mit V

$$XV = U, \tag{2.11}$$

d.h. die unbekannten Vektoren $\mathbf{u}_k$ erweisen sich als Linearkombinationen der Spaltenvektoren $\mathbf{x}_j$ von X. Weiter folgt $U' = (XV)' = V'X'$, so dass

$$V'X'XV = U'U = \Lambda. \tag{2.12}$$

folgt. Dann ergibt sich aber wegen A1, d.h. $VV' = I_n$, nach Multiplikation von links mit V die Beziehung

$$X'XV = V\Lambda. \tag{2.13}$$

Für einen bestimmten Spaltenvektor $\mathbf{v}_k$ von V erhält man dann

$$X'X\mathbf{v}_k = \lambda_k\mathbf{v}_k, \quad k = 1, \ldots, n \tag{2.14}$$

Eigenvektoren und Eigenwerte: Man kann die Matrix $C = X'X$ als Transformationsmatrix für $\mathbf{v}_k$ auffassen, und (2.14) besagt, dass C den Vektor $\mathbf{v}_k$ in den Vektor $\lambda_k\mathbf{v}_k$ überführt, d.h. in einen Vektor mit derselben Orientierung wie $\mathbf{v}_k$, dessen Länge sich aber um den Faktor λ_k von $\mathbf{v}_k$ unterscheidet. Die Vektoren $\mathbf{v}_k$ sind charakteristisch für C, und deshalb heißen sie *charakteristische Vektoren* oder *Eigenvektoren* von C. Sie lassen sich numerisch bestimmen, worauf hier nicht weiter eingegangen werden muß bzw. kann (sie werden mit dem Computer berechnet). Damit hat man V auf der Basis der Annahmen A1 und A2 bestimmt, und wegen $XV = U$ hat man damit auch U bestimmt: ein Vektor $\mathbf{u}_k$ von U ist eine Linearkombination $\mathbf{u}_k = X\mathbf{v}_k$ der Datenvektoren $\mathbf{x}_j$ von X.

Geometrische Eigenschhaften der Eigenvektoren Es kann gezeigt werden[10], dass der Eigenvektor $\mathbf{v}_1$ im (X_1, X_2)-Koordinatensystem die Orientierung der Achse L_1 angibt, in der die Punktekonfiguration der Fälle die maximale Ausdehnung hat. Diese Ausdehnung ist durch $\lambda_1 = \|\mathbf{u}_1\|^2$ gegeben; die Komponenten von $\mathbf{u}_1$ sind die Koordinaten der Punkte auf L_1. $\mathbf{v}_2$ wiederum bestimmt die Orientierung der Achse L_2, in der die Punktekonfiguration der Fälle die zweitgrößte Ausdehnung hat. Die Ausdehung der Konfiguration in dieser Richtung ist durch $\lambda_2 = \|\mathbf{u}_2\|^2$ gegeben; die Komponenten von $\mathbf{u}_2$ sind die Koordinaten der Punkte auf L_s, etc. Sind die Messwerte in X spaltenzentriert, so ist der Mittelwert jeder Spalte von X gleich Null. Diese Eigenschaft vererbt sich auf die $\mathbf{u}_k$. Die $\lambda_k = \|\mathbf{u}_k\|^2$ entsprechen dann den Varianzen (bis auf den Faktor $1/m$) der Koordinaten. Die Achsen L_k repräsentieren dann eine Zerlegung der durch

$$Var_{ges} = \sum_{k=1}^{n} \lambda_k \tag{2.15}$$

gegebenen Gesamtvarianz der Daten.

Es kann gezeigt werden, dass die symmetrische Matrix $C = X'X$ (es ist ja $C' = (X'X)' = X'X'' = X'X$) eine Menge von Ellipsoiden mit identischer

[10] http://www.uwe-mortensen.de/LineareAlgebraNeua.pdf, p. 74

Orientierung definiert; jeder Punkt der Punktekonfiguration liegt auf einer dieser Ellipsen. Die Hauptachsen dieser Ellipsoide haben die Orientierungen der L_k. Da die ursrpünglichen Koordinatenachsen in die Achsen L_k transformiert (= rotiert) werden, wird der hier besprochene Ansatz zur Bestimmung latenter Variablen auch als *Hauptachsentransformation* bezeichnet.

Die Singularwertzerlegung (SVD)[11] Ein Vektor $\mathbf{x}$ wird *normiert* (genauer: auf die Länge 1 normiert), wenn man seine Komponenten x_i durch seine Länge $\|\mathbf{x}\|$ dividiert, denn dann gilt

$$\|\frac{1}{\|\mathbf{x}\|}\mathbf{x}\| = \frac{1}{\|\mathbf{x}\|}\|\mathbf{x}\| = 1.$$

Die Länge der $\mathbf{u}_k$ ist durch $\|\mathbf{u}_k\| = \sqrt{\lambda_k} = \lambda_k^{1/2}$ gegeben, denn es ist ja $\lambda_k = \|\mathbf{u}_k\|^2$. Man kann also $\mathbf{u}_k$ normieren, indem man die Komponenten durch $\lambda_k^{1/2}$ dividiert bzw. mit $\lambda_k^{-1/2}$ multipliziert. Bezeichnet man mit $\Lambda^{-1/2}$ eine Diagonalmatrix mit den $\lambda_k^{-1/2}$ als Diagonalelementen, so kann man die Normierung durch das Produkt

$$U\Lambda^{-1/2} = Q \tag{2.16}$$

ausdrücken: Q ist die Matrix der normierten $\mathbf{u}_k$. Multipliziert man diese Gleichung von rechts mit $\Lambda^{1/2}$, so erhält man für U den Ausdruck[12]

$$U = Q\Lambda^{1/2}. \tag{2.17}$$

Substituiert man die rechte Seite für U in der Gleichung $X = UV'$, so erhält man die

Singularwertzerlegung

$$X = Q\Lambda^{1/2}V'. \tag{2.18}$$

Die Diagonalwerte $\lambda_k^{1/2}$, also die Wurzeln aus den Eigenwerten von $X'X$ (und XX', wie gleich gezeigt wird) heißen *Singularwerte*. V ist die Matrix der Eigenvektoren von $X'X$, und Q erweist sich als die Matrix er Eigenvektoren von XX': Bildet man nämlich das Produkt XX', so findet man

$$XX' = Q\Lambda^{1/2}VV'\Lambda^{1/2}Q = Q\Lambda Q' \tag{2.19}$$

woraus wegen der Orthonormalität von Q durch Multiplikation mit Q von rechts sofort

$$XX'Q = Q\Lambda \tag{2.20}$$

folgt. Dies bedeutet, dass die Spaltenvektoren $\mathbf{q}_k$ von Q die Eigenvektoren von XX' sind. Da Λ eine (n, n)-Matrix ist, muß Q eine (m, n)-Matrix sein, d.h. es

[11] Die allgemein gebräuchliche Abkürzung 'SVD' steht für den englischen Ausdruck *Singular Value Decomposition*.

[12] Die Multiplikation einer Matrix M *von rechts* mit einer Diagonalmatrix D, so bedeutet dies die Längenskalierung der Spaltenvektoren $\mathbf{m}_k$ von M mit dem entsprechenden Diagonalelement d_{kk} von D, als $M\mathbf{d}_k = d_k\mathbf{d}_k$, $\mathbf{d}_k$ der k-te Spaltenvektor von D; die KOmponenten dieses Vektors sind alle gleich Null, bis auf die k-te Komponente d_{kk}.

werden nur so viele Eigenvektoren von Q betrachtet, wie es Eigenwerte $\lambda_k \neq 0$ gibt. Diese Eigenwerte sind für XX' und $X'X$ identisch!

Die Elemente u_{ik} der Matrix U in der Darstelllung $X = UV'$ von X sind die Komponenten der Spaltenvektoren $\mathbf{u}_k$ von U . Man kann u_{ik} als *Score* des i-ten Falls auf der k-ten latenten Dimension betrachten (die Rede ist oft von *Faktorscores* der Fälle, wobei aber nicht klar ist, ob damit nicht die Komponenten q_{ik} der normierten Vektoren $\mathbf{q}_k$ gemeint sind, gerade in anwendungsbezogenen Darstellungen der PCA wird oft keine eindeutige, formale Definition gegeben!). Jedenfalls gibt es zwei Arten, die SVD zu interpretieren:

$$X = Q\Lambda^{1/2}V' = \begin{cases} QA', & A = V\Lambda^{1/2}, \quad \text{"Ladungen"} \\ UV', & U = Q\Lambda^{1/2}, \quad \text{"Factorscores"} \end{cases} \tag{2.21}$$

Zusammenfassend kann man also sagen, dass die SVD mit der Matrix V der Eigenvektoren von $X'X$ (Kovarianzen oder Korrelationen zwischen den Variablen) die *Orientierungen* der latenten Variablen liefern; die Komponenten des Spaltenvektors $\mathbf{v}_k$ repräsentieren die Variablen auf der k-ten latenten Dimension, die Komponenten q_{ik} von $\mathbf{q}_k$ repräsentieren die Fälle auf der k-ten latenten Dimension. Aber $\mathbf{q}_k'\mathbf{q}_k = \mathbf{v}_k'\mathbf{v}_k = 1$ für $k = 1, \ldots, n$, d.h. die Repräsentationen von Fällen bzw. Variablen sind normiert, so dass das Ausmaß, in dem die verschiedenen latenten Variablen in die gemessenen Variablen bzw. die einzenen Fälle eingehen aus diesen Repräsentationen noch nicht hervorgeht. Diese Ausmaße ergeben sich durch eine "Gewichtung" der $\mathbf{v}_k$ bzw. $\mathbf{q}_k$. Formal besteht diese Gewichtung in einer Skalierung der Längen entweder der $\mathbf{v}_k$ oder der $\mathbf{q}_k$. So kann man etwa die $\mathbf{v}_k$ skalieren. Die SVD sagt, wie diese Längenskalierung aussieht, wenn man etwa an einer Repräsentation der Variablen interessiert ist: Man bildet die Matrix $A = V\Lambda^{1/2}$. so dass man für den ersten Spaltenvektor $\mathbf{a}_1$ von A den Vektor $\mathbf{a}_1 = \mathbf{v}_1\sqrt{\lambda_1} = \sqrt{\lambda_1}\mathbf{v}_1$ erhält, analog dazu $\mathbf{a}_2 = \sqrt{\lambda_2}\mathbf{v}_2$, etc . Alternativ dazu kann man die Spaltenvektoren $\mathbf{q}_k$ skalieren, indem man das Produkt $U = Q\Lambda^{1/2}$ bildet. Der (zentrierte) Messwert x_{ij} ergibt sich den Regeln der Matrixmultiplikation entsprechend dann als Skalarprodukt

$$x_{ij} = \tilde{\mathbf{q}}_i'\tilde{\mathbf{a}}_j, \text{ oder } x_{ij} = \tilde{\mathbf{u}}_i'\tilde{\mathbf{v}}_j. \tag{2.22}$$

$\tilde{\mathbf{q}}_i$ ist der i-te Zeilenvektor von Q, $\tilde{\mathbf{a}}_j$ ist der j-te Zeilenvektor von A (die Zeilen von A repräsentieren die j-te Variable auf den verschiedenen latenten Dimensionen (man hat also einen Spezialfall von (2.6)). Die Interpretationen von $\tilde{\mathbf{u}}_i$ und $\tilde{\mathbf{v}}_j$ sind analog. Ist $X = Z$ die Matrix der spaltenstandardisierten Messwerte, so ergibt sich für die Ladung a_{jk} der j-ten Variablen auf der k-ten latenten Dimension aus $Z = QA'$ zunächst $Z' = AQ'$ und dann $Z'Q = A$, d.h.

$$\mathbf{z}_j'\mathbf{q}_k = a_{jk} = \|\mathbf{z}_j\|\|\mathbf{q}_k\| \cos\theta_{jk} \tag{2.23}$$

wobei θ_{jk} der Winkel zwischen dem Vektor $\mathbf{z}_j$ und dem latenten Vektor $\mathbf{q}_k$ ist. Die Ladung a_{jk} (die Ausprägung der k-ten latenten Variable in der j-ten gemessenen Variable) entspricht also der "Korrelation" zwischen den standardisierten Messungen der j-ten Variable und der k-ten latenten Dimension. Für $\theta_{jk} = 0$

wird demnach a_{jk} maximal, da $\mathbf{z}_j$ und $\mathbf{q}_k$ dann parallel sind, die standardisierte Messung z_{ij} des i-ten Falls entspricht dann der Ausprägungen q_{ik} des i-ten Falls auf der k-ten latenten Dimension.

Das Skalarprodukt $\mathbf{x}'\mathbf{y}$ zweier Vektoren $\mathbf{x}$ und $\mathbf{y}$ kann stets als Ähnlichkeitsmaß für die beiden Vektoren angesehen werden; da für Skalarprodukte stets die Gleichung

$$\mathbf{x}'\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta_{xy} \tag{2.24}$$

gilt, wobei θ_{xy} der Winkel zwischen den beiden Vektoren ist wird $bx'\mathbf{y}$ also maximal, wenn $\theta_{xy} = 0$ ist, denn dann ist $\cos\theta_{xy} = 1$; andernfalls ist $\cos\theta xy < 1$. Der Messwert x_{ij} wird also gemäß (2.22) maximal (relativ zu ihren Längen), wenn $\tilde{\mathbf{q}}_i$ und $\tilde{\mathbf{a}}_j$ parallele Vektoren sind, sie also dieselbe Orientierung haben, wenn also die Anteile der Merkmale, die durch die latenten Variablen abgebildet werden, beim i-ten Fall genau so verteilt sind wie sie in der j-ten Variablen enthalten sind! Überdies wird klar, dass die Fälle und die Variablen im selben Merkmalsraum dargestellt werden, denn $\tilde{\mathbf{q}}_i$ und $\tilde{\mathbf{a}}_j$ sind ja, wie (2.24) zeigt, Vektoren in ein und demselben Merkmalsraum (analog für $\tilde{\mathbf{u}}_i$ und $\tilde{\mathbf{v}}_j$).

Der Biplot Wenn die Vektoren $\tilde{\mathbf{q}}_i$ und $\tilde{a}_j$ Elemente desselben Vektorraums sind, so liegt es nahe, sie auch simultan in einem Koordinatensystem darzustellen; die Rede ist dann von einem *Biplot*. Dazu werden sowohl die Fälle wie die Variablen in einem durch die ersten beiden latenten Variablen definierten Koordinatensystem repräsentiert, was einerseits stets möglich ist, andererseits aber nur dann ein vollständiges Bild der beiden Punktekonfigurationen ergibt, wenn der Rang r der Datenmatrix gleich 2 ist. Im Falle $r > 2$ ist also Vorsicht geboten; überraschenderweise stellt aber der Fall $r \approx 2$ bei der Analyse von Häufigkeitstabellen, also bei Korrespondenzanalysen sehr häufig einer sehr gute Approximation dar. Bei der PCA tritt dieser Fall seltener ein, auch wenn oft ein überraschend kleiner Wert für r bei relativ großer Anzahl von Variablen gefunden wird[13]

Der Biplot erlaubt die Untersuchung der Beziehungen zwischen Fallen und Variablen und wurde anscheinend zuerst von Gabriel (1971) diskutiert, wie sie sich in den in Gleichung (2.22) definierten Skalarprodukten ausdrückt. Wie schon gesagt ist diese Repräsentation der Daten exakt, wenn der Rang der Datenmatrix gleich 2 ist, andernfalls liefert sie nur eine Approximation. Kleine Winkel zwischen einem Fallvektor und einem Variablenvektor deuten auf eine Ähnlichkeit[14] zwischen den entsprechenden Fällen (z.B. Individuen) und Variablen, d.h. wenn die Ausprägung (Komponenten) eines Falles auf den latenten Variablen proportional zu den Ausmaßen, mit denen die Variablen die latenten Variablen erfassen ist; nach (2.24) ist sie maximal relativ zu den Längen der Vektoren, wenn der Winkel zwischen den Vektoren gleich Null ist, – dann ist der Kosinus des Winkels gleich 1. Der Biplot spielt insbesondere bei der Analyse von Kontingenztabellen eine zentrale Rolle, weshalb im Kapitel über die Korrespondenzanalyse detaillierter

[13] Ein Beispiel aus der Medizin wird in http://www.uwe-mortensen.de/PCANeuAa.pdf diskutiert, vergl. dort die Seite 22 etc.

[14] Zum Skalarprodukt als Audruck der Ähnlichkeit von Vektoren vergl. http://www.uwe-mortensen.de/LineareAlgebraNeua.pdf, p. 23

auf ihn eingegangen wird.

Hauptkomponenten (PCA) Die übliche Anwendung der SVD als PCA fokussiert auf den Fall $X = QA'$, nämlich auf die Repräsentation der Variablen im Koordinatensystem $\mathbf{a}_1, \ldots, \mathbf{a}_n$ der latenten Variablen, wobei die $\mathbf{a}_k$ die Spaltenvektoren von V sind:

$$\mathbf{a}_k = \sqrt{\lambda_k}\mathbf{v}_k, \quad k = 1, \ldots, n \tag{2.25}$$

Die Komponenten a_{jk}, $j = 1, \ldots, n$ sind die *Faktorladungen* der Variablen auf der k-ten latenten Dimension. Sind die x_{ij} standardisierte Messwerte (also $x_i = z_{ij}$, wobei eine Spaltenstandardisierung $(X_{ij} - \bar{x}_j)/s_j$ gemeint ist), so ist $X'X = mR$, R die Matrix der Korrelationen zwischen den Variablen. Aus (2.21) folgt dann

$$\frac{1}{m}X'X = R = \frac{1}{m}AQ'QA' = \frac{1}{m}AA' \tag{2.26}$$

Für die Korrelation $r_{jj'}$ zwischen der j-ten und der j'-ten Variablen erhält man demnach

$$r_{jj'} = \tilde{\mathbf{a}}'_j\tilde{\mathbf{a}}_{j'} = \sum_{k=1}^{n} a_{jk}a_{j'k} = \sum_{k=1}^{n} \lambda_k v_{jk}v_{j'k} \tag{2.27}$$

Da $r_{jj'} = 1$ folgt insbesondere

$$\frac{1}{m}\sum_{k=1}^{n} a_{jk}a_{jk} = \sum_{k=1}^{n} a_{jk}^2 = \frac{1}{m}\sum_{k=1}^{n} \lambda_k v_{jk}^2 = 1 \tag{2.28}$$

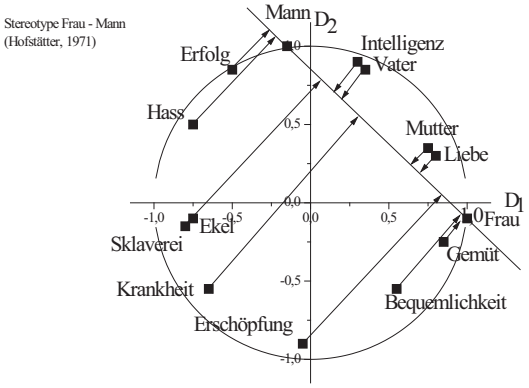
d.h.

$$\frac{1}{m}\|\tilde{\mathbf{a}}_j\|^2 = 1, \tag{2.29}$$

was bedeutet, dass die Variablen durch Punkte auf einer n-dimensionalen Hyperkugel liegen. Kann A durch eine Matrix vom Rang 2 approximiert werden, so liegen die Variablen auf einem Kreis.

Die Abildung 3 illustriert diesen Sachverhalt. Die Variablen sind hier Begriffe, die in Bezug auf die Ausprägung von elementaren Prädikaten (klein, gesund, ...) auf Rating-Skalen beurteilt wurden. Diese Prädikate spielen die Rolle der "Fälle", die beurteilten Begriffe die der "Variablen". Die PCA lieferte insgesamt drei latente Variablen, d.h. die Begriffe können durch Punkte auf einer Kugel repräsentiert werden, wobei aber die dritte Dimension kaum zur Gesamtvarianz der Daten beiträgt. Die Darstellung der Begriffe in einem 2-dimensionalen, durch die ersten beiden latenten Variablen definierten Koordinatensystem erzeugt gewissermaßen seitlichen Blick in die Kugel: die Punkte liegen innerhalb eines Kreises, aber dicht am Kreis. Diese Konfiguration ergibt sich, wenn die latente Struktur im wesentlichen 2-dimensional ist und die dritte nur durch "Fehler" oder "Rauschen" in den Messwerten erzeugt wird. Dem Befund nach sind die Begriffe 'männlich' und ´'weiblich' semantische Basisdimensionen, im Prinzip lassen sich die übrigen Begriffe durch die Ausprägungen auf diesen beiden Dimensionen – also durch Linearkombinationen dieser Basisvektoren – definieren. Der Polaritätsgedanke ist andererseits insbesondere in der geisteswissenschaftlich orientierten Psychologie als hermeneutisches Prinzip beliebt; schon Goethe soll dieses Prinzip

Abbildung 3: 2-dimensionale Struktur von Stereotypen nach Hofstätter (1971); die eingezeichnete Gerade repräsentiert die seit Goethe popularisierte Polaritätstheorie der Geschlechter: männlich versus weiblich. Obwohl die eigentliche Struktur der Stereotype 2-dimensional ist, scheint eine 1-dimensionale Struktur, eben die Polarität der Geschlechter die Daten ebenfalls zu erklären, – aber nur in Bezug auf bestimmte Begriffe, aber eben nicht ganz, da sie die Gesamtheit der Daten nicht erklärt.



formuliert haben. Diesem Prinzip zufolge sind die Begriffe 'männlich" und 'weiblich' polare Gegensätze. Zur Illustration sind die beiden Punkte 'Mann' und 'Frau' durch eine Gerade verbunden worden und die Punkte der übrigen Begriffe werden auf diese Gerade projiziert; der gesamte 2-dimensionale Raum wird auf einen 1-dimensionalen Teilraum projiziert. Es ergibt sich ein plausibles Bild, so dass die Polaritätstheorie als mit den Daten kompatibel erscheint. Gleichwohl ist die Polaritätstheorie in einem grundsätzlichen Sinne falsch, weil sie verdeckt, dass die Ausprägungen von 'männlich' und 'weiblich' im semantischen Raum *unabhängig* (im Sinne von unkorreliert) voneinander variieren, während die Polaritätstheorie eine deterministische Kovariation der Merkmale 'männlich' und 'weiblich' voraussagt: mehr 'weiblich' bedeutet automatisch weniger 'männlich' und umgekehrt. Gerade eine solche Abhängigkeit wird vom 2-dimensionalen Modell *nicht* vorhergesagt. Weitere Ausführungen bzw. Illustrationen der geisteswissenschaftlichen Polaritätstheorie findet man in Welleks (1966) *Schichtentheorie der Persönlichkeit*, die im Skriptum "Erklären und Verstehen"[15] zu finden sind.

Der Fall $X = U'V$ wird analog diskutiert; üblicherweise zielt er auf die Möglichkeit, Typen von Fällen zu finden.

[15] http://www.uwe-mortensen.de/RingvorlesungNeu.pdf, p. 18

3 Korrespondenzanalyse

3.1 Der Ansatz

Die zu analysierenden Daten seien nun die Häufigkeiten n_{ij} in einer Kontingenztabelle K der Art der Tabelle 6. Die Zeilenkategorien R_i (R von engl. rows) , $i = 1, \ldots, I$, (Körperbautypen) spielen nun die Rolle der Fälle, und die Spaltenkategorien S_j, $= 1, \ldots, J$ (psychische Erkrankungen) sind die "Variablen". Die Frage ist, ob es Beziehungen zwischen dem Körperbautyp und der Art der Erkrankung gibt. Der übliche Ansatz, die Frage nach der Existenz von Beziehungen zwischen den Zeilen- und den Spaltenkategorien zu stellen, besteht darin die χ^2-Statitik für die Tabelle zu berechnen. Ist der χ^2-Wert hinreichend groß im Sinne von signifikant, wird die Nullhypothese, dass keinerlei Beziehung existiert, verworfen.

Damit ist aber noch nichts über die Struktur der Beziehungen gesagt. Ein erster Ansatz könnte darin bestehen, Schätzungen für bedingte Wahrscheinlichkeiten $P(S_j|R_i)$ und $P(R_i|C_j)$ zu berechnen, wobei $P(S_j|R_i)$ die bedingte Wahrscheinlichkeit ist, dass eine Person vom Körperbautyp R_i aus der Subpopulation der Erkankten eine Erkrankung vom Typ S_j hat, und umgekehrt die zu dieser bedingten Wahrscheinlichkeit "inverse" bedingte Whrscheinlichkeit $P(R_i|S_j)$, nämlich die bedingte Wahrscheinlichkeit, dass eine zufällig gewählte Person mit der Erkrankung S_j auch den Körperbau vom Typ R_i hat. Diese beiden bedingten Wahrscheinlichkeiten sind im allgemeinen verschieden, was eine intuitive Bewertung der Häufigkeiten in Bezug auf Abhängigkeiten zwischen den beiden Kategorienarten erschwert; wegen $P(S_j|R_i) = P(R_i \cap S_j)/P(R_i)$ und $P(S_i|C_j = P(R_i \cap S_j)/P(S_j)$, wobei $\cap$ für "und" steht, folgt[16] die Beziehung

$$P(S_j|R_i) = P(R_i|S_j)\frac{P(S_j)}{P(R_i}, \tag{3.1}$$

wobei $P(R_i)$ und $P(S_j)$ die *unbedingten* Wahrscheinlichkeiten sind, dass eine zufällig gewählte Person vom Körperbautyp R_i bzw. vom Erkrankungstyp S_j ist. Diese Wahrscheinlichkeiten lassen sich anhand der Häufigkeiten aus der Tabelle schätzen: mit $N = \sum_i \sum_j n_{ij}$ hat man

$$p_{ij} = \hat{p}(R_i \cap S_j) = \frac{n_{ij}}{N}, \quad r_i = \hat{p}(R_i) = \frac{1}{N}\sum_j n_{ij}, \quad c_j = \hat{p}(S_j) = \frac{1}{N}\sum_i n_{ij} \tag{3.2}$$

und

$$\hat{p}(R_i|S_j) = \frac{p_{ij}}{c_j}, \quad \hat{p}(S_j|R_i) = \frac{p_{ij}}{r_i} \tag{3.3}$$

Allerdings ist die Diskussion der insgesamt $I \times J$ bedingten Wahrscheinlichkeiten nicht nur mühsam, insbesondere bei größeren Tabellen, sondern auch ineffektiv.

[16] Beide Ausdrücke nach $P(R_i \cap S_j)$ auflösen und dann die Ausdrücke für $P(R_i \cap S_j)$ gleichsetzen.

Tabelle 2: Die Matrix P mit Zeilen- (R_i) und Spaltenkategorien (S_j)

	S_1	S_2	$\cdots$	S_J	Σ
R_1	p_{11}	p_{12}	$\cdots$	p_{1J}	r_1
R_2	p_{21}	p_{22}	$\cdots$	p_{2J}	r_2
$\vdots$			$\cdots$		$\vdots$
R_I	p_{I1}	p_{I2}	$\cdots$	p_{IJ}	r_I
Σ	c_1	c_2	$\cdots$	c_J	1

So kommt die Frage auf, ob für die Daten nicht eine PCA-ähnliche, auf der Singularwertzerlegung beruhende Analyse durchgeführt werden kann; sie würde die Beziehungen zwischen Zeilen- und Spaltenkategorien durch die Wirkung voneinander unabhängiger, additiv wirkender latenter Variablen erklären. Während aber die latenten Variablen bei der PCA die Gesamtvarianz "erklären", sollte bei einer Häufigkeitstabelle die χ^2-Statistik als Maß für die Ausprägung von Abhängigkeiten zwischen den Kategorien sein, das durch die latenten Variablen erklärt wird.

Das χ^2 einer Tabelle ist durch

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(N\,p_{ij} - N r_i c_j)^2}{N\,r_i c_j} \\
&= N\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \qquad (3.4)
\end{aligned}$$

definiert. Offenbar ist

$$\frac{1}{N}\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \qquad (3.5)$$

der bekannte Kontingenzkoeffizient für die Tabelle. Summiert man nur über einen der beiden Indices i bzw. j, so erhält man die partiellen χ^2-Werte

$$\frac{1}{N}\chi_i^2 = \sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \qquad (3.6)$$

$$\frac{1}{N}\chi_j^2 = \sum_{i=1}^{I} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \qquad (3.7)$$

3.1.1 SVD und Skalenwerte

Man kann nun statt der n_{ij} die *Residuen*

$$x_{ij} = \frac{n_{ij} - n_{i.}n_{.j}/N2}{\sqrt{n_{i.}n_{.j}/N}} = \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \tag{3.8}$$

als Elemente einer Matrix X definieren. Der Ausdruck 'Residuum' für ein x_{ij} reflektiert den Sachverhalt, dass x_{ij} das, was von n_{ij} übrigbleibt, wenn der unter H_o erwartete Wert für n_i subtrahiert wird (und dann noch mit dem Reziprokwert von $\sqrt{n_{i.}n_{.j}/N}$ gewichtet wird). Es erweist sich als nützlich, die x_{ij} zu einer Matrix X zusammenzufassen. Dazu wird zunächst der Ausdruck (3.8) für x_{ij} ein wenig umformuliert und die Matrix E mit den Elementen $e_{ij} = r_i c_j$ definiert; dann hat man

$$\begin{aligned} x_{ij} &= \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} \\ &= \frac{1}{\sqrt{r_i}}(p_{ij} - e_{ij})\frac{1}{\sqrt{c_j}} \end{aligned}$$

Die $1/\sqrt{r_i}$ können zu einer Diagonalmatrix $D_r^{-1/2}$ zusammengefasst werden, und die $1/\sqrt{c_j}$ zu einer Diagonalmatrix $D_c^{-1/2}$,

$$D_r^{-1/2} = \begin{pmatrix} 1/\sqrt{r_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{r_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{r_I} \end{pmatrix} \tag{3.9}$$

$$D_c^{-1/2} = \begin{pmatrix} 1/\sqrt{c_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{c_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{c_J} \end{pmatrix} \tag{3.10}$$

Damit ergibt sich

$$X = D_r^{-1/2}(P - E)D_c^{-1/2}. \tag{3.11}$$

$|x_{ij}|$, der Absolutbetrag von x_{ij}, ist also klein, wenn $n_{ij} \approx n_{i.}n_{.j}/N$ ist, und um so größer, je mehr sich n_{ij} und $n_{i.}n_{.j}/N$ voneinander unterscheiden.

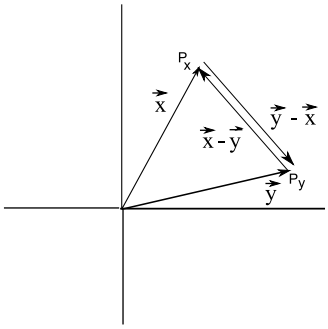
Nach (2.18), Seite 12, ist die SVD von X durch

$$X = Q\Lambda^{1/2}V'$$

gegeben.

Man kann den Zeilenvektor $\tilde{\mathbf{q}}_i$ von Q als Vektor von Koordinaten der i-ten Zeilenkategorie deuten, und ebenso den Vektor $\tilde{\mathbf{v}}_j$ als Vektor von Koordinaten der j-ten Spaltenkategorie, so dass Zeilen- und Spaltenkategorien durch Punkte in einem durch die latenten Variablen definierten Raum repräsentiert werden können.

Abbildung 4: Distanz d_{xy} zwischen zwei Punkten P_x und P_y als Länge der Differenz $\|\mathbf{x}-\mathbf{y}\| = \|\mathbf{y}-\mathbf{x}\|$ zwischen zwei Vektoren



Um allerdings eine Beziehung zwischen einer Kategorie und ihrem Beitrag zum Gesamt-χ^2 herstellen zu können, müssen die q_{ik} und v_{jk} geeignet skaliert werden. Dazu wird zumächst der Begriff der Distanz zwischen Punkten in einem Raum eingeführt.

Geht man von der SVD von X aus, so sind die Zeilenkategorien repräsentierenden Punkte durch die Endpunkte der Vektoren $\tilde{\mathbf{q}}_i$ definiert. Benennt man zwei Vektoren $\tilde{\mathbf{q}}_i$ und $\tilde{\mathbf{q}}_{i'}$ zur Vereinfachung in $\mathbf{x}$ und $\mathbf{y}$ um, so kann man die Distanz zwischen den Endpunkten P_x und P_y mit der Länge $\|\mathbf{x}-\mathbf{y}\|$ gleichsetzen, s. Abbildung 4. Die Komponenten des Vektors $\mathbf{x}-\mathbf{y}$ sind durch die Differenzen $x_i - y_i$ definiert

$$\mathbf{x}-\mathbf{y} = (x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n)' \tag{3.12}$$

und

$$\|\mathbf{x}-\mathbf{y}\|^2 = (\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y}) = \sum_{i=1}^{n}(x_i - y_i)^2 \tag{3.13}$$

d.h. das Quadrat der Länge ist durch den (auf n Dimensionen verallgemeinerten) Satz des Pythagoras gegeben. Der Vektor $\mathbf{x}-\mathbf{y}$ verbindet die Punkte P_x und P_y durch eine Gerade, genauer durch einen Abschnitt einer Geraden, auf der sowohl P_x wie auch P_y liegen. Nach Euklid ist die kürzeste Verbindung zwischen zwei Punkten eine Gerade, weshalb, wenn dieser Distanzbegriff zugrunde gelegt wird, auch von der *euklidischen Distanz* die Rede ist. Ein Raum, in dem die Distanzen zwischen irgendwelchen Punkten durch euklidische Distanzen definiert sind, heißt demnach *euklidischer Raum*. Für euklidische *Distanzen* gilt

$$d(x,y) = d_{xy} = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})} = \|\mathbf{x}-\mathbf{y}\|, \tag{3.14}$$

wobei
(i) $d_{xy} \geq 0$, denn als Wurzel einer Summe von Quadraten kann die Distanz nicht negativ werden. Außerdem muß
(ii) $d(x,y) = d(y,x)$ sein, denn es ist ja $(x_i - y_i)^2 = (y_i - x_i)^2$ für alle i. Schließlich muß
iii $d(x,y) \geq d(x,z) + d(z,y)$ (Dreiecksungleichung) gelten, wie man sofort anhand

der Abb. 4 einsieht, indem man irgendeinen Punkt P_z (der Endpunkt eines Vektors $\mathbf{z}$) in der Ebene wählt, z.B. den Ursprung: $d(x,z)+d(z,y)$ bedeutet ja, dass man einen Umweg über P_z nimmt, um von P_x zu P_y zu kommen. Die Identität $d(x,y) = d(x,z)+d(z,y)$ gilt genau dann, wenn der Punkt $\mathbf{z}$ auf der Geraden liegt, die P_x und P_y verbindet. Von Distanzmaßen, die diese drei Bedingungen erfüllten, sagt man, dass sie eine *Metrik* definieren; zusammenfassend hat man also die Definition einer

Metrik

1. $d(x,y) \geq 0$
2. $d(x,y) = d(y,x)$ (Symmetrie)
3. $d(x,y) \leq d(x,z)+d(z,y)$ (Dreiecksungleichung)

Euklidische Distanzen definieren eine *euklidische Metrik*. Die euklidische Metrik ist keineswegs die einzige Metrik; wenn z.B. der betrachtete Raum in sich gekrümmt ist, gelangt man von einem Punkt P_x zu einem anderen Punkt P_y nur auf gekrümmten Bahnen, so dass die Distanzen nicht mehr als euklidische Distanzen angegeben werden können[17].

Für die Korrespondenzanalyse sollen Koordinaten f_{ik} (i-te Zeilenkategorie, k-t latente Dimension) und g_{jk} (j-te Spaltenkategorie, k-te latente Dimension) derart gefunden werden, dass die Länge der Vektoren, die die Zeilen- bzw. Spaltenkategorien repräsentieren, den Beitrag der jeweiligen Kategorie zum Gesamt-χ^2 entsprechen. Die Koordinaten f_{ik} und g_{jk} definieren dann eine χ^2-*Metrik*.

Zur Vorbereitung dieser Skalierung müssen einige Begriffe eingeführt werden. Die Vektoren

$$\mathbf{r}_i = \begin{pmatrix} p_{i1}/r_i \\ p_{i2}/r_i \\ \vdots \\ p_{iJ}/r_i \end{pmatrix} = \begin{pmatrix} n_{i1}/n_{i+} \\ n_{i2}/n_{i+} \\ \vdots \\ n_{iJ}/n_{i+} \end{pmatrix}, \quad \mathbf{s}_j = \begin{pmatrix} p_{1j}/c_j \\ p_{2j}/c_j \\ \vdots \\ p_{Ij}/c_j \end{pmatrix} = \begin{pmatrix} n_{1j}/n_{+j} \\ n_{2j}/n_{+j} \\ \vdots \\ n_{Ij}/n_{+j} \end{pmatrix} \tag{3.15}$$

heißen i-tes *Zeilenprofil* ($\mathbf{r}_i$) bzw. j-tes *Spaltenprofil* ($\mathbf{s}_j$)[18]. Zur Verdeutlichung: Die p_{ij}/r_i sind bedingte Wahrscheinlichkeiten. So sei ϕ_i ein eine Zeile der Tabelle kennzeichnender Körperbautyp und ψ_j sei eine psychische Erkrankung. Dann ist

$$P(\psi_j|\phi_i) = \frac{P(\phi_i \cap \psi_j)}{P(\phi_i)} = \frac{n_{ij}/N}{n_{i+}/N} = \frac{p_{ij}}{r_i} \tag{3.16}$$

$$P(\phi_i|\psi_j) = \frac{P(\phi_i \cap \psi_j)}{P(\psi_j)} = \frac{n_{ij}/N}{n_{+j}/N} = \frac{p_{ij}}{c_j} \tag{3.17}$$

Man könnte nun die Komponenten der Vektoren $\mathbf{r}_i$ als Koordinaten eines Punktes in einem geeignet gewählten Koordinatensystem betrachten. Die Distanz zwischen den Endpunkten von $\mathbf{r}_i$ und $\mathbf{r}_{i'}$ ist dann in einer euklidischen Metrik

[17] Dies trifft gelegentlich auf psychologische Räume zu, die sich durch multidimensionale Skalierungen von "Objekten" ergeben; die Distanzen zwischen den Objekten repräsntieren dann Unähnlichkeiten zwischen ihnen.

[18] Man sollte also darauf achten, das Zeilenprofil $\mathbf{r}_i$, das also ein Vektor ist, nicht mit der relativen Häufigkeit $r_i = \sum_j n_{ij}/N$ zu verwechseln!

durch die Länge der Vektordifferenz $\mathbf{r}_i - \mathbf{r}_{i'}$ gegeben; das Quadrat der Länge dieses Differenzenvektors ist

$$\Delta^2_{ii'} = \|\mathbf{r}_i - \mathbf{r}_{i'}\|^2. \tag{3.18}$$

Damit hätte man allerdings noch keine Beziehung zum partiellen χ^2_i hergestellt. Es zeigt sich, daß diese Beziehung hergestellt wird, wenn man von den Zeilen- und Spaltenprofilen zu modifizierten Vektoren übergeht (für die Spaltenprofile sind die Betrachtungen analog):

$$\rho_i = \begin{pmatrix} p_{i1}/(r_i\sqrt{c_1}) \\ p_{i2}/(r_i\sqrt{c_2}) \\ \vdots \\ p_{iJ}/(r_i\sqrt{c_J}) \end{pmatrix} \tag{3.19}$$

Der Differenzenvektor $\delta_{ii'} = \rho_i - \rho_{i'}$ hat dann die Komponenten

$$\frac{p_{ij}}{r_i\sqrt{c_j}} - \frac{p_{i'j}}{r_{i'}\sqrt{c_j}} = \frac{1}{\sqrt{c_j}}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right), \quad j = 1, \ldots, J \tag{3.20}$$

Eine Komponente von $\delta_{ii'}$ ist also die Differenz zweier bedingter Wahrscheinlichkeiten $P(\psi_j|\phi_i)$ und $P(\psi_j|\phi_{i'})$ (vergl. (3.16)), gewogen mit dem Faktor $1/c_j$, c_j die (Schätzung der) Wahrscheinlichkeit, bei einer Beobachtung ein Element der j-ten Spaltenkategorie zu finden. Das Quadrat der Länge von $\rho_i - \rho_{i'}$ (also der Distanz der Endpunkte der Vektoren ρ_i und $\rho_{i'}$) ist dann durch

$$\delta^2_{ii'} = \|\rho_i - \rho_{i'}\|^2 = \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}}\right)^2 \tag{3.21}$$

Jetzt werde der Vektor $\mathbf{c} = (c_1, c_2, \ldots, c_J)'$ betrachtet. Es ist $c_j = \sum_i n_{ij}/N = n_{+j}/N$, d.h. c_j ist eine Schätzung der Wahrscheinlichkeit, bei einer Beobachtung ein Element der j-ten Spaltenkategorie zu finden). c_j ist gleichzeitig der Mittelwert der Häufigkeiten n_{ij} über alle i, also über alle Zeilenkategorien, so dass $\mathbf{c}$ gleich dem mittleren Zeilenvektor ist. Dann kann man die Distanz des i-ten Zeilenprofils zu $\mathbf{c}$ betrachten:

$$\delta^2_{i\cdot} = \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - c_j\right)^2 \tag{3.22}$$

Definition 3.1 *$\delta_{ii'}$ heißt* χ^2-Distanz *zwischen den Zeilenkategorien R_i und $R_{i'}$. $\delta_{i\cdot}$ ist die χ^2-Distanz von R_i zum Ursprung des Koordinatensystems.*

Anmerkung Die χ2-Distanz zwischen Spalten S_j und $S_{j'}$ und einer Spalte S_j und dem Koordinatenursprung ist analog definiert. Dazu müssen die Komponenten der Vektoren $\mathbf{s}_j$ (vergl (3.15) noch durch $\sqrt{r_i}$ dividiert werden; $\mathbf{s}_j$ geht dann in einen Vektor σ_j mit den Komponenten $p_{ij}/(c_j\sqrt{r_i})$ über. Die Vektoren ρ_i und

σ_j haben im Allgemeinen nicht dieselbe Anzahl von Komponenten, da $I \neq J$; dies bedeutet, dass *keine Distanz* zwischen einem Zeilen- und einem Spaltenmerkmal berechnet werden kann; diese Distanzen sind nicht definiert.

Es kann nun gezeigt werden (Anhang 6.1.1, Seite 67), dass das χ^2 der Tabelle durch

$$\chi^2 = \sum_{i=1}^{I} r_i \delta_{i\cdot}^2 \tag{3.23}$$

gegeben ist. Damit hat man eine Zerlegung des Gesamt-χ^2 in Komponenten, die durch die Zeilenkategorien erzeugt werden. Die Zerlegung des χ^2 in additive Komponenten, die durch die Spaltenkategorien erzeugt werden, ist analog definiert.

Die Anwendung der SVD Im folgenden Satz werden die Koordinaten f_{ik} und g_{jk} definiert, die die gewünschte Beziehung zu den χ^2-Komponenten liefern:

Satz 3.1 *Es sei $X = Q\Lambda^{1/2}V'$ und q_{ik} sei das Element in der i-ten Zeile und j-ten Spalte von Q. Weiter sei*

$$f_{ik} = q_{ik}\frac{\sqrt{\lambda_k}}{\sqrt{r_i}}, \quad i = 1, \ldots, I, \quad k = 1, \ldots, s \tag{3.24}$$

und

$$d_{ii'} = \left(\sum_{j=1}^{J} (f_{ij} - f_{i'j})^2 \right)^{1/2} \tag{3.25}$$

Dann gilt

$$d_{ii'} = \delta_{ii'} \tag{3.26}$$

Beweis: s. Anhang 6.1.2, Seite 68. □

Die Definition der Skalenwerte für die Spaltenkategorien ist analog zu der für die Zeilenkategorien:

$$g_{jk} = v_{jk}\frac{\sqrt{\lambda_k}}{\sqrt{c_j}}, \quad j = 1, \ldots, J \tag{3.27}$$

v_{jk} ein Element der Matrix V. Dann gilt

$$d_{jj'} = \delta_{jj'} \tag{3.28}$$

Mit der in (3.11) gegebenen Definitionen von $D_r^{-1/2}$ und $D_c^{-1/2}$ lassen sich die die f_{ik} als Elemente einer Matrix F und die g_{jk} als Elemente einer Matrix G angeben, die gemäß

$$F = D_r^{-1/2} Q \Lambda^{1/2} \tag{3.29}$$

$$G = D_c^{-1/2} V \Lambda^{1/2} \tag{3.30}$$

bestimmt werden. Anders als in der PCA werden in der KA sowohl die Koordinaten für die Zeilen wie die für die Spalten mit $\Lambda^{1/2}$ skaliert, darüber hinaus

werden die Zeilen von $Q\Lambda^{1/2}$ mit $D_r^{-1/2}$ und die Zeilen von $V\Lambda^{1/2}$ mit $D_c^{-1/2}$ skaliert. Dieser Sachverhalt reflektiert die Tatsache, dass man bei einer KA nicht entweder an den Zeilen- oder an den Spaltenkategorien interessiert ist, sondern sowohl an den Zeilen- wie auch den Spaltenkategorien interessiert ist. Die Definitionen (3.29) und (3.30) erweisen sich als nützlich bei der Repräsentation der Kategorien in einem Biplot.

3.1.2 Der Biplot

Biplots wurden bereits im Abschhnitt über die PCA im Zusammmenhang mit der Darstellung eines Messwertes x_{ij} als Skalarprodukt von latenten Vektoren eingeführt, vergl. Gleichung (2.22), Seite 13). Ein Biplot ist eine simultane Repräsentation von Zeilen- und Spaltenkategorien durch Punkte in einem gemeinsamen Koordinatensystem. Im Rahmen der Korrespondenzanalyse sollen dabei die Vektoren für die Zeilen- und Spaltenkategorien so gewählt werden, dass der Beitrag der Kategorien zum Gesamt-χ^2 der Tabelle deutlich wird. Diese Darstellung wird erreicht, wenn man die Zeilenkategorien durch die Zeilenvektoren der Matrix F und die Spaltenkategorien durch die Zeilenkategorien der Matrix G repräsentiert. Die "Ähnlichkeit" einer Zeilen- und einer Spaltenkategorie läßt sich dann analog zur Gleichung (2.22) durch Skalarprodukte der Art

$$s_{ij} = \tilde{\mathbf{f}}_i' \tilde{\mathbf{g}}_j = \|\tilde{\mathbf{f}}_i\| \|\tilde{\mathbf{g}}_j\| \cos\theta_{ij} \tag{3.31}$$

ausdrücken. Dabei ist wieder der Winkel zwischen den entsprechenden Vektoren von besonderem Interesse. Denn s_{ij} ist maximal relativ zum Produkt der Längen $\|\mathbf{x}\|$ und $\|\mathbf{y}\|$, wenn $\cos\theta_{xy} = 1$, d.h. wenn $\theta_{xy} = 0$ ist; dann sind die Vektoren parallel zueinander, – man kann auch sagen, dass sie dann auf einer Geraden liegen. Der Biplot wird detailliert in Abschnitt 5.1 illustriert, wobei es um die Beziehung zwischen Körperbau únd psychischer Erkrankung geht. Hier soll noch kurz auf die allgemeine Frage nach der Beziehung zwischen der Datenrepräsentation im Biplot und den bedingten Wahrscheinlichkeiten für die Zeilen- und Spaltenkategorien eingegangen werden

Die s_{ij} lassen sich in einer Matrix

$$S = FG' = (s_{ij}) \tag{3.32}$$

zusammenfassen. Natürlich ist $\tilde{\mathbf{f}}_i' \tilde{\mathbf{g}}_j = \tilde{\mathbf{g}}_j' \tilde{\mathbf{f}}_i$, d.h. die Ähnlichkeit der Vektoren ist eine symmetrische Größe, wie überhaupt die Distanzen im Raum der latenten Variablen durch eine Metrik gekennzeichnet sind, d.h. $d_{ii'} = d_{i'i}$ etc. Andererseits gilt für die bedingten Wahrscheinlichkeiten im Allgemeinen $P(\psi_j|\phi_i) \neq P(\phi_i\psi_j)$, d.h. sie sind asymmetrisch. Also ergibt sich die Frage, in welcher Beziehung bedingte Wahrscheinlichkeiten und Skalarprodukte zueinander stehen.

Da sich die Koordinaten F und G aus der SVD der Matrix X der Residuen ergeben, muß dazu zunächst die Beziehung zwischen F und G einerseits und X bestimmt werden. Nach (3.11)und der Definition der SVD hat man

$$X = D_r^{-1/2}(P - E)D_c^{-1/2} = Q\Lambda^{1/2}V'$$

und wegen (3.29) und (3.30) hat man

$$F = D_r^{-1/2} Q \Lambda^{1/2}, \quad G = D_c^{-1/2} V \Lambda^{1/2}.$$

Dann folgt

$$P - E = D_r^{1/2} Q \Lambda^{1/2} V' D_c^{1/2},$$

d.h. aber man hat die

Rekonstitutionsformel

$$P - E = D_r^{1/2} F \Lambda^{-1/2} G' D_c^{1/2}, \tag{3.33}$$

d.h.

$$X = D_r^{-1/2}(P - E) D_c^{-1/2} = F \Lambda^{-1/2} G' \tag{3.34}$$

Das Residuum x_{ij} ergibt sich also als Skalarprodukt des i-ten Zeilenvektors von $F\Lambda^{-1/2}$ mit dem j-ten Spaltenvektor von G', d.h. dem j-ten Zeilenvektor von G, oder, alternativ dazu, als Skalarprodukt des Zeilenvektors $\tilde{\mathbf{f}}_i$ mit dem j-ten Zeilenvektor von $G\Lambda^{-1/2}$. Zur Vorbereitung sei angemerkt, dass

$$W_{(j,i)} = D_r^{-1} P \tag{3.35}$$

$$W_{(j,i)} = P D_c^{-1} \tag{3.36}$$

die bedingten Wahrscheinlichkeiten $P(\psi_j|\phi_i) = p_{ij}/r_i$ (in $W_{(j,i)}$) und $P(\phi_i|\psi_j) = p_{ij}/c_j$ (in $W_{(j,i)}$) enthalten, wie man durch Ausrechnen der Matrixprodukte auf den rechten Seiten der Gleichungen leicht überprüft.

Eine erste Möglichkeit, die Beziehung zwischen den bedingten Wahrscheinlichkeiten, z.B. denen in $W_{(j,i)}$, darzustellen ergibt sich aus der Rekonstitutionsformel (3.34). Es folgt

$$D_r^{-1/2} P D_c^{-1/2} = F \Lambda^{-1/2} G' + D_r^{-1/2} E D_c^{-1/2}$$

und nach Multiplikation von rechts mit $D_c^{1/2}$ erhält man

$$D_r^{-1/2} P = F \Lambda^{-1/2} G' D_c^{1/2} + D_r^{-1/2} E \tag{3.37}$$

Nochmalige Multiplikation von links mit $D_r^{-1/2}$ liefert schließlich

$$D_r^{-1} P = D_r^{-1/2} F \Lambda^{-1/2} G' D_c^{1/2} + D_r^{-1} E \tag{3.38}$$

für die bedingten Wahrscheinlichkeiten $P(\psi_j|\phi_i)$ in Abhängigkeit von $F\Lambda^{-1/2}G'$. Die Herleitung von PD_c^{-1} ist analog. Hier gehen nicht die Skalarprodukte FG' ein, sondern die Skalarprodukte $F\Lambda^{-1/2}G'$; eine direktere Beziehung zwischen FG' und den bedingten Wahrscheinlichkeiten folgt aus den Übergangsgleichungen, in denen F in Abhängigkeit von G und G in Abhängigkeit von F dargestellt werden:

Die Übergangsgleichungen

$$G = D_c^{-1} P' F \Lambda^{-1/2}, \tag{3.39}$$
$$F = D_r^{-1} P G \Lambda^{-1/2} \tag{3.40}$$

ausgeht (Herleitung in Abschnitt 6.2.1, insbesondere von Gleichung 6.12, von Seite 71 an). Dann folgt

$$\begin{aligned} FG' &= D_r^{-1} P G \Lambda^{-1/2} \Lambda^{-1/2} F' P D_c^{-1/2} \\ &= D_r^{-1} P G \Lambda^{-1} F' P D_c^{-1/2} \end{aligned} \tag{3.41}$$

Setzt man auf der rechten Seite die Ausdrücke (3.29) für F und (3.30) für G ein, so erhält man

$$\begin{aligned} FG' &= D_r^{-1} P D_c^{-1/2} V \Lambda^{1/2} \Lambda^{-1} \Lambda^{1/2} Q' D_r^{-1/2} P D_c^{-1/2} \\ &= D_r^{-1} P (D_c^{-1/2} V Q' D_r^{-1/2}) P D_c^{-1/2} \\ &= D_r^{-1} P (AB') P D_c^{-1/2} \end{aligned} \tag{3.42}$$

mit

$$A = D_c^{-1/2} V, \quad B = D_r^{-1/2} Q. \tag{3.43}$$

Man kann zur Verdeutlichung

$$FG' = W_{(j,i)} (AB') W_{(i,j)} \tag{3.44}$$

schreiben um hervorzuheben, wie die bedingten Wahrscheinlichkeiten $P(\psi_j|\phi_i)$ und $P(\phi_i|\psi_j)$ in die Definition der Skalarprodukte $\tilde{\mathbf{f}}_i' \tilde{\mathbf{g}}_j$ eingehen.

3.2 Weitere Diskussion einer Lösung

Wie bei der Faktorenanalyse wird man versuchen, die Daten mit einem Modell von maximaler Ökonomie, d.h. mit einer kleinstmöglichen Anzahl von Dimensionen zu deuten. Dazu wird zunächst angegeben, in welcher Weise die durch F und G gegebenen Koordinaten zu einer additiven Zerlegung des χ^2 führen.

3.2.1 Die Zerlegung des χ^2

In der Korrespondenzanalyse werden bestimmte Größen nach Größen der Mechanik benannt, so die r_i unf c_j als *Massen*, denen in der Physik gewisse Eigenschaften entsprechen, etwa die Trägheit (Im Lateinischen und Englische *inertia*, die durch das χ^2 bzw. durch ein partielles χ^2 ausgedrückt wird. Der Koordinatenursprung wird auch *Baryzentrum* genannt, weil er mit dem Schwerpunkt einer mit verschiedenen Massen belasteten Ebene entspricht. Mit $In(i)$ wird die *Teilinertia* χ_i^2 bezeichnet, etc.

Für die Zerlegung des χ^2 gelten die folgenden Aussagen (vergl. Anhang, Abschn. 6.1.3), Seite 69), wobei F_k der k-te Spaltenvektor von F ist:

$$F_k^{'} D_r F_k = \lambda_k \tag{3.45}$$

$$In(i) = \frac{\chi_{i\cdot}^2}{N} = \sum_{k=1}^{s} \lambda_k u_{ik}^2 \tag{3.46}$$

$$In(j) = \frac{\chi_{\cdot j}^2}{N} = \sum_{k=1}^{s} \lambda_k v_{jk}^2 \tag{3.47}$$

$$In(K) = \frac{\chi^2}{N} = \sum_{k=1}^{s} \lambda_k \tag{3.48}$$

Die Gesamt-Inertia und damit das Gesamt-χ^2 ist also nach (3.48) durch die Summe der Eigenwerte λ_k von $X'X$ bzw. XX' (die von Null verschiedenen Eigenwerte dieser beiden Matrizen sind gleich groß!) gegeben. Es kann nun eine Reihe von Qualitätsmaßen definiert werden, anhand derer die Güte der Repräsentation der Zeilen- bzw. Spaltenkategorien durch Punkte in einem Raum beurteilt werden kann.

3.2.2 Trägheitsanteil

Definition 3.2 *Der Quotient*

$$\pi_k \stackrel{def}{=} \frac{\lambda_k}{\sum_k \lambda_k} = \frac{\lambda_k}{In(K)} \tag{3.49}$$

heißt Trägheits- *oder* Inertia-Anteil*; π_k ist der Anteil der Gesamt-Inertia, der durch die k-te latente Variable erzeugt wird.*

Anmerkung: In einigen Programmpaketen, z.B. Statistica, wird der Trägheitsanteil als Prozentwert ausgegeben, also als $100\pi_k$.

λ_k charakterisiert die k-te latente Variable, und somit gibt π_k den Anteil an Abhängigkeiten der Tabelle, die "zu Lasten" der k-ten latenten Variablen gehen. Die Bedeutung von π_k ist analog zum Anteil der durch die k-te Achse erklärten Gesamtvarianz bei der Hauptachsentransformation (als Approximation an die Faktorenanalyse) von *Meßwerten.*

3.2.3 Relative Trägheit

Definition 3.3 *Die Quotienten*

$$\rho_{i\cdot} \stackrel{def}{=} \frac{In(i)}{In(K)} \tag{3.50}$$

$$\rho_{\cdot j} \stackrel{def}{=} \frac{In(j)}{In(K)} \tag{3.51}$$

heißen relative Trägheiten *bzw.* relative Inertiae*; $\rho_{i\cdot}$ ist die relative Inertia für die i-te Zeilen, $\rho_{\cdot j}$ ist die für die j-te Spaltenkategorie.*

Die relative Inertia ist der Anteil der Inertia oder Trägheit, die ein Punkt (Zeilen- oder Spaltenpunkt) an der Gesamtträgheit der Tabelle hat, und zwar *unabhängig von der Anzahl der für die Interpretation der Daten angenommen Dimensionen.* Es ist natürlich

$$\rho_{i\cdot} = \frac{\chi^2_{i\cdot}}{\chi^2}, \quad \rho_{\cdot j} = \frac{\chi^2_{\cdot j}}{\chi^2}, \tag{3.52}$$

da sich N bei den Trägheiten herauskürzt. $\chi^2_{i\cdot}$ ist das χ^2 für die i-te Zeile, $\chi^2_{\cdot j}$ ist das χ^2 für die j-te Spalte.

Nach (3.45) der Eigenwert λ_k, $k = 1, \ldots, s$, gerade durch $F_k^{'} D_r F_k$ gegeben; die Komponenten f_{ik} von F_k sind die Koordinaten der Zeilenkategorien R_i auf der k-ten latenten Variablen. Ausgeschrieben heißt (3.45)

$$\lambda_k = f_{k1}^2 r_1 + f_{k2}^2 r_2 + \cdots + f_{kI}^2 r_I \tag{3.53}$$

Setzt man diesen Ausdruck in (3.49) ein, so erhält man

$$\pi_k = \frac{f_{k1}^2 r_1 + f_{k2}^2 r_2 + \cdots + f_{kI}^2 r_I}{In(K)} = \frac{f_{k1}^2 r_1}{In(K)} + \cdots + \frac{f_{kI}^2 r_I}{In(K)}$$

Der Anteil π_k setzt sich demnach additiv wiederum aus den Anteilen

$$\pi_{i\cdot k} \stackrel{def}{=} f_{ik}^2 r_i / In(K), \quad i = 1, \ldots, I \tag{3.54}$$

zusammen; $\pi_{i\cdot k}$ ist der Anteil der Gesamt-Inertia, den die i-te Kategorie R_i bezüglich der k-ten latenten Variablen erzeugt.

Definition 3.4 *Der Anteil $\pi_{i\cdot k}$ heißt* relative Trägheit *oder* relative Inertia *der i-ten Zeilenkategorie für die k-te Dimension.*

Die relative Trägheit $\pi_{\cdot jk}$ für die k-te Dimension der j-ten Spaltenkategorie ist analog definiert.

3.2.4 Qualität

In (**??**) ist die χ^2-Distanz zwischen der i-ten Zeilenkategorie und dem mittleren Zeilenprofil angegeben worden. Nach (**??**) bzw. (3.28) entsprechen die χ^2-Distanzen aber den euklidischen Distanzen zwischen den repräsentierenden Punkten, wenn die Koordinaten durch die Matrizen F und G gegeben sind, also durch f_{ik} und g_{jk}, $k = 1, \ldots, s$. Da man aber eine ökonomische Darstellung sucht, wird man die Anzahl der Dimensionen so klein wie möglich wählen. Es sei s_a die Anzahl der Dimensionen, die man für die Approximation der Daten wählt, $s_a < s$. Die Distanzen zwischen den Punkten bzw. Profilen werden nun im allgemeinen weniger genau reproduziert. Dies führt zu der folgenden

Definition 3.5 *Es sei $\hat{d}_{i\cdot}$ die Distanz der i-ten Kategorie (dem i-ten Profil) zum mittleren Profil, wenn $s_a < s$ Dimensionen für die Darstellung der Kategorien gewählt werden, und $d_{i\cdot}$ sei die entsprechende Distanz, wenn alle s Dimensionen berücksichtigt werden. Dann heißt der Quotient*

$$q_{i\cdot} = \frac{\hat{d}_{i\cdot}}{d_{i\cdot}} \tag{3.55}$$

die Qualität *der approximierenden Repräsentation für die i-te Zeilenkategorie. Die Qualität $q_{\cdot j}$ für die j-te Spaltenkategorie ist analog definiert.*

Benutzt man also alle Dimensionen für die Repräsentation, so sind alle Qualitäten gleich 1.

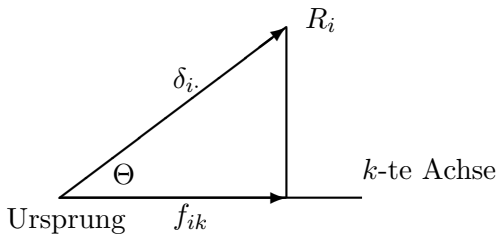
3.2.5 Der $\cos^2$-Anteil

Nach (3.23), Seite 23, ist

$$IN(K) = \sum_{i=1}^{I} r_i \delta_{i\cdot}^2.$$

d.h. die Gesamt-Inertia ist gleich der gewogenen Summe der χ^2-Distanzen der Zeilenkategorien R_i vom mittleren Zeilenprofil. $\delta_{i\cdot}$ ist die χ^2-Distanz der i-ten

Abbildung 5: Projektion von R_i auf die k-te Achse



Kategorie R_i vom Ursprung des Koordinatensystems, und $f_{i1}, \ldots, f_{is}$ sind die Koordinaten der i-ten Kategorie R_i. f_{ik} ist die Projektion des R_i repräsentierenden Punktes auf die Achse, die die k-te latente Variable repräsentiert. Es sei Θ der Winkel zwischen dem Vektor vom Koordinatenursprung zu dem Punkt R_i und dem Vektor vom Ursprung bis zu dem durch f_{ik} definierten Punkt auf der k-ten Achse. Dann ist $\cos\Theta = f_{ik}/\delta_{i\cdot}$. Quadriert man diesen Kosinus und erweitert mit r_i, so erhält man

$$\cos^2(\theta) = \frac{f_{ik}^2}{\delta_{i\cdot}^2} = \frac{f_{ik}^2 r_i}{\delta_{i\cdot}^2 r_i} \tag{3.56}$$

Der Wert von $\cos^2(\Theta)$ gibt also den Anteil an, der von der Koordinate f_{ik} der i-ten Kategorie R_i auf der k-ten Achse an der χ^2-Distanz dieser Kategorie vom Koordinatenurspung erklärt wird. Da $0 \leq \cos^2(\Theta) \leq 1$ ist folgt, daß der Fall

$\cos^2(\theta) = 1$ anzeigt, daß die Kategorie R_i gerade auf der k-ten Achse liegt, also genau durch diese Dimension "erklärt" wird (oder umgekehrt diese Dimension definiert!). Der andere Extremfall, $\cos^2(\Theta) = 0$ bedeutet, daß die k-te Dimension gar nicht in R_i enthalten ist.

3.2.6 Die Verallgemeinerte SVD

Im Anhang wird die *Single Value Decomposition* (SVD) eingeführt. Nach (**??**), Seite **??** im Anhang gilt für eine reelle Matrix X stets die Zerlegung $X = U\Lambda^{1/2}V'$, wobei U_0 und V_0 in (**??**) wieder in U und V umbenannt worden sind. U ist die Matrix der Eigenvektoren von XX', V ist die Matrix der Eigenvektoren von $X'X$, und $\Lambda^{1/2}$ ist die Diagonalmatrix der Wurzeln aus den Eigenwerten von XX' und $X'X$. Auf Seite 19 sind in Gleichung (3.8) die Größen $x_{ij} = (p_{ij} - r_i c_j^{'})/r_i c_j^{'}$ eingeführt worden. Geht man von

$$X = U\Lambda^{1/2}V'$$

aus, so folgt

$$X = D_r^{-1/2}(P - rc')D_c^{-1/2} = U\Lambda^{1/2}V'. \tag{3.57}$$

Dann folgt weiter

$$P - rc' = D_r^{1/2}U\Lambda^{1/2}V'D_c^{1/2}. \tag{3.58}$$

Es seien nun

$$A = D_r^{1/2}U, \quad B = D_c^{1/2}V. \tag{3.59}$$

Da U und V orthonormal sind, gilt $U'U = I$, $V'V = I$, I die Einheitsmatrix. Aber $U = D_r^{-1/2}A$, $V = D_c^{-1/2}B$. Dann folgt

$$U'U = A'D_r^{-1}A = I, \quad V'V = B'D_c^{-1}B = I, \tag{3.60}$$

und statt (3.58) läßt sich

$$P - rc' = A\Lambda^{1/2}B' \tag{3.61}$$

schreiben. (3.61) liefert eine Zerlegung nicht der gewichteten Residuen, sondern der ungewichteten Residuen $p_{ij} - r_i c_j$. Diese Zerlegung ergibt sich aus einer Skalierung der durch die SVD gegebenen Koordinaten U und V.

Definition 3.6 *Die Gleichung (3.61) heißt, zusammen mit (3.60), die* generalisierte SVD *der Matrix $P - rc'$. (vergl. Greenacre (1984), p. 87).*

Die Koordinaten F und G lassen sich dann in der Form

$$F = D_r^{-1}A\Lambda^{1/2} \tag{3.62}$$

$$G = D_c^{-1}B\Lambda^{1/2} \tag{3.63}$$

anschreiben (Greenacre (1984), P. 89).

4 Beziehungen zu anderen Verfahren und Anwendungen

4.1 Korrespondenzanalyse und Kanonische Korrelation

Bei der Kanonischen Korrelation werden zwei Datensätze, erhoben an den gleichen Personen, aufeinander bezogen. Man hat demnach zwei Datenmatrizen X und Y, wobei X eine $m \times p$ und Y eine $m \times q$-Matrix ist. Demnach existieren (Transformations-)Matrizen A und B derart, dass

$$U = XA', \quad V = YB', \tag{4.1}$$

wobei U eine $(m \times p)$- und V eine $(m \times q)$-Matrix ist. Die Spaltenvektoren von U sind orthogonal, ebenso die Spaltenvektoren von V. A und B sind derart, dass die Kanonischen Korrelationen $r(\vec{U}_k, \vec{V}_k)$ die jeweils maximal möglichen sind. Die kanonischen Variablen sind die Spalten $\vec{a}_r$ der Matrizen $A = [\vec{a}_1, \ldots, \vec{a}_s]$ und $\vec{b}_r$ von $B = [\vec{b}_1, \ldots, \vec{b}_2]$, und es gilt

$$R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx} A = A\Lambda^2 \tag{4.2}$$
$$R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy} B = B\Lambda^2 \tag{4.3}$$

gegeben, d.h. als Eigenvektoren der Matrizen

$$R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx} \quad \text{und} \quad R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}.$$

Die zugehörigen kanonischen Korrelationen sind die Diagonalelemente von Λ.

Bei der Korrespondenzanalyse werden Kontingenztabellen analysiert, und ein Zusammenhang mit der Kanonischen Korrelation, bei der metrische Daten analyisiert werden, ist zunächst nicht offenkundig. Tatsächlich ist dieser Zusammenhang aber leicht herzustellen. Dazu betrachte man die "Einsortierung" einer Person oder eines Objektes in eine Kontigenztabelle: man stellt fest, dass sie einerseites in die i-te Kategorie R_i und andererseits in die j-te Kategorie S_j fällt. Damit gehört sie in die (i, j)-te Zelle der Kontingenztabelle. Man kann dies durch Nebeneinanderschreiben der beiden Klassen von Kategorien für die k-te Person P, $k = 1, 2, \ldots, m$, verdeutlichen: Schreibt man die Kategorisierungen für alle Per-

P	R_1	$\cdots$	R_{i-1}	R_i	R_{i+1}	$\cdots$	R_I	S_1	$\cdots$	S_{j-1}	S_j	S_{j+1}	$\cdots$	S_J
k	0	$\cdots$	0	1	0	$\cdots$	0	0	$\cdots$	0	1	0	$\cdots$	0

sonen untereinander an, so entstehen zwei nebeneinander geschriebene Matrizen, die aus zwei Teilmatrizen bestehen: die Spalten der ersten repräsentieren die Katgegorien R_i, die der zweiten die Kategorien S_j. Jede Zeile dieser Dopplematrix enthält genau zwei Einsen: die erste indiziert die Kategorisierung hinsichtlich der R-Kategorien, die zweite die Kategorisierung hinsichtlich der S-Kategorien.

Die erste Teilmatrix, d.h. die für die R-Kategorisierungen, werde nun mit Z_1 bezeichnet, und die zweite für die S-Katgegorisierungen werde mit Z_2 bezeichnet;

die Gesamtmatrix läßt sich dann in der Form $Z = [Z_1, Z_2]$ anschreiben. Man verifiziert nun leicht, dass die Kontigenztabelle K durch das Matrixprodukt

$$K = Z_1' Z_2 \tag{4.4}$$

gegeben ist. Abgesehen davon, dass hier die Spalen von Z_1 und Z_2 nicht standardisiert wurden entspricht K damit der Matrix R_{xy} bei der Kanonischen Korrelation, wenn man X mit Z_1 und Y mit Z_2 identifiziert.

Man kann nun von der Matrix, d.h. der Kontingenztabelle K, die die absoluten Häufigkeiten enthält, zu den relativen Häufigkeiten übergehen, indem man durch die Gesamtzahl m der Beobachtungen (= Personen) dividiert; man erhält die Matrix P:

$$P = \frac{1}{m} Z_1' Z_2. \tag{4.5}$$

Weiter findet man, dass

$$D_r = \frac{1}{m} Z_1' Z_1, \quad D_c = \frac{1}{m} Z_2' Z_2, \tag{4.6}$$

d.h. die Diagonalmatrizen D_r und D_c enthalten in den Diagonalen die Summen der Spalten von Z_1 bzw. von Z_2. Diese Summen geben die Häufigkeit an, mit der eine Kategorie in Z_1 bzw. Z_2 vorgekommen ist. In Abschnitt 6.2.2 wurde gezeigt, dass die Skalenwerte F und G für die Zeilen- bzw. Spaltenkategorien die Lösungen der Eigenvektorgleichungen

$$\begin{aligned} (D_r^{-1} P D_c^{-1} P') F &= F\Lambda \\ (D_c^{-1} P' D_r^{-1} P) G &= G\Lambda \end{aligned}$$

sind. Wgen (4.5) und (4.6) sieht man, dass diese Gleichungen den Gleichungen (4.2) und (4.3) äqivalent sind. Die Spalten von F und G entsprechen also kanonischen Variablen, und die Inertiaanteile in Λ entsprechen kanonischen Korrelationskoeffizienten.

4.2 Multiple Korrespondenzanalyse

Die multiple Korrespondenzanalyse wird auf Indikatormatrizen angewendet. Solche Matrizen sind im Prinzip bereits in Abschnitt 4.1 betrachtet worden: Es werden Q Kategorien betrachtet, von denen jede eine Anzahl von Unterkategorien hat. Eine Person oder ein gemessenes Objekt wird genau einer Unterkategorie jeder dieser Kategorien zugeordnet, wobei die Zuordnung durch eine 1 indiziert wird. J_j ist die Anzahl der Unterkategorien der j-ten Kategorie. Eine Zeile, korrespondierend zu einer Person oder einem Objekt, enthält dann bis auf die insgesamt Q Einsen nur Nullen, vergl. Tabelle 3;

4.2.1 Spezialfall: die bivariate Indikatormatrix ($Q = 2$)

Dieser Fall repräsentiert die Rohdaten für eine $(J_1 \times J_2)$-Kontingenztabelle. Ist

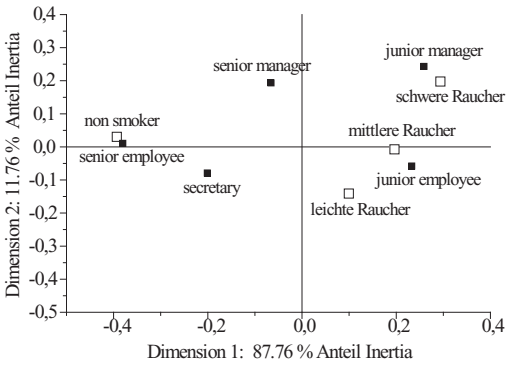
Tabelle 3: Q-variate Indikatormatrix

	$\overbrace{\quad}^{J_1}$	$\overbrace{\quad\quad}^{J_2}$		$\overbrace{\quad}^{J_Q}$
1	1 0 0 0	0 1 0 0 0	$\cdots$	0 1 0
2	0 1 0 0	0 0 1 0 0	$\cdots$	1 0 0
$\vdots$				
i	0 0 1 0	0 1 0 0 0	$\cdots$	0 1 0
$\vdots$				
I	1 0 0 0	0 0 0 1 0	$\cdots$	0 0 1

Tabelle 4: Status und Rauchgewohnheit; SM = Senior Manager, JM = Junior Manager, SE = senior employee, JE = junior employee, SC = secretary, no = Nichtraucher, li = light, med = medium, hv = heavy. (1, 1): Person ist SM und li-Raucher, (5,4): Person ist Sekretär(in) und schwere(r) Raucher(in).

Klassif.	SM	JM	SE	JE	SC	no	li	med	hv	Σ
(1, 1)	1	0	0	0	0	1	0	0	0	2
(1, 1)	1	0	0	0	0	1	0	0	0	2
$\vdots$										
(1,2)	1	0	0	0	0	0	1	0	0	2
$\vdots$										
(1, 3)	1	0	0	0	0	0	0	1	0	2
$\vdots$										
(5, 4)	0	0	0	0	1	0	0	0	1	2
(5, 4)	0	0	0	0	1	0	0	0	1	2
$\sum$	11	18	51	88	25	61	45	62	45	386

Abbildung 6: Typ der Anstellung und Rauchverhalten



Z_1 die Teilindikatormatrix für die erste Kategorie it J_1 Unterkategorien, und Z_2 die Teilindikatormatrix für die zweite Kategorie mit J_2 Unterkategorien, so ist die Indikatormatrix in der Form

$$Z = [Z_1, Z_2] \tag{4.7}$$

darstellbar; Tabelle 4 erläutert die Konstruktion einer solchen Indikatormatrix. Die Kontingenztabelle K ergibt sich durch das Produkt

$$K = Z_1' Z_2. \tag{4.8}$$

Die Tabelle 5 ist die der Tabelle 4 entsprechende Kontingenztabelle. Man be-

Tabelle 5: Typen von Rauchern in einer größeren Firma

Klassif.	Nichtr (nr)	leicht (li)	mittel (med)	schwer (hv)	$\sum$
Sen. Man.	4	2	3	2	11
Jun. Man.	4	3	7	4	18
Sen. empl.	25	10	12	4	51
Jun. empl.	18	24	33	13	88
Secret.	10	6	7	2	25
$\sum$	61	45	62	25	193

achte, dass nach Tabelle 5 die Gesamtzahl der Fälle gleich 193 ist, dass aber in Tabelle 4 ein Gesamtwert von 386 auftritt. Das liegt daran, dass in der Tabelle 4 jeder Fall zweimal auftritt, d.h. in jeder Zeile von Z tritt ein Fall einmal für die erste Katgegorie (Status), und ein weiteres Mal für die Raucherkategorie auf. Dementsprechend treten die Zeilensummen der Tabelle 4 als Spaltensummen für Z_1 auf, und die Spaltensummen von Tabelle 5 treten als Spaltensummen von Z_2 in Tabelle 4 auf.

Es sei n_{ij} die Anzahl der Fälle in der (i,j)-ten Zelle von K. Dann gibt es genau n_{ij} Zeilen in Z, bei denen in der i-ten Spalte von Z_1 und in der j-ten Spalte von Z_2 (d.h. in der $(i+j)$-ten Spalte von Z) eine 1 auftritt. Es ist möglich, die Matrix Z (also *nicht* die Tabelle K) ebenfalls einer Korrespondenzanalyse zu unterziehen. Die Zeilen- und die Spaltenkategorien von Z erscheinen dann wieder als Punkte in einem Biplot; dabei werden die n_{ij} identischen Zeilen auf einen Punkt, den Punkt (i,j), abbgebildet. Die Spalten werden ebenfalls auf Punkte abgebildet, die der Spalte entsprechend eine Subkategorie einer Kategorie entsprechen.

Ist N die Gesamtzahl der Fälle, so ist also $2N$ gleich der Summe aller Einträge in Z; dementsprechend ist die Matrix P^Z durch

$$P^Z = \frac{1}{2N} Z \tag{4.9}$$

gegeben. Die Diagonalmatrizen der relativen Zeilen- und Spaltenhäufigkeiten von Z sind dann

$$D_r^Z = \frac{1}{N} I, \quad D_c^Z = \frac{1}{2} \begin{pmatrix} D_r & 0 \\ 0 & D_c \end{pmatrix}, \tag{4.10}$$

mit D_r und D_c die Diagonalmatrizen der relativen Zeilen- bzw. Spaltenhäufigkeiten von K. Nach (6.20) gilt für die Koordinaten F der Zeilenkategorien von K die Beziehung

$$(D_r^{-1} P D_c^{-1} P') F = F\Lambda.$$

Diese Beziehung gilt für die Analyse einer beliebigen Matrix, also auch für Z, so dass die entsprechenden Koordinaten Γ^Z von Z in analoger Form durch die Beziehung

$$\left[2 \begin{pmatrix} D_r^{-1} & 0 \\ 0 & D_c^{-1} \end{pmatrix} \frac{1}{2} Z' N \frac{1}{2N} Z \right] \Gamma^Z = \Gamma^Z D_\lambda^Z \tag{4.11}$$

gegeben sind. Die hier auftretende Matrix

$$B^* = \frac{1}{2} Z' N \frac{1}{2N} Z$$

kann, wie man nach kleiner Rechnung nachweist, in der Form

$$B^* = \begin{pmatrix} Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2 \end{pmatrix} \tag{4.12}$$

geschrieben werden; dies ist eine Supermatrix, deren Elemente selbst Matrizen sind. B^* ist als Burt-Matrix bekannt, nach Burt (1950). Dabei ist $Z_1'Z_2$ gerade die Kontingenztabelle K, vergl. (4.7), und $Z_1'Z_1$ und $Z_2'Z_2$ sind die Diagonalmatrizen der Zeilen- bzw. Spaltenhäufigkeiten von K. In entsprechender Weise kann man Γ^Z aufteilen:

$$\Gamma^Z = \begin{pmatrix} \Gamma_1^Z \\ \Gamma_2^Z \end{pmatrix}, \tag{4.13}$$

wobei Γ_1^Z J_1 Zeilen und Γ_2^Z J_2 Zeilen hat. (4.11) läßt sich dann in der kompakten Form

$$\frac{1}{2N} \begin{pmatrix} D_r^{-1} & 0 \\ 0 & D_c^{-1} \end{pmatrix} \begin{pmatrix} Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2 \end{pmatrix} \begin{pmatrix} \Gamma_1^Z \\ \Gamma_2^Z \end{pmatrix} = \begin{pmatrix} \Gamma_1^Z \\ \Gamma_2^Z \end{pmatrix} D_\lambda^Z \tag{4.14}$$

schreiben. Multipliziert man diese Gleichungen aus, so erhält man

$$\frac{1}{2N}(D_r^{-1}Z_1'Z_1\Gamma_1 + D_r^{-1}Z_1'Z_2\Gamma_2) = \Gamma_1 D_\lambda \quad (4.15)$$

$$\frac{1}{2N}(D_c^{-1}Z_2'Z_1\Gamma_1 + D_c^{-1}Z_2'Z_2\Gamma_2) = \Gamma_2 D_\lambda. \quad (4.16)$$

Es ist aber $Z_1'Z_1/N = D_r$, $Z_2'Z_2/N = D_c$ und $Z_1'Z_2/N = P$, so dass

$$\Gamma_1^Z + D_r^{-1}P\Gamma_2^Z = 2\Gamma_1^Z D_\lambda^Z \quad (4.17)$$

$$D_c^{-1}P'\Gamma_1^Z + \Gamma_2^Z = 2\Gamma_2^Z D_\lambda^Z \quad (4.18)$$

folgt. Multipliziert man (4.17) von links mit $D_c^{-1}P'$ und setzt man den Ausdruck für $D_c^{-1}P'\Gamma_1^Z$ aus (4.18) ein, so erhält man

$$D_c^{-1}P'D_r^{-1}P\Gamma_2^Z = \Gamma_2^Z(2D_\lambda^Z - I)(2D_\lambda^Z - I). \quad (4.19)$$

Auf analoge Weise erhält man

$$D_r^{-1}PD_c^{-1}P'\Gamma_1^Z = \Gamma_1^Z(2D_\lambda^Z - I)(2D_\lambda^Z - I). \quad (4.20)$$

Die Gleichungen (4.19) und (4.20) sind Gleichungen für die Eigenvektoren Γ_1^Z und Γ_2^Z; gleichzeitig entsprechen diese Gleichungen den Gleichungen (6.20) und (6.21), so dass die Lösungen dafür auch Lösungen für (4.19) und (4.20) sind, d.h. es muß $\Gamma_1^Z = F$, $\Gamma_2^Z = G$ und $(2D_\lambda^Z - I)(2D_\lambda^Z - I) = \Lambda$ gelten. Hieraus folgen die Beziehungen

$$\lambda = (2\lambda^Z - 1)^2, \text{ oder } \lambda^Z = (1 \pm \lambda^{1/2})/2. \quad (4.21)$$

Für die Koordinaten ergeben sich also für die Indikatormatrix die gleichen Werte wie für die Kontingenztabelle; die Eigenwerte und damit die erklärten Inertiaanteile unterscheiden sich aber. Eine ausführliche Diskussion geometrischer Aspekte der Lösung findet man bei Greenacre (1984), S. 133.

4.2.2 Multivariate Indikatormatrizen und Burt-Matrizen

Für den allgemeinen Fall von $q = 1, \ldots, Q$ Kategorien mit jeweils J_q Subkategrien wird zunächst die entsprechende Burt-Matrix eingeführt: es ist

$$B = Z'Z = \begin{pmatrix} Z_1'Z_1 & Z_1'Z_2 & \cdots & Z_1'Z_Q \\ Z_2'Z_1 & Z_2'Z_2 & \cdots & Z_2'Z_Q \\ & & \vdots & \\ Z_Q'Z_1 & Z_Q'Z_2 & \cdots & Z_Q'Z_Q \end{pmatrix}. \quad (4.22)$$

Dabei ist jedes Element $Z_j'Z_k$, $j \neq k$ eine 2-dimensionale Kontgingenztabelle mit den J_j Subkategorien der j-ten Kategorie als Zeilenkategorien und den J_k Subkategorien der k-ten Kategorie als Spaltenkatgegorien; diese Kontingenztabellen repräsentieren die Assoziation zwischen den Kategorien j und k, "gemittelt" (= summiert) über die Personen. Darüber hinaus enthält B die Häufigkeiten, mit der eine Subkategorie der q-ten Kategorie mit einer Subkategorie einer anderen

Kategorie q' vorkommt. Z.B. können die Kategorien Ratingskalen repräsentieren; die q-te und die q'-te Kategorie entsprechen dann zwei verschiedenen Skalen, und B enthält die Häufigkeiten, mit denen ein Skalenwert S_{iq} mit einem Skalenwert $S_{jq'}$ vorkommt. Die Matrizen $Z_q'Z_q$ sind Diagonalmatrizen der Spaltensummen von Z_q.

Die Burt-Matrix B ist eine symmetrische Matrix. Die Singularwertzerlegung einer solchen Matrix liefert notwendig gleiche Skalenwerte für die Zeilen und Spalten, - einfach weil Zeilen und Spalten die gleiche Bedeutung haben. Wie im Fall $Q = 2$ gilt allgemein, dass die Skalenwerte identisch sind mit denen, die sich bei der Analyse der Indikatormatrix Z ergeben, und für die Inertia-Werte gilt

$$\lambda^B = (\lambda^Z)^2. \tag{4.23}$$

Eine Illustration der Konstruktion und Anwendung einer Burt-Matrix wird in Beispiel 5.4, Seite 52, gegeben.

5 Beispiele

5.1 Körperbau und Charakter

Beispiel 5.1 Westphal (1931) erstellte die schon auf Seite 38 vorgestellte Tabelle 6 zum Zusammenhang zwischen Körperbau und Charakter, wie er etwa von Kretschmer (1961) diskutiert wurde (vergl. Hofstätter (1971), p. 330). Westphal klassifizierte insgesamt 8099 Patienten in Psychiatrischen Landeskrankenhäusern (i) nach ihrem Körperbautyp, und (ii) nach der bei ihnen diagnostizierten Störung. Die Zeile "erw." in Tabelle 6 enthält die Häufigkeiten, die man im Falle stochastischer Unabhängigkeit gemäß der allgemeinen Formel $P(A \cap B) = P(A)P(B)$, d.h.

$$\hat{n}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{N}, \quad n_{i\cdot} = \sum_j n_{ij}, \quad n_{\cdot j} = \sum_i n_{ij} \tag{5.1}$$

erwarten würde, wären Körperbau und psychische Erkrankung stochastisch unabhängig voneinander; $N = 8099$ ist die Gesamtzahl der Beobachtungen in der Tabelle. Tabelle 7 zeigt die relativen Häufigkeiten der Kombinationen von Körperbau und Erkrankung, d.h. die Werte $p_{ij} = n_{ij}/N$, zusammen mit den relativen Häufigkeiten r_i der Zeilen- bzw. c_j der Spaltenkategorien[19].

Wie bereits der Tabelle 6 zu entnehmen ist, bilden die Leptosomen einerseits die größte Teilstichprobe, wenn man nach Körperbau klassifiziert, und die Schizophrenen andererseits, wenn man nach Erkrankung klassifiziert. Mit Ausnahme der Pykniker wird die Schizophrenie bei allen Körperbautypen am häufigsten diagnostiziert, und bei den Pyknikern wird die Schizophrenie immerhin mit zweitgrößter Häufigkeit festgestellt. Um die Art der Assoziationen genauer zu erfassen, kann

[19] Im Jargon der Korrespondenzanalytiker sind dies die *Massen* der Kategorien; dieser Sprachgebrauch bezieht sich auf Analogien zu bestimmten physikalischen Begriffsbildungen, die hier nicht weiter elaboriert werden sollen, weil sie kaum hilfreich für das Verständnis der KA als Methode sind.

Tabelle 6: Körperbau und psychische Erkrankung: beobachtete und erwartete Häufigkeiten

		Erkrankung (ψ)			
Typ (ϕ)		man./dep.	Epilepsie	Schizophr.	Σ
pyknisch	n_{ij}	**879**	83	717	1679
erw.	$\hat{n}_{ij}$	282	312	1085	
athletisch	n_{ij}	91	435	**884**	1410
erw.	$\hat{n}_{ij}$	237	262	911	
leptosom	n_{ij}	261	378	**2632**	**3271**
erw.	$\hat{n}_{ij}$	549	608	2114	
dysplastisch	n_{ij}	15	444	**550**	1009
erw.	$\hat{n}_{ij}$	170	187	652	
atypisch	n_{ij}	115	165	**450**	730
erw.	$\hat{n}_{ij}$	123	136	471	
	Σ	1361	1505	**5233**	$N = 8099$

man versuchen, die bedingten Wahrscheinlichkeiten, d.h. die Zeilen und Spaltenprofile (s. die Definitionen in (3.15), Seite 21 zu untersuchen. Es ist $\hat{p}$ statt P

Tabelle 7: Körperbau (ϕ) und psychische Erkrankung (ψ): relative Häufigkeiten

	Erkrankung (ψ)			
Typ (ϕ)	man.-dep.	Epilepsie	Schizophr.	$r_i = \hat{p}(\phi_i)$
pyknisch	**.1085**	.0102	.0885	.2070
athletisch	.0112	.0537	**.1091**	.1740
leptosom	.0322	.0467	**.3250**	.4039
dysplastisch	.0018	**.0548**	**.0679**	.1246
atypisch	.0142	.0204	**.0556**	.0901
$c_j = \hat{p}(\psi_j)$	.1680	.1858	.6461	1.000

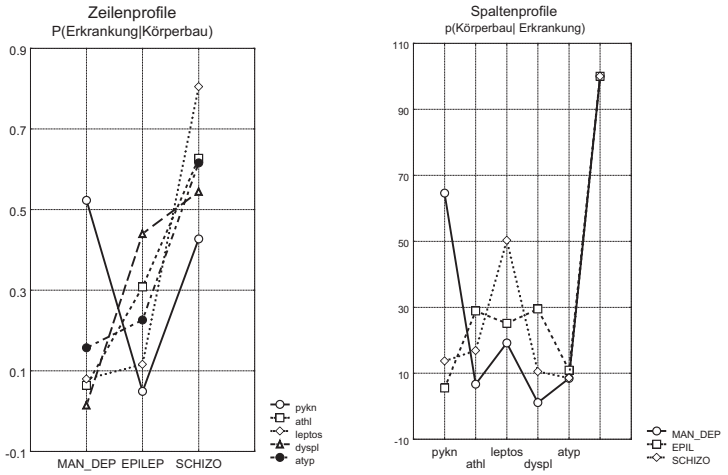
für die Wahrscheinlichkeiten geschrieben worden, um anzuzeigen, dass es sich um Stichprobenschätzungen handelt. Die Quotienten 261/3271 ergeben sich aus der Formel $p(A|B) = p(A \cap B)/p(B)$.

Abb. 7 zeigt die Zeilen- und Spaltenprofile der Datenmatrix. Es werden beispielhaft einige der bedingten Wahrscheinlichkeiten $p(\psi_j|\phi_i)$ und ihrer "Inversen" $p(\phi_i|\psi_j)$ diskutiert.

$$P(\psi_j|\phi_i) = P(\phi_i|\psi_j)\frac{P(\psi_j)}{P(\phi_i)}, \quad P(\phi_i|\psi_j) = P(\psi_j|\phi_i)\frac{P(\phi_i)}{P(\psi_j)} \tag{5.2}$$

Offenbar ist $P(\psi|\phi) \neq P(\phi|\psi)$, wenn $P(\phi) \neq P(\Psi)$, und dies ist für die gegebenen

Abbildung 7: Zeilen- und Spaltenprofile



Daten der Fall. Ein Beispiel ist

$$\hat{p}(\text{mp}|\text{pyk}) = \frac{879}{1679} = .573, \quad \hat{p}(\text{pyk}|\text{mp}) = \frac{879}{1361} = .646 \tag{5.3}$$

$$\hat{p}(\text{schiz}|\text{pyk}) = \frac{717}{1670} = .427, \quad \hat{p}(\text{pyk}|\text{schiz}) = \frac{717}{5233} = .137 \tag{5.4}$$

$$\hat{p}(\text{mp}|\text{lept}) = \frac{261}{3271} = .080, \quad \hat{p}(\text{lept}|\text{mp}) = \frac{261}{1361} = .191 \tag{5.5}$$

$$\hat{p}(\text{schiz}|\text{lept}) = \frac{2632}{3271} = .805, \quad \hat{p}(\text{lept}|\text{ schiz}) = \frac{2632}{5233} = .503 \tag{5.6}$$

Eine Bewertung der kretschmerschen Theorie auf der Basis von bedingten Wahrscheinlichkeiten scheint nicht ganz einfach zu sein. Nach (5.5) ist $\hat{p}(\text{mp}|\text{lept}) = .080$ und nach (5.6) ist $\hat{p}(\text{schiz}|\text{lept}) = .805$, was als unterstützend für die Theorie gelten kann (einen kleinen Prozentsatz von Fehlklassifikationen gibt es stets), andererseits hat man nach (5.3) $p(\text{mp}|\text{pyk}) = .573$ und nach (5.4) $\hat{p}(\text{schiz}|\text{pyk}) = .427$, d.h. die Zuordnung von Pyknikern zur manisch-depressiven Erkrankung ist deutlich weniger eindeutig als die Zuordnung von Leptosomen zur Schizophrenie. Falls ein Teil der Motivation für die Formulierung kretschmerschen Theorie darin zu bestand, von der Physis auf die Psyche zu schließen zu können, so legen die Daten nahe, dass derartige Diagnosen it einiger Unsicherheit behaftet sind. Schon die "inversen" Wahrscheinlichkeiten $p(\text{pyk}|\text{mp}) = .646$ verweisen auf eine gewisse Mehrdeutigkeit, Aussagen wie "Ein(e) PatientIn ist manisch-depressiv dann und nur dann, wenn sie/er eine pyknischen Körperbau hat" sind mit der Theorie sicherlich nicht vereinbar. Es ist nicht bekannt, ob Kretschmer in seinen Schriften auf die Frage der Asymmetrie der bedingten Wahrscheinlichkeiten und ihrer

Inversen eingegangen ist, aber sie könnten bei der Klassifikation der PatientInnen durch Westphal eine Rolle gespielt haben: z.B. bei einem Patienten, der als eindeutig schizophren diagnostiziert worden war, kann er eher einen Leptosomen "gesehen" haben als er bei einem Patienten, der als eindeutig manisch-depressiv diagnostiziert wurde, den Körperbautyp als 'pyknisch' identifizieren konnte. Es ist ein bekannter Urteilsfehler, die Wahrscheinlichkeit von Konjunktionen von Merkmalen mit bedingten Wahrscheinlichkeiten zu verwechseln. Die westphalsche Tabelle läßt allerdings keine Rückschlüsse auf derartige Fehlklassifikationen zu.

Ein alternativer Ansatz zur Interpretation der Daten könnte darin bestehen, Vierfelder-Korrelationen (ϕ-Koeffizienten) zu betrachten, wie es schon Hofstätter (1971) getan hat. Tatsächlich ergibt sich die Formel, wenn man die Formel für den Produkt-Moment-Korrelationskoeffizienten auf $(0, 1)$-Daten anwendet, wobei die 0 für "nicht vorhanden" und die 1 für "vorhanden" steht. Das Problem ist, dass man nun $I \cdot J$ (jeder Körperbautyp kann mit jeder Erkrankung korreliert werden) zu interpretieren und dabei die Frage nach der Signifikanz der einzelnen ϕ-Koefizienten zu berücksichtigen hat. Der von der Korrespondenzanlyse gelieferte Biplot gibt ein direktes Bild der Abhängigkeiten in der Häufigkeitstabelle.

Bei der Korrespondenzanalyse geht man von dem schon der PCA unterliegenden Modell von additiv wirkenden, von einander unabhängigen, im Sinne von nicht korrelierenden latenten Variablen aus. Für die hier diskutierten Daten repräsentieren die latenten Variablen Aspekte des Körperbaus in Kombination mit psychischen Merkmalen. Die SVD für X wird zur Erinnerung noch einmal vorgestellt:

$$X = Q\Lambda^{1/2}V',$$

wobei X die Matrix der Residuen ist:

$$x_{ij} = \frac{n_{ij} - n_{i.}n_{.j}/N}{\sqrt{n_{i.}n_{.j}/N}} \tag{5.7}$$

Die Tabelle 8 zeigt die x_{ij}. Diese Tabelle unterscheidet sich von der Tabelle 7 der

Tabelle 8: Residuen x_{ij} nach (5.7)

Typ	man./dep.	Epilepsie	Schizophr.
pyknisch	**.395**	-.144	-.124
athletisch	-.105	**.119**	-.010
leptosom	-.137	-.104	**.125**
dysplastisch	-.132	**.208**	-.044
atypisch	-.008	**.027**	-.011

relativen Häufigkeiten; offenbar kann die intuitive Abschätzung der Relationen zwischen Zeilen- und Spaltenkategorien zu Fehlschlüssen führen. Für das Merkmal 'pyknisch' erzeugt das psychische Merkmal 'manisch-depressiv' das maximale

Residuum, für das Merkmal 'athletish' erzeugt das Merkmal 'Epilepsie' das maximale Residuum, für die Kategorie 'leptosom' ist es die Schizophrenie, die das maximale Residuum generiert, und für die Dysplastiker wie die Atyischen ist es ebenfalls die Epilepsie, für die das größte Residuum generiert wird. Insgsamt sind es die Paare 'pyknisch-man. depressiv', 'leptosom-schizophren' und 'dysplastisch-epileptisch', die die größten Abhängigkeiten zeigen. Die Tsbelle 8 verweist zwar auf die Merkmalspaare, zwischen denen die stärkste Abhängigkeit existiert, aber sie liefert noch keinen Hinweise auf die Relationen zwischen diesen Paaren. Diese Relationen werden im Biplot deutlich.

Die Koordinaten F für die Zeilenkategorien und G sind in den Gleichungen (3.29) und (3.30) angegeben worden und werden hier ebenfalls für eine direkte Referenz noch einmal wiederholt:

$$\begin{aligned} F &= D_r^{-1/2} Q \Lambda^{1/2}, \quad \mathbf{f}_k = (f_{1k}, \ldots, f_{Ik})' \\ G &= D_c^{-1/2} V \Lambda^{1/2}, \quad \mathbf{g}_k = (g_{1k}, \ldots, g_{Jk})' \end{aligned}$$

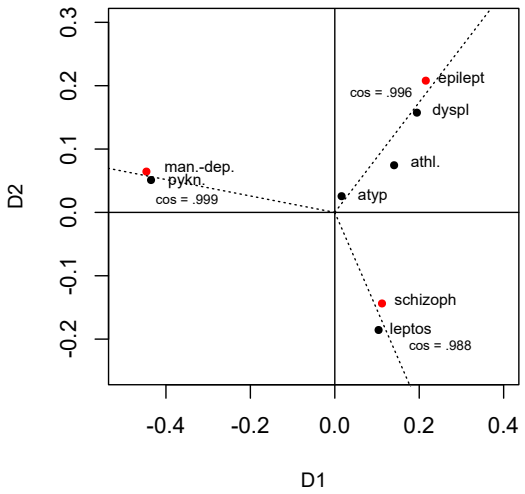
wobei Q, V und $\Lambda^{1/2}$ durch die SVD gegeben sind. Um die PCA mit der CA zu vergleichen sind Skalenwerte für die Zeilen- und Spaltenvektoren (i) nach Maßgabe der PCA, und (ii) entsprechend der CA berechnet worden, wobei für die PCA die Matrix $A = V\Lambda^{1/2} = [\mathbf{a}_1, \mathbf{a}_2]$ der Ladungen (Skalen für die Spaltenkategorien) und $F_0 = Q\Lambda^{1/2} = [\mathbf{f}_{01}, \mathbf{f}_{02}]$ der Faktorenscores (Skalen für die Zeilenkategorien) berechnet wurden. Für beide Analsen wurde der entsprechende Biplot bestimmt. Tabelle 9 enthält die Werte:

Tabelle 9: Skalenwerte bei PCA $(\mathbf{w}_1, \mathbf{w}_2)$ und $(\mathbf{v}_1, \mathbf{g}_1)$ und $(\mathbf{v}_2, \mathbf{g}_2)$ und CA-Skalierung $(\mathbf{f}_1, \mathbf{f}_2)$, bzw.

lat. Dimension	D_1	D_2	D_3	
Eigenwerte	.258	.068	3.6e-07	
Singularwerte	.508	.262	.0006	
	D_1		D_2	
Körperbau	$\mathbf{f}_{01}$	$\mathbf{f}_1$	$\mathbf{f}_{01}$	$\mathbf{f}_2$
pykn.	-.435	-.956	.052	.112
athl.	.140	.357	.074	.178
lept.	.104	.663	-.186	-.292
dysp.	.195	.551	.177	.446
atyp.	.016	.053	.026	.085
	D_1		D_2	
Erkrankung	$\mathbf{a}_1$	$\mathbf{g}_1$	$\mathbf{a}_2$	$\mathbf{g}_2$
man.-dep.	-.446	-1.085	.064	.157
Epilepsie	.216	.500	.208	.482
Schizophr.	.111	.139	-.143	-.179

Kommentar: Die numerisch existierende dritte latente Dimension reflektiert offenbar reines Rauschen in den Daten ohne jede systematische Bedeutung (der

Abbildung 8: PCA-Biplot: Kretschmertypen nach Westphal (1931). Beide Achsen mit $\Lambda^{1/2}$ skaliert



zum dritten Singularwert .006 korrespondierende Eigenwert ist gleich .00000036), d.h. es genügt, für die Interpretation der Daten nur die ersten beiden Dimensionen zu betrachten.

Zur Erinnerung: für die PCA ist die i-te Komponente von $\mathbf{f}_0$ durch $f_{0,ik} = \sqrt{\lambda_k} q_{ik}$ gegeben, $k = 1, 2$, und für die CA ist $f_{ik} = \sqrt{\lambda_k / r_i} q_{ik}$. Es zeigt sich, dass für die gegebenen Daten die Beziehung zwischen $\mathbf{f}_{01}$ und $\mathbf{f}_1$ und ebenso zwischen $\mathbf{f}_{02}$ und $\mathbf{f}_2$ in guter Näherung linear, so dass man sehr ähnliche Biplots erwarten kann.

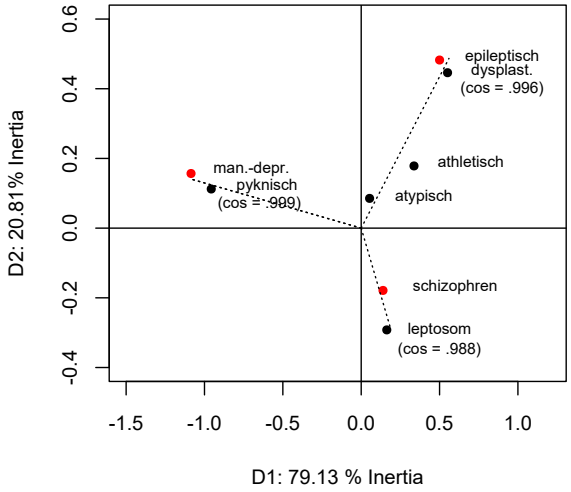
Zunächst fällt die außerordentliche Ähnlichkeit der Biplots auf. Die Maximierung der Varianz der Projektionen der Punkte, die die Kategorien repräsentieren, auf die erste latente Dimension entspricht demnach weitgehend der Maximierung des χ^2-Anteils, dass durch die erste latente Dimension erzeugt wird, eine analoge Aussage gilt für die zweite Dimension.

Die Tabelle 10 gibt einige "globale" Statistiken an: Da die Tabelle nur drei

Tabelle 10: Globale Statistiken

Dim.	χ^2	Anteil Inertia (π_k)	Eigenwert
1	$\chi^2_1 =$ 2090.152	.791	.258
2	$\chi^2_2 =$ 551.407	.021	.068
$\sum$	$\chi^2_g =$ 2641.559	1.000	In(K) = .326

Abbildung 9: CA-Biplot: Kretschmertypen nach Westphal (1931). cos bezeichnet den Kosinus des Winkel zwischen den Vektoren zu den beiden Merkmalen; da $\cos\theta = 1$ für $\theta = 0$ bedeuten angegeben Werte, dass die Winkel zwischen den Vektoren für die jeweiligen Paare von Merkmalen praktisch gleich Null sind, d.h. die Vektoren liegen in sehr guter Näherung auf einer Geraden (gepunktete Linie).



Spalten hat, gibt es zwei von Null verschiedene Eigenwerte. Der erste Eigenwert λ_1 ist ca 3.8-mal so groß wie der zweite λ_2, und dementsprechend sind die Inertia und das χ^2_1 für die erste Dimension ca 3.8-mal so groß wie die jeweiligen Größen für die zweite Dimension. Die Unterschiede zwischen den Kategorien werden also *zum größten Teil durch Unterschiede bezüglich der ersten latenten Variablen erzeugt.*

Korrespondierend zu den Residuen in Tabelle 5.7 zeigen die Biplots die äußerst enge Beziehung zwischen den Kategorien pyknisch& manisch-depressiv, dysplastisch & epileptisch sowie leptosom & schizophren. Die Enge der jeweiligen Beziehung läßt sich durch das Skalarprodukt zwischen den Vektoren, die die jeweils zwei Kategorien repräsentieren, ausdrücken: relativ zur Länge der Vektoren wird das Skalarprodukt maximal, wenn der Winkel zwischen den Vektoren gleich Null ist; der Kosinus des Winkels ist dann gleich 1, wie der generelle Ausdruck

$$\mathbf{x}'\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \cos\theta_{xy}, \quad \cos\theta = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|},$$

für ein Skalarprodukt zeigt. Der Kosinus ist gleich dem Skalarprodukt der auf die Länge 1 normierten Vektoren $\mathbf{x}$ und $\mathbf{y}$. Der Wert des Kosinus (cos) für jedes der drei Paare ist in den Biplots eingetragen worden; die Werte sprehen dafür, dass die beiden Vektoren in extrem guter Näherung parallel sind, also auf einer Geraden liegen (das sind die punktierten Geraden). Die Koordinaten der Endpunkte

der beiden Vektoren unterscheiden sich nur um einen gemeinsamen Faktor, d.h. die Merkmalspaare sind *strukturell identisch*: das *das Verhältnis* der Anteile der Merkmale an den beiden latenten Dimensionen, wie sie durch die Koordinaten f_{ik} und g_{jk} gegeben sind, für die beiden Merkmale dasselbe ist:

$$\frac{f_{i1}}{f_{i2}} \approx \frac{g_{j1}}{g_{j2}} \tag{5.8}$$

damit $i \rightarrow$ pyknisch, $j \rightarrow$ manisch-depressiv, analog für die Paare dysplastisch-epileptisch und leptosom-schizophren. Das $\approx$-Zeichen statt des $=$-Zeichens wurde hier verwendet, weil die Beziehung wegen der unvermeidlichen Stichprobenfluktuationen nur approximativ gelten kann. Die Athleten liegen in guter Näherung zwischen den Atypischen und den Dysplastikern, – ob die geringe Abweichung von der Geraden systematisch oder eher zufällig ist, ist anhand der Daten schwer zu entscheiden.

Wenn also zum Beispiel Pykniker und Manisch-Depressive in strukturell identischer Weise durch die beiden latenten Dimensionen definiert werden, so ist die Frage, was die latenten Variablen überhaupt bedeuten, denn die eine Kategorie bezeichnet ein physisches, die andere Kategorie ein psychisches Merkmal. Dazu muß man sich zuerst vergegenwärtigen, dass ein Skalarprodukt nur das Ausmaß von Ähnlichkeit ("Korrelation") der beiden Vektoren, deren Skalarprodukt gebildet werden, angibt, und Korrelationen bedeuten ja nicht die Identität der durch die Vektoren repräsentierten Merkmale, sondern eben nur ihre Kovariation. Die kann zustande kommen, wenn beide Merkmale durch ein drittes, "latentes" (weil nicht direkt gemessenes) Merkmal bestimmt werden. Man könnte zum Beispiel die Hypothese aufstellen, dass die Kopplung zwischen physischen und psychischen Aspekten sich auf bestimmte genetische Konstellationen bezieht, die sich einerseits auf physische und andererseits auf psychische Aspekte auswirken. Eine solche Hypothese muß natürlich durch weitere Untersuchungen gestützt werden.

Ein Skalarprodukt ist ein *symmetrisches* Ähnlichkeitsmaß für eine Merkmalskombination. Die bedingten Wahrscheinlichkeiten für dieselbe Merkmalskombination sind im Allgemeinen nicht symmetrisch,

$$P(\psi|\phi) \neq P(\phi|\psi), \quad \text{d.h. } P(\psi|\phi) = P(\phi|\psi)\frac{P(\psi)}{P(\phi)},$$

man kann also sagen, dass die Asymmetrie durch unterschiedliche Werte für die absoluten Wahrscheinlichkeiten erzeugt wird. In die graphische Repräsentation der Kategorien im Biplot gehen diese Asymmetrien offenbar nicht ein. Der Grund dafür wird insbesondere in der Gleichung (3.41), Seite 26 deutlich, die hier noch einmal angeschrieben wird:

$$FG' = D_r^{-1}P(G\Lambda^{-1}F')PD_c^{-1/2};$$

hier wird einerseits die Matrix G der Koordinaten für die Spaltenkategorien mit $D_r^{-1}P$, der Matrix der bedingten Wahrscheinlichkeiten für die Spaltenmerkmale gewichtet, und andererseits wird die Matrix F' mit der Matrix der bedingten

Wahrscheinlichkeiten $PD_c^{-1/2}$ für die Zeilenmerkmale gewichtet. Beide Arten bedingter Wahrscheinlichkeiten gehen in die Definitionen von F und G ein und werden bei der Bildung der Skalarprodukte FG' gewissermaßen wieder herausgekürzt.

Es gibt noch eine Reihe von Statistiken, anhand derer man die Güte einer Lösung diskutieren kann. In der Tabelle 11 sind diese Statistiken zusammengefaßt worden. Die Spalte f_{i1} enthält die Koordinaten für die erste Achse, f_{i2} die für die zweite Achse.

Tabelle 11: Koordinaten und Statistiken: Körperbau und Erkrankung

Körperbau	$\pi_{i\cdot 1}$	$\cos^2\Theta_{11}$	$\pi_{i\cdot 2}$	$\cos^2\Theta_{21}$	$\rho_{i\cdot}$
pyknisch	.7340	.9863	.0385	.0136	.5888
athletisch	.0776	.7831	.0806	.2169	.0776
leptosom	.0417	.2384	.5052	.7616	.1384
dysplast.	.1465	.6032	.3654	.3968	.1922
atypisch	.0010	.2720	.0103	.7280	.0030
Erkrankung	$\pi_{\cdot j1}$	$\cos^2\Theta_{12}$	$\pi_{\cdot j2}$	$\cos^2\Theta_{22}$	$\rho_{\cdot j}$
man.-dep.	.7712	.9796	.0607	.0203	.6229
Epilepsie	1803	.5189	.6338	.4811	.2751
Schizophr.	.0484	.3775	.3054	.6245	.1021

Tabelle 12: Erläuterung

$\pi_{i\cdot k}$, $\pi_{\cdot jk}$	relative Inertia: Anteil d. Inertia einer Kategorie für k-te Dim.
$\cos^2\Theta$	Koordinatenanteile: Projekt. eines Datenvektors auf eine Achse
$\rho_{i\cdot}$, $\rho_{\cdot j}$	Anteil der Kategorie an $In(K)$

Es sollen noch die Qualitätsmerkmale der Repräsentation betrachtet werden. Zunächst ist dabei an die Qualität $q_{i\cdot}$ (vergl. Gleichung (3.55)) zu denken. Die Qualität $q_{i\cdot}$ gibt an, wie gut eine Kategorie durch einen Raum der gewählten Dimensionalität repräsentiert wird. Hier werden alle Dimensionen - also zwei - berücksichtigt, deswegen sind die Qualitätsmaße für alle Kategorien gleich 1, d.h. die Kategorien werden durch die Punkte perfekt repräsentiert.

Pykniker und Dysplastiker haben auf der ersten Achse nicht nur die betragsmäßig größten Koordinaten, sondern diese beiden Typen erzeugen auch die größten Anteile an der Gesamt-Inertia: $\rho_{1\cdot}$(pykn) = .558 und $\rho_{4\cdot}$(dyspl) = .1922 (bei dieser Zerlegung betrachtet man *entweder* die Zeilen- *oder* die Spaltenkategorien, nicht die Kombinationen Zeilen/Spaltenkategorie). Man beachte gleichwohl, daß die Pykniker einen gut 3-mal so großen Anteil an der Gesamt-Inertia erzeu-

gen wie die Dysplastiker. Die Leptosomen, obwohl insgesamt am häufigksten in der Gesamtstichprobe vertreten, erzeugen einen geringeren Anteil an der Gesamt-Inertia.

Es ist noch von Interesse, die relativen Inertiae pro Kategorie und Dimension, d.h. die $\pi_{i\cdot k}$, zu betrachten. Sie ist für die erste Dimension für die Pykniker am größten ($\pi_{1\cdot 1} = .7340$), gefolgt von der für die Dysplastiker ($\pi_{4\cdot 1} = .4469$); erst dann tragen die Athleten und die Leptosomen zu dieser Dimension bei. Der Beitrag der Pykniker ist fast 18-mal ($\pi_{1\cdot 1}/\pi_{3\cdot 1} = 17.601$) so groß wie die der Leptosomen, der der Dysplastiker ist immer noch 3.5-mal so groß.Bei der zweiten Dimension findet man, daß die Beiträge der Leptosomen ($\pi_{3\cdot 2} = .5052$) und der Dysplastiker ($\pi_{4\cdot 2} = .3654$) dominieren. Diese Werte stützen den Ansatz, die zweite Dimension durch die Polarität von Leptosomen einerseits und Dysplastikern andererseits zu definieren.

Zum Abschluß sollen noch die $\cos^2\Theta_{i1}$-Werte betrachtet werden. Es gilt

$$\cos^2\Theta_{i1} + \cos^2\Theta_{i2} = 1$$

für alle $i = 1, \ldots, 5$). Der pyknische Typ wird zu fast 99 % auf der ersten Achse repräsentiert ($\cos^2\Theta_{11} = .9863$) und kaum durch die zweite ($\cos^2\Theta_{12} = .0136$). Man beachte, daß die Athleten wegen $\cos^2\Theta_{21} = .7831$ mehr auf der ersten Achse als auf der zweiten Achse ($\cos^2\Theta_{22} = .2169$) abgebildet werden. Die relative Inertia $\pi_{2\cdot 1} = .0776$ ist gleichwohl geringer als die relative Inertia der Dyplastiker ($\pi_{4\cdot 1} = .1465$), die auf dieser Achse weniger ausgeprägt abgebildet werden. Dies verdeutlicht, daß die Güte der Repräsentation eines Punktes auf einer Achse, wie sie durch den $\cos^2$-Wert dargestellt wird, noch nicht viel über das Ausmaß aussagt, mit dem ein Punkt zum Gesamt-χ^2 bzw. zur Geamt-Inertia beiträgt. Der Beitrag zum Gesamt-χ^2 wird durch die Länge eines Vektors, dessen Endpunkt eine Kategorie repräsentiert, bestimmt, denn diese Länge entspricht der χ^2-Distanz zwischen dem durchschnittlichen Profil und dem Profil dieser Kategorie.

In gleicher Weise kann man den Raum der Erkrankungen diskutieren. Hier wird die erste Dimension durch die Polarität manisch-depressiv versus Epilepsie charakterisiert, die zweite durch die Polarität Epilepsie und Schizophrenie. Man beachte, daß die zweite Dimension die Schizophrenie von den beiden anderen Erkrankungen separiert. Zerlegt man die Gesamt-Inertia in Anteile zu Lasten der Erkrankungen, so sieht man, daß über 60% ($\rho_{\cdot 1} = .6229$) durch die manisch-depressive Erkrankung erzeugt werden, der zweitgrößte Anteil wird durch die Epilepsie generiert; die Schizophrenie hat an dieser Dimension den geringsten Anteil. Die Interpretation der übrigen Statistiken wird analog zu der bei den Zeilenkategorien, d.h. den Körperbautypen, vorgenommen und braucht hier nicht im Einzelnen durchgeführt zu werden. □

5.2 Interviews zur Abtreibung

Beispiel 5.2 Marascuilo & McSweeny (1977, p. 242) berichten die folgenden Daten aus einer Umfrage, in der u.a. 500 Männer nach ihrer Einstellung zur Abtreibung interviewt wurden: es wurde ihnen die Frage:

> Does a woman have the right to decide whether an unwanted birth can be terminated during the first three month of pregnancy?
>
> Yes ☐ No ☐ No Opinion ☐

vorgelegt. Die Tabelle 13 enthält die Häufigkeiten der Antworten, aufgeschlüsselt nach dem religiösen Bekenntnis der Befragten, sowie die Zeilen- und Spalten-χ^2 und das Gesamt-χ^2 (= 40.175).

Tabelle 13: Religiöse Präferenz und Einstellung zur Abtreibung (Marscuilo und McSweeny 1977)

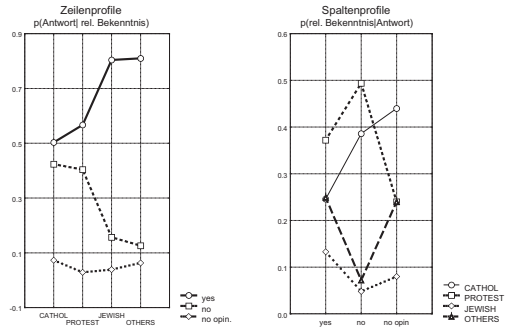
Response	Catholic	Protestant	Jewish	Others	$\sum$	Zeilen-χ^2
Yes	76	115	41	77	309	12.635
No	64	82	8	12	166	23.818
No Opinion	11	6	2	6	25	3.721
$\sum$	151	203	51	95	500	
Spalten-χ^2	8.627	5.932	7.683	18.132		40.175

Tabelle 14: Koordinaten und Statistiken: Einstellung und religiöses Bekenntnis

Antworten	f_{i1}	f_{i2}	$\pi_{i\cdot 1}$	$\cos^2\Theta_{11}$	$\pi_{i\cdot 2}$	$\cos^2\Theta_{21}$	$\rho_{i\cdot}$
Yes	.2010	-.0218	.3442	.9883	.0377	.0112	.3145
No	-.3784	-.0173	.6552	.9979	.0128	.0021	.5929
no opin.	.0277	.3848	.0005	.0052	.9495	.9948	.0926
Bekenntnis	g_{j1}	g_{j2}	$\pi_{\cdot j1}$	$\cos^2\Theta_{12}$	$\pi_{\cdot j2}$	$\cos^2\Theta_{22}$	$\rho_{\cdot j}$
Catholic	-.2123	.1099	.1876	.7887	.4676	.2113	.1876
Protest	-.1416	-.0905	.1122	.7101	.4263	.2899	.1122
Jewish	.3837	-.0586	.2070	.1020	.9772	.0449	.0228
Others	.4340	.0501	.4933	.9868	.0612	.0132	.4933
f_{ik}, g_{jk}	Koordinaten der Kategorien						
$\pi_{i\cdot k}$, $\pi_{\cdot jk}$	relative Inertia: Anteil d. Inertia einer Kategorie für k-te Dim.						
$\cos^2\Theta$	Koord'anteile: Projekt. eines Datenvektors auf eine Achse						
$\rho_{i\cdot}$, $\rho_{\cdot j}$	Anteil der Kategorie an $In(K)$						

Die relevanten Statistiken werden in Tabelle 14 angegeben. Abbildung 10 zeigt die Zeilen- und Spaltenprofile der Daten. Es sei daran erinnert, daß die Punkte in den Profilen bedingten Wahrscheinlichkeiten entsprechen. So geben die Zeilenprofile die jeweilige bedingte Wahrscheinlichkeit an, katholischen, protestantischen etc Bekenntnisses zu sein unter der Bedingung,"ja", "nein" zu sagen oder

Abbildung 10: Zeilen- und Spaltenprofile

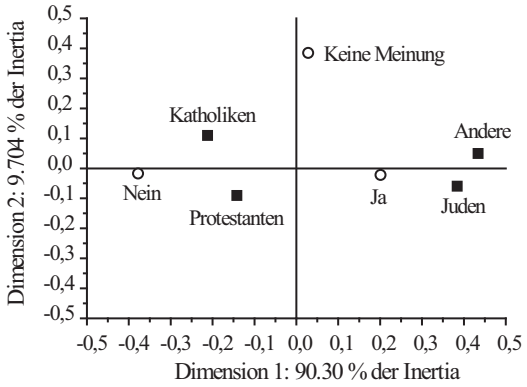


keine Meinung zu haben. Die Spaltenprofile dagegen bilden die bedingte Wahrscheinlichkeit, "ja" oder "nein" zu sagen oder keine Meinung zu äußern unter der Bedingung, ein bestimmtes religiöses Bekenntnis zu haben, ab. Betrachtet man zunächst die "ja"- und "nein"- Verläufe bei den Zeilenprofilen, so sieht man, daß die Verläufe zwar eine qualitative ähnlichkeit haben, aber nicht parallel verlaufen, was auf eine Abhängigkeit von Gruppenzugehörigkeit (religiöses Bekenntnis) und Antwort hinweist. Die bedingte Wahrscheinlichkeit, "nein" zu sagen, ist bei den Christen - d.h. Katholiken und Protestanten - größer als bei den anderen beiden Gruppen, und korrespondierend dazu ist die bedingte Wahrscheinlichkeit, "ja" zu sagen, geringer. Bei den Gruppen "Jewish" und "Others" ist es gerade umgekehrt. Einen qualitativ ganz anderen Verlauf zeigt das Profil für "no opinion". Interessanterweise ist die bedingte Wahrscheinlichkeit, keine Meinung zu haben, bei den Katholiken am größten, gefolgt von der der Protestanten, und bei der jüdischen Gruppe ist sie am geringsten.

Die Spaltenprofile zeigen, daß die Häufigkeiten in der Datentabelle im wesentlichen durch die Zugehörigkeit zu einer von zwei Gruppen strukturiert sein müssen: die der Christen einerseits und die der Nicht-Christen andererseits. Insbesondere das Profil der Protestanten ist gegenläufig zu dem der jüdischen Befragten; das Profil der Katholiken ist nur partiell parallel dem der Protestanten, da die Katholiken einen ungleich höheren Anteil von "no opinion"-Personen haben; diese Unterscheidung ist vermutlich charakteristisch für amerikanische Katholiken und Protestanten.

Abbildung 11 zeigt den Biplot für die Daten. In bezug auf die religiösen Gruppen läßt sich sagen, daß die erste Dimension Christen und Nichtchristen voneinander trennt. Der wesentliche Teil der Struktur in den Daten ist durch diese Dimension auch schon erklärt, da sie für über 90 % der Inertia der Datentabelle verantwortlich ist. Die Abweichungen der Gruppenkategorien (religiöses Bekenntnis) von der ersten Dimension könnten auf zuffällige Effekte zurückzuführen sein. Schaut man aber auf die relativen Inertiae $\pi_{i\cdot 1}$ und $\pi_{i\cdot 2}$, so sieht man, daß die relative Inertia für die Katholiken auf der ersten Dimension weniger als halb

Abbildung 11: Religiöses Bekenntnis und Einstellung zur Abtreibung: Biplot

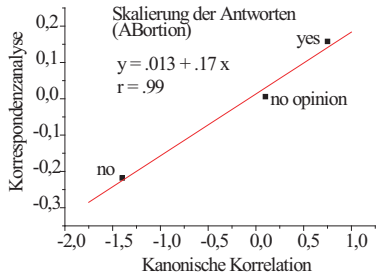


so groß ist wie die für die zweite Dimension, - in der Tat ist ja die bedingte Wahrscheinlichkeit, Katholik zu sein, wenn man *keine Meinung* hat, größer als die entsprechende bedingte Wahrscheinlichkeit, wenn man "ja" oder "nein" sagt! Bei den Protestanten beträgt die relative Inertia für die erste Dimension weniger als ein Drittel von der für die zweite Dimension. Bei der jüdischen Gruppe ist das Verhältnis sogar nur ein Fünftel. Diese beiden Gruppen haben auf der zweiten Gruppe negative Skalenwerte: eine dezidierte Ansicht - entweder "ja" oder "nein" - kommt bei ihnen also häufiger vor, d.h. die bedingten Wahrscheinlichkeiten für diese beiden Antworten sind für diese beiden Gruppen größer. Nur bei der "Others"-Gruppe ist das Verhältnis umgekehrt: der Inertia-Anteil für die erste Dimension ist 8-mal so groß wie der für die zweite Dimension.

Schaut man im Biplot auf die Antworten, so wird deutlich, daß die erste Achse durch die Antworten "nein" und "ja" bestimmt zu sein scheint. Die logische Beziehung zwischen "yes", "no" und "no opinion" verlangt, daß "no opinion" zwischen "yes" und "no" liegt, und für die erste Dimension gilt dies auch. Interessant ist aber, daß durch "no opinion" *eine neue*, zweite Dimension definiert wird. In der Tat weicht das Profil für "no opinion" in seiner Form von den "yes"- und "no"- Profilen ab. Aus den Zeilenprofilen kann man aber ablesen, daß die Wahrscheinlichkeit, die Antwort "no opinion" unter Bedingung, ein bestimmtes religiöses Bekenntnis zu haben, für alle Bekenntnisse ungefähr gleich ist, d.h. der Anteil der Personen, die keine Meinung zu der gestellten Frage haben, ist für die untersuchten Populationen (religiösen Bekenntnisse) jeweils gleich groß. Damit unterscheidet sich das Profil für "no opinion" qualitativ sowohl vom "yes"- wie vom "no"-Profil, und dies ist der (formale) Grund, weshalb "no opinion" eine neue Dimension generiert.

Der Inertia-Anteil $\pi_{i\cdot 2} = .9495$ zeigt, daß der χ^2- bzw. Anteil der Inertia, der durch "no opinion" erzeugt wird, in allererster Linie durch die zweite Dimension generiert erzeugt wird, der entsprechende Anteil für die erste Dimension ist dagegen vernachlässigbar ($\pi_{i\cdot 1} = .0005$). Die Inertia-Anteile für "yes" und "no"

Abbildung 12: Religiöse Präferenz und Einstellung zur Abtreibung

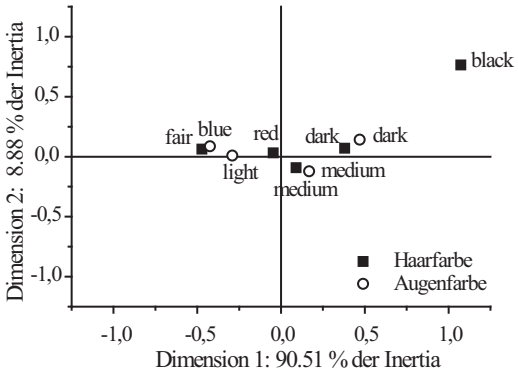


werden dagegen in erster Linie durch die erste Dimension erzeugt. Die $\cos^2$-Werte für diese beiden Antworten entsprechen natürlich dem visuellen Eindruck. Es sei noch darauf hingewiesen, daß die "no"- Antwort einen nahezu doppelt so großen Inertia-Anteil erzeugt wie die "Yes"- Antwort.

Ein zweites Merkmal des Biplots in Abb. 11 ist noch von Interesse. Die religiösen Gruppen liegen *nicht zwischen* den beiden Polen "yes" und "no". Dies könnte man vermuten, da man ja nicht mehr als zustimmen kann, und nicht nicht weniger als ablehnen, und nicht alle Mitglieder der verschiedenen Gruppen entweder zustimmen oder ablehnen. Gleichwohl liegt der Punkt für "Jewish" und "Others" rechts von "yes". Dies darf nicht weiter verwirren: man darf ja nicht vergessen, daß die Distanzen zum Mittelpunkt bzw. Ursprung des Koordinatensystems die relevanten Größen sind: diese euklidischen Distanzen entsprechen χ^2-Distanzen der Profile zum entsprechenden mittleren Profil, und diesen Distanzen entsprechen die χ^2-Anteile, die zu Lasten der einzelnen Kategorien gehen. Es ist dieser Sachverhalt, der die Lage der Punkte, die die Kategorien repräsentieren, bestimmt. Betrachtet man die Projektionen der religösen Gruppen auf die erste Achse, so liegen die Katholiken näher am Punkt für "nein" als die Protestanten. In der Tat ist die bedingte Wahrscheinlichkeit, einen Katholiken vor sich zu haben, wenn die befragte Person mit"nein" geantwortet hat, größer, als wenn die befragte Person "ja" geantwortet hätte. Auf der anderen Seite liegt die Projektion der jüdischen Gruppe näher an der der "ja"-Antwort als die der Others-Gruppe, wobei die Projektion der "ja"-Antwort *kleiner* ist als die der beiden Gruppen. Dies zeigt, daß man die Gruppen nicht *zwischen* den Polen "nein" und "ja" anordnen kann. Wie die Spaltenprofile zeigen, ist die bedingte Wahrscheinlichkeit,"ja" zu sagen, für die "Jewish"- und die "Others"-Gruppe praktisch *gleich groß*.

Man kann die Daten auch einer Kanonischen Korrelationsanalyse unterziehen; die Details werden hier nicht dargestellt, die Analyse wird in Marascuilo und Levin (1983), p. 451, vorgestellt. Die Abbildung 12 zeigt die Beziehung zwischen den Skalenwerten der ersten Dimension nach der Kanonischen Korrelation (x-Achse) und der Korrespondenzanalyse (y-Achse). Bis auf eine Skalentransformation liefern die beiden Analysen offenbar das gleiche Bild. □

Abbildung 13: Beziehung zwischen Augen- und Haarfarbe (Maung (1941), nach Daten von Tocher (1908)



5.3 Genetische Zusammenhänge

Beispiel 5.3 Das Interesse an genetischen Zusammenhängen hat früh dazu geführt, dass statistische Verfahren zur Analyse der Daten entwickelt wurden. Tocher (1908) legte eine große Studie zur Genetik der Pigmentierung schottischer Schulkinder durch, die Fischer (1940) und Maung (1941) weiter zu analysieren versuchten. Maung betrachtete die Tabelle Das χ^2 für diese Tabelle ist hochsignifikant:

Tabelle 15: Haar- und Augenfarben schottischer Schulkinder (Tocher (1908), Maung (1941))

		Haarfarbe				
		fair	red	medium	dark	black
Augen-farbe	Blue	1368	170	1041	398	1
	Light	2577	474	2703	932	11
	Medium	1390	420	3826	1842	33
	Dark	454	255	1848	1506	112

$\chi^2 = 2466.14$ bei $df = 12$ Freiheitsgraden hat unter H_0 (kein Zusammenhang) eine Wahrscheinlichkeit von $p = .000$. Der statistische Zusammenhang zwischen Merkmalen läßt sich oft durch Korrelationen ausdrücken, – aber wie will man im Falle einer solchen Tabelle die Korrelation zwischen Augen- und Haarfarbe berechnen? Die Korrespondenzanalyse liefert Skalen, in Bezug auf die die Merkmale verglichen werden können. Das Resultat der Korrespondenzanalyse dieser Daten wird in Abbildung 13 präsentiert. Offenbar erklärt die erste Dimension gut 90 % des Gesamt-χ^2. Nur das sehr dunkle Haar ("black") scheint eine systematische Abweichung von der ersten Dimension zu erzeugen.

Die Häufigkeitsverteilungen in der Tabelle 15 weisen bereits auf eine enge

Kopplung zwischen bestimmten Augen- und Haarfarben hin. Im Biplot (13) wird dieser Zusammenhang sehr deutlich dargestellt. Träten Augenfarbe und Haarfarbe bei den Individuen unabhängig voneinander auf, hätte man eine 2-dimensionale Konfiguration erhalten.

Die Abstände zwischen den Projektionen der Augen- bzw. der Haarfarbe auf die Achse(n) sind durch die entsprechenden χ^2-Distanzen definiert. Solche Distanzen sind ein Maß für die Unterschiedlichkeit der entsprechenden Zeilen- bzw. Spaltenprofile. Die Unterschiede zwischen den Zeilenprofilen (Augenfarben) sind so, dass sie einen Übergang von "Blue" zu "Dark" über "Light" und "Medium" implizieren; die χ^2-Distanzen reflektieren also gewissermaßen genetische Nachbarschaften. □

5.4 Anthropometrische Messungen und Burt-Matrizen

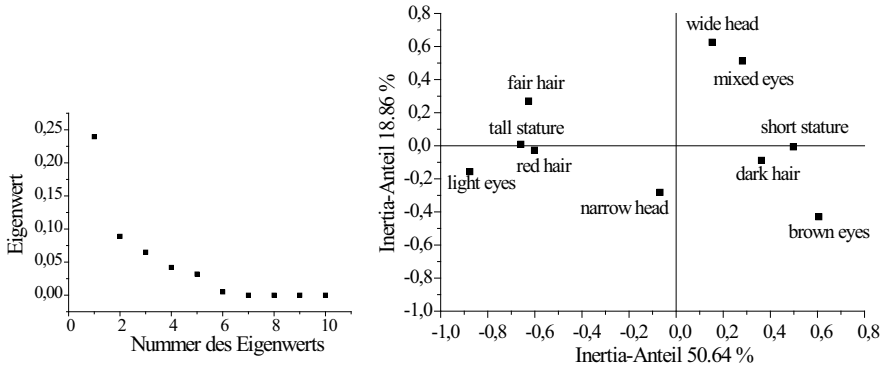
Beispiel 5.4 Burt (1950) bestimmte für 100 zufällig in Liverpool ausgewählte Personen (i) die Haarfarbe (fair, red, dark), (ii) die Augenfarbe (light, medium, brown), (iii) Kopfform (narrow, wide), und (iv) die Statur (tall, short). Es ergibt sich zunächst eine 4-dimensionale Kontingenztabelle mit jeweils 3, 3, 2 und 2 Kategorien, die die Burt-Matrix 16 findet man in Basilevsky (1994), p.532. Die

Tabelle 16: Burt-Matrix zu genetischen Abhängigkeiten (Burt, 1950)

	fh	rh	dh	le	me	be	nh	wh	ts	ss
fair hair	22	0	0	14	6	2	14	8	13	9
red hair	0	15	0	8	5	2	11	4	10	5
dark hair	0	0	63	11	25	27	44	19	20	43
light eyes	14	8	11	33	0	0	27	6	29	4
mixed eyes	6	5	25	0	36	0	20	16	10	26
brown eyes	2	2	27	0	0	31	22	9	4	27
narrow head	14	11	44	27	20	22	69	0	30	39
wide head	8	4	19	6	16	9	0	31	13	18
tall stature	13	10	20	29	10	4	30	13	43	0
short stature	9	5	43	4	26	27	39	18	0	57

Tabelle hat die Eigenwerte .2395, .0892, .0647, .0422, .0319, .0054, .000, .000, .000. Die erste Dimension separiert die Personen in zwei Klassen: die erste Klasse ist durch eine "tall stature", die zweite durch eine "short stature" charakterisiert; die Lage der Staturmerkmale ("tall' versus "short') ist bemerkenswert: sie liegen genau auf der ersten Achse. Die maximal differierenden Projektionen auf die erste Achse werden allerdings durch Augenfarben ("light" versus "brown") erzeugt. Mit der Art der "stature" einhergehen einerseits die Merkmale "fair hair", "red hair" und insbesondere "light eyes", andererseits "dark hair" und "brown eyes". Die Kopfform – "wide head" versus "narrow head" scheint, gekoppelt mit den Augenfarben "mixed" versus "brown" – eine zweite Dimension aufzumachen, wobei

Abbildung 14: Eigenwerte und (Bi)Plot der Merkmale für die Daten aus Tabelle 16; Daten: vergl. Basilevsky (1994)



es eine geringfügige Assoziation des "wide head" mit dem "short stature"-Pol und des "narrow head" mit dem "tall stature"-Pol der ersten Achse zu geben scheint. Der Verlauf der Eigenwerte legt nahe, dass es noch weitere Dimensionen gibt, oder aber die relativ suggestive Struktur der ersten beiden Dimensionen durch zufällige Kombinationen überlagert wird. □

5.5 Zeitliche Entwicklungen

Die folgenden Beispiele zeigen, dass die Korrespondenzanalyse auch dazu dienen kann, zeitliche Entwicklungen zu verdeutlichen.

5.5.1 Kriminelle Delikte Jugendlicher

Beispiel 5.5 Andersen (1989), p. 340, gibt eine Tabelle an, die die Häufigkeiten krimineller Delikte Jugendlicher enthält, bei denen die Anklage *vor* der Gerichtsverhandlung fallengelassen wurde: Abb. 15 zeigt die Zeilen- und Spaltenprofile

Tabelle 17: Straftaten dänischer Jugendlicher

	Alter						
Jahr	15	16	17	18	19	$\sum$	Zeilen-χ^2
1955	141	285	320	441	427	1614	13.335
1956	144	292	342	441	396	1615	6.019
1957	196	380	424	462	427	1889	4.277
1958	212	424	399	442	430	1907	14.779
$\sum$	693	1381	1485	1786	1680	7025	
Spalten-χ^2	7.088	11.967	2.884	9.020	7.450		χ^2 =38.410

Tabelle 18: Koordinaten und Statistiken: Straftaten dän. Jugendlicher

Jahr	f_{i1}	f_{i2}	$\pi_{i\cdot 1}$	$\cos^2\Theta_{11}$	$\pi_{i\cdot 2}$	$\cos^2\Theta_{21}$	$\rho_{i\cdot}$	q_i
1955	.088	-.022	.361	.939	.223	.058	.347	.996
1956	.058	.016	.157	.908	.124	.071	.157	.978
1957	-.039	.027	..082	.669	.391	.315	.111	.984
1958	-.085	-.021	.399	.938	.262	.061	.385	.999
Altersgruppe	g_{j1}	g_{j2}	$\pi_{\cdot j1}$	$\cos^2\Theta_{12}$	$\pi_{\cdot j2}$	$\cos^2\Theta_{22}$	$\rho_{\cdot j}$	q_j
A 15	-.101	-.007	.203	.992	.011	.005	.185	.998
A 16	-.091	-.018	.331	.959	.128	.037	.312	.996
A 17	-.023	.037	.023	.281	.594	.710	.075	.991
A 18	.073	.007	.255	..980	.024	.009	.234	.989
A 19	.062	-.022	.188	.877	.242	.112	.194	.989
f_{ik}, g_{jk}	Koordinaten der Kategorien							
$\pi_{i\cdot k}$, $\pi_{\cdot jk}$	relative Inertia: Anteil d. Inertia einer Kategorie für k-te Dim.							
$\cos^2\Theta$	Koord'anteile: Projekt. eines Datenvektors auf eine Achse							
$\rho_{i\cdot}$, $\rho_{\cdot j}$	Anteil der Kategorie an $In(K)$							
q_i, q_j	Qualität							

der Häufigkeiten.

Man sieht einen Anstieg für jedes Jahr mit dem Alter einerseits und für jedes Alter mit den Jahren 1955 - 1958 andererseits. Das Gesamt-$\chi^2 = 38.410$ ist hochsignifikant, die Abhängigkeiten sind also sicherlich nicht zufällig. Die Korrespondenzanalyse liefert die Eigenwerte $\lambda_1 = .00494$, $\lambda_2 = .00049$ und $\lambda_3 = .00004$. Die Summe der Eigenwerte beträgt .00543, und somit erklärt die erste Dimension $\lambda_1/.00543 \approx .91$ der Gesamt-Inertia (vergl. (**??**), p. **??**). Die zweite Dimension hat einen Anteil von .08972 an der Gesamt-Inertia, und damit erklären beide Dimensionen zusammen einen Anteil von .99 der Gesamt-Inertia. Die dritte Dimension kann also vernachlässigt werden, - möglicherweise sogar die zweite.

Die Betrachtung der Profile liefert einen ersten Einblick in die Struktur der Daten. Für die Jahre 1957 und 1958 ergibt sich ein monotoner Anstieg der (bedingten) relativen Häufigkeit von Straftaten, bei denen die Anklage vor der Gerichtsverhandlung fallengelassen wurde, mit der Altersgruppe: je höher die Altersgruppe, desto höher die Anzahl der Straftaten, wobei die Gruppe der 19-jährigen allerdings bereits geringfügig weniger Straftaten der betrachteten Art[20] aufweist. Für die Jahre 1955 und 1956 fallen die 17-jährigen auf: 1955 werden von dieser

[20]Im Folgenden wird die Bedingung, daß die Anklage vor der Gerichtsverhandlung fallengelassen wurde, der Einfachheit weggelassen. Die Nebenbedingung, daß die Anklage fallengelassen wurde, ist aber wichtig für die Interpretation: es ist ja möglich, daß nicht die Anzahl der Straftaten steigt oder fällt, sondern daß die Polizei ihre Politik gegenüber den verschiedenen Altersgruppen geändert hat.

Abbildung 15: Dänische Jugendkriminalität und Polizeiverhalten: Profile

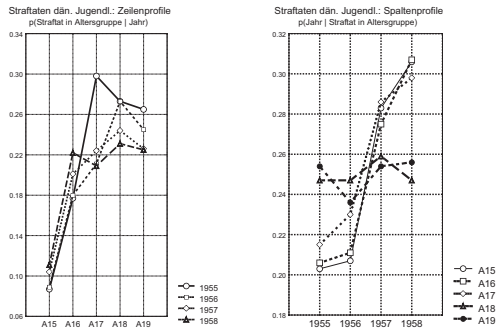
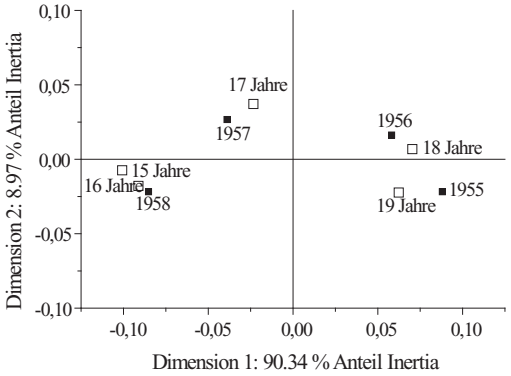


Abbildung 16: Dänische Jugendkriminalität und Polizeiverhalten: Biplot



Gruppe die meisten Straftaten begangen, während diese Gruppe im darauffolgenden Jahr 1957 die nach den 15-jährigen geringste Anzahl von Straftaten begeht!

Die (Spalten-)Profile für die Gruppe der 15- und 16-jährigen sind sehr ähnlich: in den Jahren 1955 und 1956 begehen diese Gruppen die wenigsten Straftaten, im Jahr 1957 begehen dann die 15- und 16-jährigen schon mehr als die 18- und 19-jährigen. Im Jahr 1958 haben sie dann den Abstand zu den 18- und 19-jährigen noch einmal vergrößert. Die Anzahl der Straftaten, die von 18- und 19-jährigen begangen werden, ist über die Jahre 1955 bis 1958 zwar nicht konstant, die relative Häufigkeit schwankt aber um einen Wert von $\approx .25$. Die Profile der 15/16-jährigen einerseits und der 18/19-jährigen andererseits bilden zwei klar unterschiedene Klassen von Profilen.

Das Profil der 17-jährigen ist zwar dem der 15/16-jährigen ähnlich, nähert sich aber insbesondere für das Jahr 1956 dem der 18/19-jährigen an; es liegt näher an dem Profil, das sich ergibt, wenn man über alle Profile mittelt.

Betrachtet man nun die Repräsentation der Jahre im 2-dimensionalen Koordinatensystem, das von der Korrespondenzanalyse geliefert wird, so ergibt die Projektion der den Jahren entsprechenden Punkte gerade die natürliche Ordnung der Jahre von 1955 bis 1958. Dies korrespondiert zu dem den Spaltenprofilen zu entnehmenden Befund, daß die 15/16-jährigen einen mit den Jahren monoton steigenden Anteil an den Straftaten haben; 1955 und 1956 gehen weniger Straftaten auf das Konto dieser Altersgruppen als auf das der 18/19-jährigen, und 1957 und 1958 deutlich mehr als auf das der 18/19-jährigen.

Betrachtet man nun die Repräsentation der Altersgruppen im 2-dimensionalen Koordinatensystem, so sieht man, daß die erste Dimension durch die Gruppen der 15/16-jährigen auf der einen Seite und die der 18/19-jährigen auf der anderen Seite definiert wird. Man könnte meinen, daß eine derart ausgeprägte Bipolarität eine ebenso ausgeprägte Gegenläufigkeit der Profile voraussetzt, aber dies ist offenbar nicht so. Die Altersgruppen korrespondieren zu den Jahreszahlen in der Weise, daß die Position auf der Achse den großen Häufigkeiten entspricht: die 15/16-jährigen haben 1958 den größten Anteil an den Straftaten, und 1955 haben die 18/19-jährigen den größten Anteil.

Die zweite Dimension wird im wesentlichen durch die Gruppe der jeweils 17-jährigen charakterisiert. Interessant ist die ähnlichkeit der Position der Gruppe der 17-jährigen zu der des Jahres 1957. Das (Zeilen-) Profil des Jahres 1957 entspricht am ehesten dem mittleren Zeilenprofil, ebenso wie das (Spalten-) Profil der 17-jährigen dem mittleren Profil der Altersgruppen am nächsten kommt. Aber diese Nähe zum mittleren Profil definiert eigentlich nur den Abstand des 1957-Punktes vom Ursprung des Koordinatensystems, noch nicht die Nähe zur Gruppe der 17-jährigen. Hierzu muß man sich daran erinnern, daß die Beziehung zwischen den Koordinaten der Zeilen- und Spaltenpunkte nicht in Termen der euklidischen Distanz zwischen den jeweiligen Punkten interpretiert werden darf. Vielmehr ist die Beziehung (6.23), d.h.

$$P = D_r F \Lambda^{-1/2} G' D_c + E,$$

der hier relevante Bezugspunkt. Hier werden die relativen Häufigkeiten der Kontingenztabelle - und damit die Häufigkeiten, denn $NP = K$ - anhand der Koordinaten F und G "zurück"gerechnet. Zur Erinnerung sei angemerkt, daß das Skalarprodukt zwischen den Vektoren für einen Zeilen- und einen Spaltenpunkt noch nicht hinreichend ist, denn es muß ja der Unterschied zwischen der Euklidischen Metrik einerseits und der χ^2-Metrik andererseits berücksichtigt werden. Schaut man nun in die Datentabelle 17, so sieht man, daß die Gruppe der 17-jährigen im Jahre 1957 im Vergleich zu den anderen Jahren die meisten Straftaten begangen hat. Allerdings hat die Gruppe der in diesem Jahr 18-jährigen im Jahr 1957 ebenfalls die meisten Straftaten, wieder im Vergleich zu den anderen Jahren, begangen, und der Punkt für die Gruppe der 18-jährigen liegt nicht nahe bei dem für 1957. Dies erscheint nur auf den ersten Blick widersprüchlich: um die Relation zwischen den Kategorien zu deuten müssen auch die *Massen* der Kategorien berücksichtigt werden; in der Beziehung zwischen F, G und P gehen sie über die Diagonalmatrizen D_r und D_c ein. So ist die Masse für das Jahr 1957 kleiner als die für das Jahr 1958, und die Masse für die Altersgruppe A 17 ist kleiner als die für A 18 bzw. A 19.

Abbildung 15 zeigt noch den Biplot für die Daten. Andersen analysiert die Daten zusätzlich anhand von log-linearen Modellen und kann so spezifische Hypothesen über den Zusammenhang zwischen Alter und Jahr testen; hierauf kann an dieser Stelle nicht eingegangen werden. □

5.5.2 Veränderungen von Meinungen

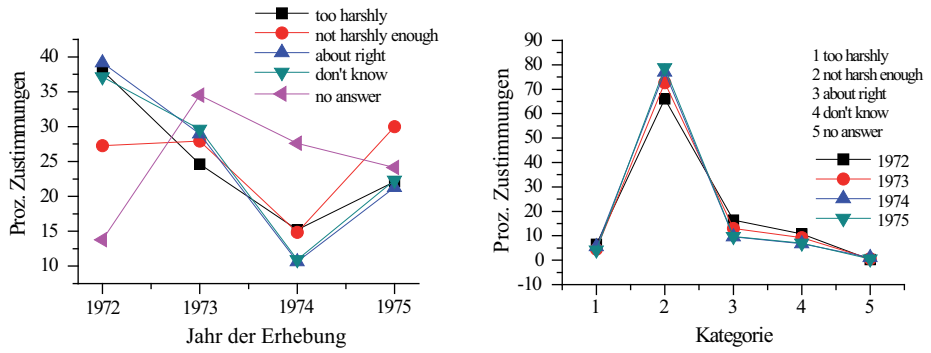
Beispiel 5.6 Die Daten der Tabelle 19 reflektieren die Verteilung von Ansichten über die Behandlung Krimineller in den USA in verschiedenen Jahren[21]. Das

Tabelle 19: Ansichten zur Behandlung Krimineller in Gefängnissen (USA)

	Jahr				
Ansicht	1972	1973	1974	1975	$\sum$
Too harshly (t.h.)	105	68	42	61	276
Not harshly enough (n.h.e.)	1066	1092	580	1174	3912
About right (a.r.)	265	196	72	144	677
Don't know (d.k.)	173	138	51	104	466
No answer (n.a.)	4	10	8	7	29
$\sum$	1613	1504	753	1490	5360
$\chi^2 = 87.360$, df $= 12$, $p = .0000$, $In(K) = \chi^2/N = .016298$					

[21] aus: Haberman, SJ Analysis of qualitative data, Vol. I, p. 120; National Opinion Research Center 1972-1975

Abbildung 17: Ansichten über die Behandlung von Kriminellen in US-Gefängnissen: Profile



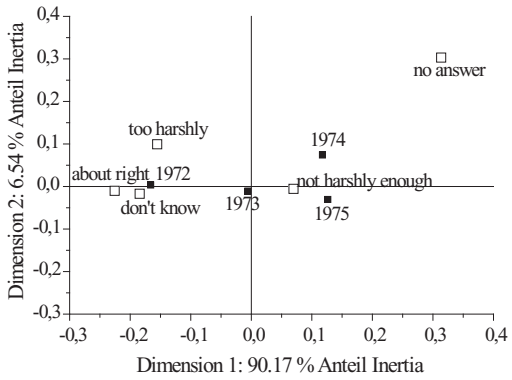
χ^2 für die Tabelle ist hochsignifikant, es muß also einen Wechsel in den Häufigkeiten, mit denen bestimmte Einstellungen vertreten werden, mit den Jahren geben. Die Korrespondenzanalyse liefert die Eigenwerte $\lambda_1 = .01469$, $\lambda_2 = .0011$ und $\lambda_3 = .000536$, $\sum_i \lambda_i = .01633$, so daß durch die erste Dimension .899 der Gesamt-Inertia χ^2/N erklärt werden, durch die zweite nur noch .067. Es sollen zuerst die Profile betrachtet werden (vergl. Abb. 17). Die Zeilenprofile - d.h. die (relativen) Häufigkeiten der Einstellungen für ein gegebenes Jahr - sind insgesamt sehr ähnlich, aber nicht streng parallel: die Kurven überschneiden sich. Insgesamt scheinen sich die Befragten sehr darüber einig zu sein, daß die Inhaftierten nicht streng genug (not harshly enough, n.h.e.) behandelt werden. 1973 ist das Jahr, das dem Durchschnitt der Häufigkeiten für diese Einstellung entspricht. "Zu streng" (too harshly, t.h.) ist eine Einstellung, deren relative Häufigkeit, gegeben ein bestimmtes Jahr, für alle Jahre nahezu konstant ist. Eine analoge Aussage gilt für "keine Antwort" (no answer, n.a.).

Bei den Spaltenprofilen weicht das Profil für "keine Antwort" (n.a.) allerdings deutlich von den anderen Profilen ab; es ist gewissermaßen gegenläufig. Für das Jahr 1972 kommt diese Einstellung am wenigsten häufig vor, im Gegensatz zu den anderen Einstellungen, die bis auf "nicht streng genug" (n.h.e.), die mit mittlerer Häufigkeit vorkommt, hier hier maximale Häufigkeit haben. Im Jahr 1973 ist dann die relative Häufigkeit von "keine Antwort" maximal. Man kann vermuten, daß die Einstellungen "zu streng" (t.h.), "einigermaßen richtig" (a.r.) und "weiß nicht" (d.k.) im Biplot einigermaßen nahe beieinander liegen werden, und "nicht streng genug" und "keine Antwort" weiter von dieser Gruppe entfernt liegen werden. Vermutlich wird das Jahr 1972 eine Position eine Position nahe bei der ersten Gruppe von Einstellungen haben und 1975 eine Position nahe bei der zweiten Gruppe. Die Tabelle 20 enthält die Koordinaten und Qualitätsmaße für die Daten. Es sei zunächst ein Blick auf die Qualitäten q_i bzw. q_j geworfen. Alle Werte sind kleiner als 1, d.h. die Kategorien werden durch die ersten beiden Dimensionen nicht perfekt abgebildet. Insbesondere die Einstellung "ungefähr rich-

Tabelle 20: Koordinaten und Statistiken: Einstellung gegenüber Kriminellen in den USA

Einstellung	f_{i1}	f_{i2}	$\pi_{i\cdot 1}$	$\cos^2\Theta_{11}$	$\pi_{i\cdot 2}$	$\cos^2\Theta_{21}$	$\rho_{i\cdot}$	q_i
too harshly	-.155	.099	.085	.638	.477	.261	.119	.899
not h. enough	.069	-.005	.241	.993	.021	.006	.218	.999
about right	-.226	-.010	.437	.995	.012	.002	.396	.007
don't know	-.184	-.017	.201	.972	.024	.008	.186	.980
no answer	.314	.303	.036	.413	.467	.386	.079	.798
Jahr	g_{j1}	g_{j2}	$\pi_{\cdot j1}$	$\cos^2\Theta_{12}$	$\pi_{\cdot j2}$	$\cos^2\Theta_{22}$	$\rho_{\cdot j}$	q_j
1972	-.166	.003	.566	.991	.003	.000	.514	.992
1973	-.006	-.011	.001	.022	.030	.079	.025	.101
1974	.118	.074	.132	.715	.725	.285	.166	.999
1975	.126	-.030	.302	.926	.241	.054	.294	.980
f_{ik}, g_{jk}	Koordinaten der Kategorien							
$\pi_{i\cdot k}$, $\pi_{\cdot jk}$	relative Inertia: Anteil d. Inertia einer Kategorie für k-te Dim.							
$\cos^2\Theta$	Koord'anteile: Projekt. eines Datenvektors auf eine Achse							
$\rho_{i\cdot}$, $\rho_{\cdot j}$	Anteil der Kategorie an $In(K)$							
q_i, q_j	Qualität							

Abbildung 18: Ansichten über die Behandlung von Kriminellen in US-Gefängnissen: Biplot

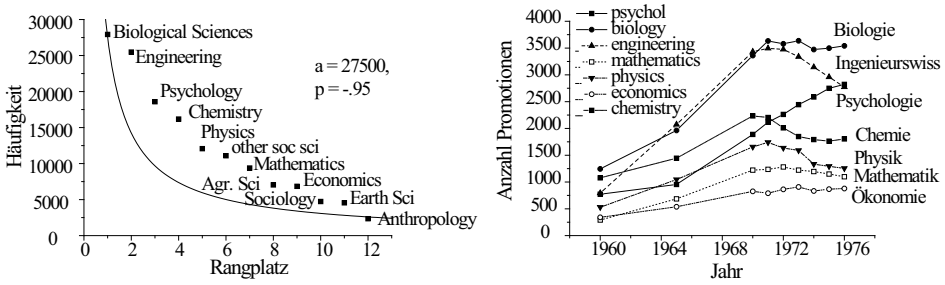


tig" (a.r.) und das Jahr 1973 werden schlecht abgebildet; bei den anderen Jahren und Einstellungen ist die Abbildung aber zufriedenstellend. Betrachet man andererseits die $\cos^2$-Werte, so sieht man, daß sie sich für alle Einstellungen und Jahre für die beiden ersten Dimensionen fast perfekt zu ergänzen, d.h. diese Dimensionen erklären die Beiträge der Kategorien zum Gesamt-χ^2 so gut wie vollständig; es ist also nicht notwendig, weitere Dimensionen zu betrachten. Schlüsselt man das χ^2 hinsichtlich der Zeilenkategorien (Einstellungen) auf, so sieht man (Spalte ρ_i), daß "ungefähr richtig" (a.r.) den größten Anteil erklärt, während "keine Antwort" nur einen vernachlässigbaren Anteil erzeugt. Dem entspricht, daß sich die Werte der Zeilenprofile für diese Einstellung über die Jahre so gut wie nicht unterscheiden. Bei den Jahreskategorien ist es das Jahr 1972, das die größten Unterschiede generiert, gefolgt von 1975. Schließlich kann man sich noch einen überblick darüber verschaffen, welche Kategorien durch welche Dimension am besten "erklärt" werden. Für die erste Dimension ist der χ^2-Anteil, der auf die erste Dimension zurückgeführt werden kann, für dieKategorie "ungefähr richtig" am größten; "keine Antwort" hat einen sehr kleinen Anteil. Bei den Jahren ist es wieder 1972, für das die erste Dimension den größten Erklärungswert hat, gefolgt von 1975. Die zweite Dimension erklärt am besten die Kategorie "zu streng" (t.h.), gefolgt von "keine Antwort", so wie die Jahre 1974 und 1975.

Betrachtet man nun den Biplot, so sieht man, daß das Zeilenprofil für 1973 ziemlich genau dem Durchschnittsprofil entspricht: der Punkt für dieses Jahr liegt im Ursprung des Koordinatensystems. Betrachtet man die Projektionen der Jahre auf die erste Achse, so sieht man, daß die Jahre auf dieser Achse gemäß ihrer natürlichen Ordnung abgebildet werden. Wie schon aufgrun dder Profile vermutet wurde, liegen die Einstellungen "ungefähr richtig", "weiß nicht" und "zu streng" nahe bei dem Punkt für 1972; das Jahr 1972 entspricht noch am ehesten der "milderen" Auffassung. Die Einstellung "nicht streng genug" (n.h.e.) dagegen liegt bei den Jahren 1974 und 1975. Man kann also vermuten, daß mit den Jahren ein Trend zu den strengeren Einstellungen erfolgt ist; in der Tat entnimmt man ja den Spaltenprofilen, daß die bedingte relative Häufigkeit für die Einstellung "nicht streng genug" im Jahr 1975 im Vergleich zu allen vorangegangenen Häufigkeiten maximal ist.

Die Kategorie "keine Antwort" nimmt eine Sonderstellung ein.Auf der ersten Dimension ist sie mehr als "nicht streng genug" ausgeprägt, auf der zweiten Dimension mehr als "zu streng". Betrachtet man die komplette 3-dimensionale Lösung (die Koordinaten für die dritte Dimension sind hier nicht aufgeführt worden), so findet man, daß auf der dritten Dimension alle Kategorien vernachlässigbare Koordinatenwerte haben, bis auf die Kategorie "keine Antwort", die hier noch den Wert -.221 hat. Die Reaktion, gar keine Antwort zu geben, reflektiert also vermutlich nicht die Einstellung "Ich weiß nicht", sondern den Wunsch, seine Ansicht nicht bekannt zu geben. Im "liberalen" Jahr 1972 wird diese Reaktion dann auch mit der geringsten Häufigkeit beobachtet, 1973 dann mit maximaler Häufigkeit; 1974 und 1975 wird die Häufigkeit dann wieder geringer. Mit dieser Kategorie wird also wohl weniger die Einstellung zu Kriminellen erfaßt, sondern die Einstellung darüber, ob man seine Ansicht bekannt geben soll. □

Abbildung 19: Verteilung der Häufigkeiten der Fächer



5.5.3 Trends in der Wissenschaft

Beispiel 5.7 Betrachtet man die Anzahlen der Promotionen in verschiedenen Fächern in aufeinander folgenden Jahren, so können sich Trends zeigen, die die Veränderung von Interessen oder ökonomischen Lagen abilden. Die *Statistical Abstracts of the Unites States, 1976, Table 958* liefern die Tabelle 21. die Verteilung der Häufigkeiten für die einzelnen Fächer.

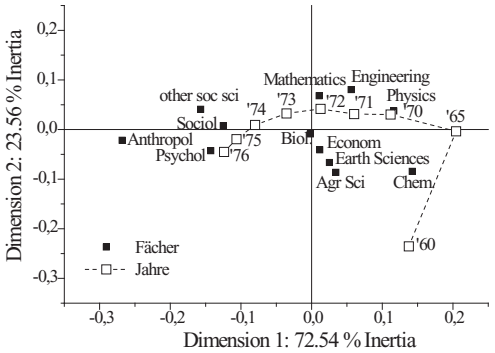
Tabelle 21: Trends bei Doktorgraden in den Jahren 1960 - 1976

	1960	1965	1970	1971	1972	1973	1974	1975	1976	Σ
Engineer.	794	2073	3432	3495	3475	3338	3144	2959	2773	25483
Mathem	291	685	1222	1236	1281	1222	1196	1149	1099	9381
Physics	530	1046	1655	1740	1635	1590	1334	1293	1254	12077
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762	1804	16178
Earth Sci	253	375	511	550	580	577	570	556	584	4556
Biol. Sci	1245	1963	3360	3633	3580	3636	3473	3498	3541	27929
Agric. Sci	414	576	803	900	855	853	830	904	908	7043
Psychology	772	954	1888	2116	2262	2444	2587	2749	2822	18594
Sociology	162	239	504	583	638	599	645	680	687	4737
Economy	341	538	826	791	863	907	833	867	879	6845
Anthropology	69	82	217	240	260	324	381	385	394	2352
other soc sci	314	502	1079	1392	1500	1609	1531	1550	1616	11093
Σ	6263	10477	17731	18880	18940	18948	18316	18352	18361	146268

Den Zeilensummen der Tabelle entnimmt man schnell, dass es offensichtlich bevorzugte Fächer wie Ingenieurswissenschaften, Biologie und Psychologie gibt, aber die zeitliche Dynamik der Entwicklung der Trends bleibt einer direkten Betrachtung der Tabelle verschlossen. Abb. 19 zeigt

Interessant ist das Ansteigen der Doktorate in der Psychologie, während insbesondere die Ingenieurswissenschaften, Mathematik und Physik ein Rückgang ab etwa 1970 bzw. 1971 zu beobachten ist; die Chemie scheint sich zu stabilisieren. Die Abbildung Häufigkeit versus Rangplatz der Fächer zeigt die über die Jahre

Abbildung 20: Biplot Doktorate - Jahre



"gemittelten" (d.h. aggregierten) Häufigkeiten für die Fächer. Die Kurve ist die Pareto-Kurve (vergl. Beispiel 5.8); diese Daten sind ein eher seltenes Beispiel für den *mangelnden* Fit des Pareto-Modells.

Die eigentlich bemerkenswerten Strukturen, die die Tabelle 21 enthält, zeigt der Biplot. Es wird ein klarer zeitlicher Verlauf sichtbar, der eine Veränderung der Präferenzen anzeigt. □

5.5.4 Trends bei Selbstmorden

Beispiel 5.8 Heuer (1979) hat Daten über Art und Anzahl von Selbstmorden in Westdeutschland vorgelegt, anhand derer die Anwendung der Korrespondenzanalyse auf mehr als 2-dimensionale Kontingenztabellen illustriert werden kann. Tabelle 22 enthält die Daten. Diese Tabelle ist 3-dimensional, die Faktoren sind Geschlecht, Altersgruppe und Methode. Die Korrespondenzanalyse ist nur auf 2-dimensionale Kontingenztabellen anwendbar. Möchte man eine Korrespondenzanalyse dieser Daten durchführen, so gibt es zwei Möglichkeiten:

1. Man aggregiert über einen der Faktoren, z.B. über das Geschlecht,

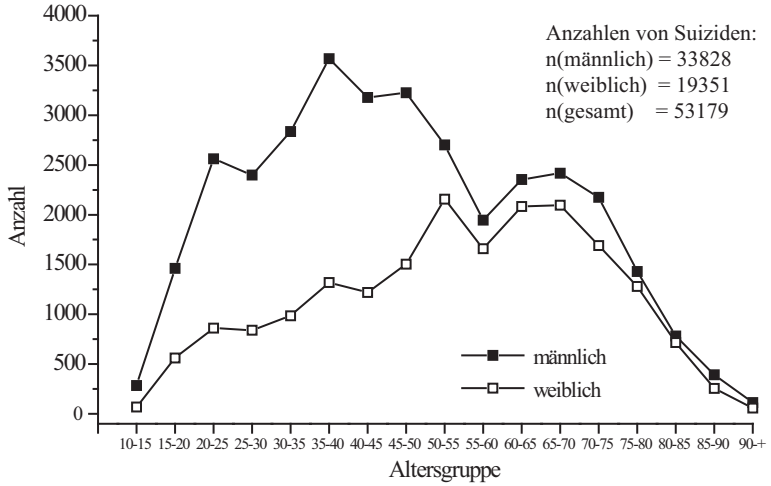
2. Man macht aus der Tabelle eine 2-dimensionale Tabelle, indem man einen der Kategoriensätze wie zwei behandelt; so kann man die Tabelle so nehmen, wie sie angeschrieben wurde, und betrachtet die Altersgruppen für die Kategorie "männlich" als Altersgruppen/männlich und die für die Kategorie "weiblich" als Altersgruppe/weiblich. Die Altersgruppen erscheinen dann zweimal im Biplot. Alternativ kann man die Daten für die Frauen neben die der Männer schreiben und erhellt so eine Tabelle, in der die Methoden zweimal erscheinen, einmal als Methoden/männlich und einmal als Methoden/weiblich.

Die Aggregation über eine Kategorienklasse, etwa Geschlecht, setzt voraus, daß sich Frauen und Männer in ihrem Suicidverhalten nicht signifikant unterscheiden.

Tabelle 22: Selbstmorde in Westdeutschland 1974-1977

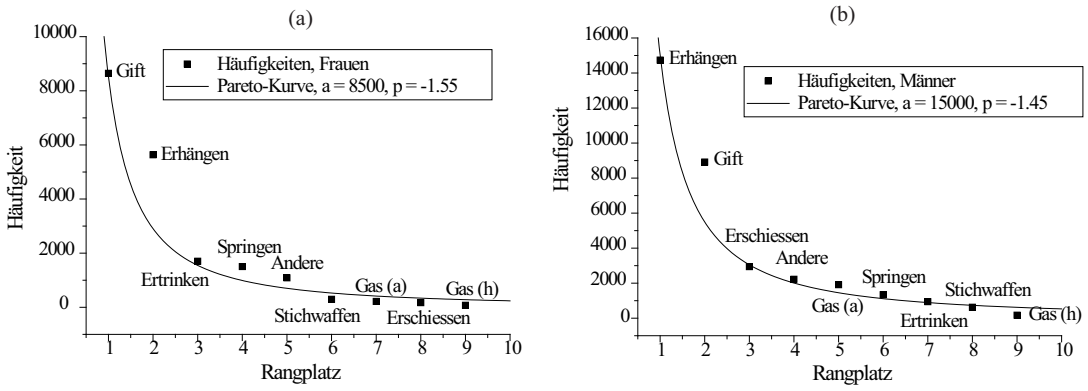
Alter/männl	Materie	Gas (h)	Gas (a)	Hängen	Ertrinken	Schußw.	Stichw.	Springen	Andere
10-15	4	0	0	247	1	17	1	6	9
15-20	348	7	67	578	22	179	11	74	175
20-25	808	32	229	699	44	316	35	109	289
25-30	789	26	243	648	52	268	38	109	226
30-35	916	17	257	825	74	291	52	123	281
35-40	1118	27	313	1278	87	293	49	134	268
40-45	926	13	250	1273	89	299	53	78	198
45-50	855	9	203	1381	71	347	68	103	190
50-55	684	14	136	1282	87	229	62	63	146
55-60	502	6	77	972	49	151	46	66	77
60-65	516	5	74	1249	83	162	52	92	122
65-70	513	8	31	1360	75	164	56	115	95
70-75	425	5	21	1268	90	121	44	119	82
75-80	266	4	9	866	63	78	30	79	34
80-85	159	2	2	479	39	18	18	46	19
85-90	70	1	0	259	16	10	9	18	10
90+	18	0	1	76	4	2	4	6	2
Alter/weibl	Materie	Gas (h)	Gas (a)	Hängen	Ertrinken	Schußw.	Stichw.	Springen	Andere
10-15w	28	0	3	20	0	1	0	10	6
15-20w	353	2	11	81	6	15	2	43	47
20-25w	540	4	20	111	24	9	9	78	67
25-30w	454	6	27	125	33	26	7	86	75
30-35w	530	2	29	178	42	14	20	92	78
35-40w	688	5	44	272	64	24	14	98	110
40-45w	566	4	2	343	76	18	22	103	86
45-50w	716	6	24	447	94	13	21	95	88
50-55w	942	7	26	691	184	21	27	129	131
55-60w	723	3	14	527	163	14	30	92	92
60-65w	820	8	8	702	245	11	35	140	114
65-70w	740	8	4	785	271	4	38	156	90
70-75w	624	6	4	610	244	1	27	129	46
75-80w	495	8	1	420	161	2	29	129	35
80-85w	292	3	2	223	78	0	10	84	23
85-90w	113	4	0	83	14	0	6	34	2
90+w	24	1	0	19	4	0	2	7	0

Abbildung 21: Häufigkeitsverteilung der Selbstmorde, aggregiert über Methoden



Aggregiert man trotz signifikanter Interaktionen Geschlecht × Altersgruppe, Geschlecht × Methode oder gar Geschlecht × Altersgruppe × Methode, so erhält man ein falsches Bild. Dieser Sachverhalt wirft ein neues Licht auf die Analysen zweidimensionaler Tabellen; ihre Analyse liefert nun dann ein adäquates Bild der Beziehungen zwischen den Zeilen- und Spaltenkategorien, wenn diese Kategorien nicht mit anderen, nicht explizit erfaßten Kategorien interagieren. Natürlich kann dies nur selten ausgeschlossen werden, so daß man bei Interpretationen von Analysen die Randbedingung "bis auf Interaktionen mit anderen Kaegorien" nicht vergessen sollte. Abb. 21 zeigt die Häufigkeiten der Selbstmorde getrennt nach Geschlechtern, aber aggregiert über die Methoden. Generell gilt, daß die Anzahl der männlichen Selbstmörder für alle Altersgruppen höher ist als die der weiblichen. Bis zum 52-ten Lebensjahr hat die Verteilung der Selbstmorde bei Männern eine andere Gestalt als die entsprechende Verteilung bei den Frauen. Bei den Männern. In der Gruppe der 35 - 40-jährigen Männer ist die Anzahl am höchsten, während die Maximalzahl der Selbstmorde bei den Frauen in der Gruppe der 50-55-jährigen liegt. Bei den Männern hat die Häufigkeitsverteilung bei den 55-60-jährigen ein lokales Minimum, um dann zu einem neuen,lokalen Maximum für die Gruppe der 65-70-jährigen anzusteigen. Auffallend ist, daß die Verteilungen für die Frauen und die Männer von der Gruppe der 55-60-jährigen an nahezu parallel verläuft. Abb. 22 zeigt die Häufigkeiten, mit denen die verschiedenen Methoden gewählt werden; die Abbildung entspricht einer Darstellung der Spaltenprofile. Allerdings wurden hier die Methoden in bezug auf die Häufigkeit, mit der sie gewählt wurden, ranggeordnet: an Platz 1 steht die am häufigsten gewählte Methode, an Platz 2 die am zweithäufigsten gewählte Methode, etc. Die durch die Punkte gelegte Kurve ist die *Pareto-Kurve*; sie ist durch die Funktion $n(r) = ar^p$ definiert. Dabei ist $n(r)$ die Häufigkeit, mit die an Rangplatz r stehende Methode zur Anwendung kam, unad a und p sind freie Parameter. Pa-

Abbildung 22: Ranggeordnete Häufigkeit der Methoden, (a) Frauen, (b) Männer.Zur Definition der Pareto-Kurve siehe Text.

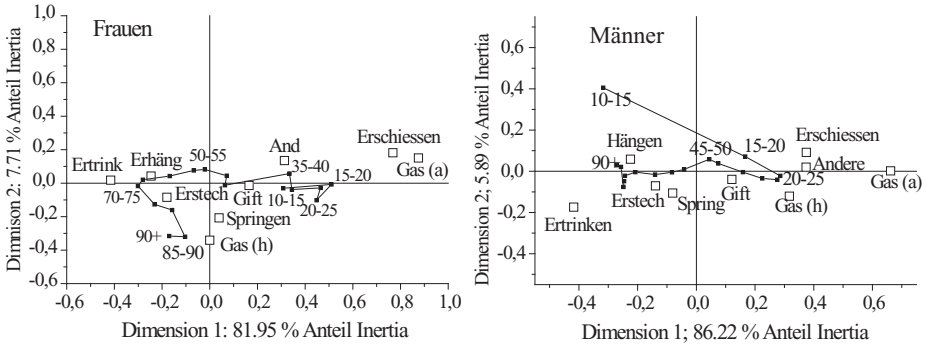


reto war ein italienischer Nationökonom, der in dieser Form das Einkommen von Bürgern eines Staates gegen den Rangplatz auftrug und vermutete, dass die politische Lage in einem Staat um so instabiler sei, je steiler diese Kurve abfällt; diese Vermutung hat sich in dieser einfachen Form nicht bestätigt. In der Linguistik ist die Kurve als Zipfsches Gesetz bekannt: trägt man die Häufigkeit, mit der Worte in einer Sprache verwendet werden, gegen ihren entsprechenden Rangplatz auf, so entspricht die entstehende Kurve bei den meisten Sprachen einer Pareto-Formel. Die Formel gilt in vielen anderen Fällen ebenfalls in guter Näherung; betrachtet man z.B. die Anzahl der Prüfungen, die von einem Professor abgenommen werden, und rangordnet man die Professoren nach ihrer Prüfungsbelastung, so entsteht häufig wieder eine Pareto-Kurve. Auch bei den hier betrachteten Daten scheint die Pareto-Kurve in guter Näherung den Daten zu entsprechen. Die Eruierung der theoretischen Implikationen der Kurve für die Suicidhäufigkeit in einer Gesellschaft wird den Leserinnen und Lesern als Denksportaufgabe überlassen.

Es wird jedenfalls noch einnmal deutlich, daß Gift und Strick die am häufigsten verwendeten Methoden sind, aber bei den Männern der Strick, bei den Frauen das Gift dominiert. Bei den Männern folgen dann das Gift und dann die Schußwaffen, während bei den Frauen der Strick und das Ertrinken folgen. Frauen und Männer unterscheiden sich also nicht nur in der Gesamthäufigkeit, sondern auch hinsichtlich der Häufigkeiten der gewählten Methoden.

Die Frage ist nun, wie die Daten einer Korrrespondenzanalyse unterzogen werden können, denn einerseits ist die Datenmatrix 3-dimensional, und andererseits setzt die Korrespondenzanalyse eine 2-dimensionale Tabelle voraus. Eine erste Möglichkeit besteht darin, über das Geschlecht zu mitteln, d.h. zu "aggregieren". Dies setzt voraus, daß es keine Interaktion zwischen dem Faktor Geschlecht und einem der beiden anderen Faktoren oder mit den beiden anderen Faktoren gibt, was aber wegen der in Abb. 22 aufscheinenden Wechselwirkung zwischen Metho-

Abbildung 23: Biplots Methode × Altersgruppen, für Frauen und Männer getrennt.



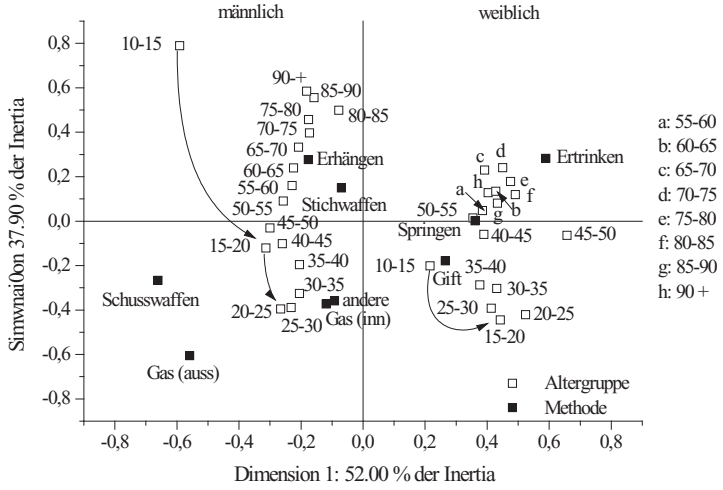
de und Geschlecht keine vernünftige Annahme zu sein scheint, wie im Übrigen durch eine log-lineare Analyse der Daten bestätigt wird: bei dieser Analyse ergibt sich die Signifikanz der Interaktion Geschlecht × Alter × Methode. In Abb. 23 wird der Biplot Alter × Methode, männlich bzw. weiblich, gezeigt. Der Verlauf der Altersgruppen "durch" die Methoden ist für die beiden Geschlechter offenbar unterschiedlich. In den Biplots wird die Beziehung jeder Methode zu den Altersgruppen gezeigt, welche Methode aber eher männlich und welche eher weiblich ist, wird weniger deutlich.

Ene Möglichkeit, die Gesamtdaten einer Korrespondenzanalyse zu unterziehen, ist, die Daten der Frauen neben die der Männer zu schreiben. Die Tabelle hat dann so viele Zeilen, wie es Altersgruppen gibt, und doppelt so viele Spalten, wie es Methoden gibt. Die Methoden erscheinen zweimal, einmal für die Männer und einmal für die Frauen. Die Korrespondenzanalyse versucht nun, für die Zeilen (Altersgruppen) Skalenwerte zu finden, die zu optimal zu bestimmten Spaltenkategorien korrespondieren. Diese sind dann Methoden, die für eines der Geschlechter charakteristisch sind. Abb. 24 zeigt den entsprechenden Biplot. Die Korrespondenzanalyse liefert eine klare Trennung der beiden Gruppen "männlich" und "weiblich". Insbesondere für die Männer ergibt sich eine klare Struktur der Altersgruppen. Abb. 23 zeigt, dass es auch für die Frauen eine Struktur der Altersgruppen gibt, sie ist aber weniger deutlich als die für die Männer. Das Bemerkenswerte an Abb. 24 ist, dass die Frauen und Männer nach Methoden getrennt werden; die Analyse zeigt im Unterschied zu den Einzelanalysen für Frauen und Männer in Abb. 23, dass eben einige Methoden für die Männer, und andere für die Frauen charaktistisch sind.

Es wird deutlich, daß die Korrespondenzanalyse ein wesentlich detaillierteres Bild der in der Tabelle 22 verborgenen Zusammenhänge liefert als eine bloße loglineare Analyse, die einem nur signalisiert, daß nur *ein* Modell akzeptabel ist, nämlich dasjenige, das eine Interaktion Geschlecht × Alter × Methode postuliert.

□

Abbildung 24: Biplot: Selbstmorde: Methode, Altersgruppen und Geschlecht



6 Anhang

6.1 Beweise

6.1.1 Gleichung (3.23)

Beweis: Es ist

$$\chi^2/N = \sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2$$

gemäß (**??**), und weiter

$$\chi^2/N = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{r_i c_j}(p_{ij} - r_i c_j)^2$$

nach (3.8). Also folgt

$$\begin{aligned}
\chi^2/N &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{r_i c_j}\left(\frac{p_{ij} r_i}{r_i} - r_i c_j\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{r_i^2}{r_i c_j}\left(\frac{p_{ij}}{r_i} - c_j\right)^2 \\
&= \sum_{i=1}^{I} r_i \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - c_j\right)^2 \\
&= \sum_{i=1}^{I} r_i \delta_{iv}^2
\end{aligned}$$

□

6.1.2 Satz 3.1

Beweis: Aus $F = D_r^{-1/2} U \Lambda^{1/2}$ folgt, daß das Element f_{is} von F durch

$$f_{is} = u_{is}\sqrt{\lambda_s}/\sqrt{r_i},$$

u_{is} das Element in der i-ten Zeile und s-ten Spalte von U, gegeben ist. Dann folgt weiter

$$FV' = D_r^{-1/2} U \Lambda^{1/2} V' = D_r^{-1/2} X$$

nach (**??**). Dann ist

$$\frac{x_{ij}}{\sqrt{r_i}} = \sum_{s=1}^{r} f_{is} v_{js}$$

und

$$\frac{x_{kj}}{\sqrt{r_k}} = \sum_{s=1}^{r} f_{ks} v_{js}$$

so daß

$$\frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{r_k}} = \sum_{s=1}^{r} (f_{is} - f_{ks}) v_{js}.$$

Aber es ist

$$\begin{aligned} \frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{r_k}} &= \frac{1}{\sqrt{r_i}} \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} - \frac{1}{\sqrt{r_k}} \frac{p_{kj} - r_k c_j}{\sqrt{r_k c_j}} \\ &= \frac{1}{\sqrt{c_j}} \left(\frac{p_{ij}}{r_i} - c_j \right) - \frac{1}{\sqrt{c_j}} \left(\frac{p_{kj}}{r_k} - c_j \right) \\ &= \frac{1}{\sqrt{c_j}} \left(\frac{p_{ij}}{r_i} - \frac{p_{ik}}{r_k} \right). \end{aligned} \tag{6.1}$$

Dann ist aber

$$\sum_{j=1}^{J} \left(\frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{r_k}} \right)^2 = \sum_{j=1}^{J} \frac{1}{c_j} \left(\frac{p_{ij}}{r_i} - \frac{p_{kj}}{r_k} \right)^2 = \delta_{ik}^2 \tag{6.2}$$

und also

$$\begin{aligned} \delta_{ik}^2 &= \sum_{j=1}^{J} \left(\sum_{s=1}^{r} (f_{is} - f_{ks}) v_{js} \right)^2 \\ &= \sum_{s=1}^{r} (f_{is} - f_{ks})^2 \sum_{j=1}^{J} v_{js}^2 \\ &+ 2 \sum_{j=1}^{J} \sum_{s<s'} (f_{is} - f_{ks})(f_{is'} - f_{ks'}) v_{js} v_{js'}. \end{aligned} \tag{6.3}$$

Aber

$$\begin{aligned} & 2\sum_{j=1}^{J}\sum_{s<s'}(f_{is}-f_{ks})(f_{is'}-f_{ks'})v_{js}v_{js'} \\ = \; & 2\sum_{s<s'}[(f_{is}-f_{ks})(f_{is'}-f_{ks'})\sum_{j=1}^{J}v_{js}v_{js'}] = 0 \end{aligned}$$

denn $\sum_j v_{js}v_{js'} = 0$ wegen der Orthogonalität der Eigenvektoren B_s und $B_{s'}$. Weiter ist $\sum_j v_{js}^2 = 1$ wegen der Normiertheit der Eigenvektoren B_s. Damit steht links in (6.3) die χ^2-Distanz zwischen der i-ten und der k-ten Zeile, und rechts die entsprechende euklidische Distanz bezüglich der Koordinaten F. □

6.1.3 Zerlegung des χ^2

Beweis: Es ist $y_{jk} = \sum_i x_{ij}x_{ik}$ und

$$\sum_{i=1}^{I} x_{ij}x_{ik} = \frac{1}{N}\sum_{i=1}^{I}\left(\frac{n_{ij}-n_{i.}n_{.j}/N}{\sqrt{n_{i.}n_{.j}/N}}\right)\left(\frac{n_{ik}-n_{i.}n_{.k}/N}{\sqrt{n_{i.}n_{.k}/N}}\right)$$

Für $j = k$ folgt sofort

$$y_{jj} = \frac{1}{N}\sum_{i=1}^{I}\frac{(n_{ij}-n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N} = \chi^2_{.j}/N \tag{6.4}$$

Andererseits ist nach (**??**) $X'X = V\Lambda V'$, und dementsprechend ist

$$y_{jk} = \sum_{s=1}^{r}\lambda_s v_{js}v_{ks},$$

und für $j = k$ folgt wiederum $y_{jj} = \sum_s \lambda_s v_{js}^2$, so daß (3.47) folgt.

Weiter ist $\tilde{y}_{il} = \sum_j x_{ij}x_{il}$, und

$$\sum_j x_{ij}x_{il} = \sum_j\left(\frac{n_{ij}-n_{i.}n_{.j}/N}{\sqrt{n_{i.}n_{.j}/N}}\right)\left(\frac{n_{il}-n_{l.}n_{.j}/N}{\sqrt{n_{l.}n_{.j}/N}}\right) \tag{6.5}$$

und für $i = l$ erhält man

$$\tilde{y}_{ii} = \sum_{j=1}^{J}\frac{(n_{ij}-n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N} = \chi^2_{i.}$$

Nach (**??**) ist $XX' = U\Lambda U'$ und mithin

$$\tilde{y}_{il} = \sum_{s=1}^{r}\lambda_s u_{is}u_{ls}$$

und somit

$$\tilde{y}_{ii} = \sum_{s=1}^{r} \lambda_s u_{is}^2 = \chi_{i.}^2.$$

Daraus folgt aber sofort

$$\sum_{i=1}^{I} \tilde{y}_{ii} = sp(X'X) = \frac{1}{N}\chi^2 = \sum_{s=1}^{r} \lambda_s$$

denn $\sum_i \sum_s \lambda_s u_{is}^2 = \sum_s \lambda_s$, da ja wegen der Normierung der A_s die Beziehung $\sum_i u_{is}^2 = 1$ gilt, und analog

$$\sum_{j=1}^{J} y_{jj} = sp(XX') = \frac{1}{N}\chi^2 = \sum_{s=1}^{r} \lambda_s$$

denn $\sum_j \sum_s \lambda_s v_{js}^2 = \sum_s \lambda_s$, ebenfalls wegen der Normierung der V_s.

Schließlich ist $F_s D_r F_s = U_s' \sqrt{\lambda_s}\sqrt{\lambda_s} U_s = U_s' U_s \lambda_s = \lambda_2$, denn die Eigenvektoren U_s sind auf die Länge 1 normiert. □

6.2 Weitere Ergebnisse

6.2.1 Die Beziehung zwischen den Koordinaten

Es wird zunächst der folgende Hilfssatz bewiesen:

Hilfssatz: Es gelten die Gleichungen

$$r'F = 0 \tag{6.6}$$
$$c'G = 0. \tag{6.7}$$

Beweis: Es sei $\vec{\mathbf{1}} = (1, 1, \ldots, 1)'$. Die Komponenten von $\vec{\mathbf{1}}$ sind also alle gleich 1. Die Dimension von $\vec{\mathbf{1}}$, d.h. die Anzahl der Komponenten, ergibt sich im Folgenden aus den entsprechenden Gleichungen, in denen $\vec{\mathbf{1}}$ auftritt.

Sicherlich gilt

$$D_r^{-1} r = \vec{\mathbf{1}}, \quad D_c^{-1} c = \vec{\mathbf{1}}, \tag{6.8}$$

wie man sich am Beispiel $J = 2$ leicht klar macht:

$$\begin{pmatrix} 1/r_1, & 0 \\ 0, & 1/r_2 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Der $\vec{\mathbf{1}}$-Vektor $D_r^{-1} r$ hat J Komponenten, $\vec{\mathbf{1}}$-der Vektor D_c^{-1} hat I Komponenten. Weiter gilt (vergl. (**??**))

$$r'\vec{\mathbf{1}} = \sum_i r_i = 1, \quad c'\vec{\mathbf{1}} = \sum_j c_j = 1. \tag{6.9}$$

Diese Gleichungen folgen sofort aus der Definition der r_i und c_j als relative Häufigkeiten.

Zum Beweis von (6.6) und (6.7) wird zunächst angemerkt, dass sich der Ausdruck für F auch in der Form $F = D_r^{-1}(P - rc')D_c^{-1}B$ anschreiben läßt: substituiert man hier $A\Lambda^{1/2}B'$ für $P - rc'$, so erhält man in der Tat

$$F = D_r^{-1}A\Lambda^{1/2}B'D_c^{-1}B = D_r^{-1}A\Lambda^{1/2}.$$

Andererseits ist

$$D_r^{-1}(P - rc')D_c^{-1}B = (D_r^{-1}P - D_r^{-1}rc')D_c^{-1}B = (D_r^{-1}P - \mathbf{1}c')D_c^{-1}B.$$

Man erhält auf diese Weise

$$\begin{aligned} F &= (D_r^{-1}P - \mathbf{1}c')D_c^{-1}B && (6.10)\\ G &= (D_c^{-1}P' - \mathbf{1}r')D_r^{-1}A, && (6.11) \end{aligned}$$

wobei der Ausdruck für G auf analoge Weise gewonnen wurde. Dann ist

$$r'F = r'(D_r^{-1}P - \mathbf{1}c')D_c^{-1}B.$$

Es ist aber $r'(D_r^{-1}P - \vec{\mathbf{1}}c') = \vec{\mathbf{1}}'P - r'\vec{\mathbf{1}}c'$, und weiter gilt $\vec{\mathbf{1}}'P = c'$, $r'\vec{\mathbf{1}} = 1$, so daß $\vec{\mathbf{1}}'P - r'\vec{\mathbf{1}}c' = 0$, nach (6.9). Damit folgt $r'F = 0$. Auf analoge Weise zeigt man $c'G = 0$.

□

Die Übergangsgleichungen. Es gilt

$$\begin{aligned} G &= D_c^{-1}P'F\Lambda^{-1/2}, && (6.12)\\ F &= D_r^{-1}PG\Lambda^{-1/2} && (6.13) \end{aligned}$$

Beweis: Aus $P - rc' = A\Lambda^{1/2}B'$ folgt

$$D_r^{-1}(P - rc')D_c^{-1} = D_r^{-1}A\Lambda^{1/2}B'D_c^{-1}, \tag{6.14}$$

bzw. in transponierter Form

$$D_c^{1/2}(P' - cr')D_r^{1/2} = D_c^{-1}B\Lambda^{1/2}A'D_r^{-1}. \tag{6.15}$$

Wegen $F = D_r^{-1}A\Lambda^{1/2}$, $G = D_c^{-1}B\Lambda^{1/2}$ erhält man daraus einerseits

$$D_r^{-1}PD_c^{-1} - D_r^{-1}rc'D_c^{-1} = FB'D_c^{-1}, \tag{6.16}$$

und andererseits

$$D_c^{-1}P'D_r^{-1} - D_c^{-1}cr'D_r^{-1} = GA'D_r^{-1}. \tag{6.17}$$

Multipliziert man (6.16) von rechts mit $B\Lambda^{1/2}$ und (6.17) mit $A\Lambda^{1/2}$, so erhält man wegen (3.60)

$$\begin{aligned} D_r^{-1}PD_c^{-1}B\Lambda^{1/2} - D_r^{-1}rc'D_c^{-1}B\Lambda^{1/2} &= F\Lambda^{1/2}\\ D_c^{-1}P'D_r^{-1}A\Lambda^{1/2} - D_c^{-1}cr'D_r^{-1}A\Lambda^{1/2} &= G\Lambda^{1/2}. \end{aligned}$$

Aber $D_c^{-1}B\Lambda^{1/2} = G$, und $D_r^{-1}A\Lambda^{1/2} = F$, und $c'G = 0$, $r'F = 0$, so daß

$$D_r^{-1}PG = F\Lambda^{1/2} \quad (6.18)$$
$$D_c^{-1}P'F = G\Lambda^{1/2} \quad (6.19)$$

folgt. Multiplikation von rechts mit $\Lambda^{-1/2}$ liefert dann gerade die Gleichungen (6.12) und (6.13). □

6.2.2 Die Koordinaten als Eigenvektoren

Das Ergebnis dieses Abschnitts ist u.a. hilfreich, wenn man die Beziehung zwischen der Korrespondenzanalyse und dem Dualen Skalieren betrachten will.

Nach (6.18) gilt $(D_c^{-1}P')F = G\Lambda^{1/2}$. Multiplikation von links mit $D_r^{-1}P$ liefert

$$D_r^{-1}P(D_c^{-1}P')F = D_r^{-1}PG\Lambda^{1/2}.$$

Nach (6.13) ist aber $D_r^{-1/2}PG\Lambda^{-1/2} = F$, d.h. aber $D_r^{-1/2}PG = F\Lambda^{1/2}$, so daß

$$(D_r^{-1}PD_c^{-1}P')F = F\Lambda \quad (6.20)$$

folgt. Demzufolge ergeben sich die Koordinaten F als Eigenvektoren der Matrix $D_r^{-1}P(D_c^{-1}P'$. Analog zeigt man, daß

$$(D_c^{-1}P'D_r^{-1}P)G = G\Lambda \quad (6.21)$$

gilt.

6.2.3 Die Rekonstitutionsformel

Die Rekonstitutionsformel erlaubt die Rekonstruktion der Datenmatrix aus den Skalenwerten.

Für den Wert x_{ij} der i-ten Zeile und der j-ten Spalte von X ergibt die Singularwertzerlegung (2.18) den Ausdruck

$$x_{ij} = q_{i1}\sqrt{\lambda_1}v_{j1} + q_{i2}\sqrt{\lambda_2}v_{j2} + \cdots + q_{in}\sqrt{\lambda_s}v_{jn}. \quad (6.22)$$

x_{ij} ergibt sich demnach als Skalarprodukt des i-ten Zeilenvektors $\tilde{\mathbf{q}}_i = (q_{i1}, \ldots, q_{in})'$ von Q und des j-ten Zeilenvektors $(\sqrt{\lambda_1}v_{j1}, \ldots, \sqrt{\lambda_r}v_{jr})'$ von $V\Lambda^{1/2}$. Andererseits sollen die Koordinaten der Kategorien durch die Elemente von F und G gegeben sein. Es ergibt sich die für die Interpretation der Ergebnisse wichtige Frage, ob sich die x_{ij} auch als Skalarprodukte der entsprechenden Zeilenvektoren von F und G darstellen lassen. Dies läßt sich leicht zeigen:

Es sei $P = (p_{ij})$ die Matrix der relativen Häufigkeiten $p_{ij} = n_{ij}/N$, und E die Matrix der erwarteten Häufigkeiten $e_{ij} = n_{i\cdot}n_{\cdot j}/N$. Dann gilt

$$X = D_r^{-1/2}(P - E)D_c^{-1/2}.$$

Wie im Anhang, Abschn. **??** gezeigt wird, ergibt sich hieraus unter Berücksichtigung von (3.29) und (3.30) die *Rekonstitutionsformel*

$$P = D_r F \Lambda^{-1/2} G' D_c + E. \tag{6.23}$$

Es sei noch darauf hingewiesen, daß eine Beschränkung auf $s_a < s$ latente Dimensionen eine Kleinste-Quadrate-Approximation der Datenmatrix K, bzw. P bzw. X darstellt (Eckart & Young (1936)).

Die Matrix X läßt sich dann aus F und G zurückrechnen: aus $F = D_r^{-1/2} U \Lambda^{1/2}$ folgt durch Multiplikation von rechts mit V'

$$FV' = D_r^{-1/2} U \Lambda^{1/2} V' = D_r^{-1/2} X,$$

woraus durch Multiplikation von rechts mit $D_r^{1/2}$

$$X = D_r^{1/2} F V'$$

folgt. Aus $G = D_c^{-1/2} V \Lambda^{1/2}$ folgt wiederum $V = D_c^{1/2} G \Lambda^{-1/2}$, so daß sich

$$X = D_r^{1/2} F \Lambda^{-1/2} G' D_c^{1/2} \tag{6.24}$$

ergibt.

Multipliziert man in der Gleichung für X von links mit $D_r^{1/2}$ und von rechts mit $D_c^{1/2}$, so erhält man

$$D_r^{1/2} X = D_c^{1/2} = P - E, \quad \text{oder} \quad P = D_r^{1/2} X D_c^{1/2} + E$$

Literatur

[1] Andersen, E. B.: Introduction to the statistical analysis of categorical data. Springer-Verlag, Berlin etc 1997

[2] Basilevsky, A.: Statistical factor analysis and related methods. Theory and applications. John Wiley & Sons, New York, 1994

[3] Burt, C. (1950) The factorial analysis of qualitative data. *British Journal of Psychology, (Statistical Section)* 3, 166–185

[4] Eckart, C. Young, G. (1939) A principal axis transformation for non-Hermitian matrices. *Am. Math. Society Bulletin*, 45, 118 – 121

[5] Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics*, 10, 422-429.

[6] Gabriel, K.R. (1971) The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. *Biometrika* 58(3), 453 – 467

[7] Golub, G.G., Reinsch, C.H. (1970) Singular value decomposition and least squares solution. *Numer. Math.*, 14, 403-20

[8] Greenacre, M.: Theory and Applications of Correspondence Analysis. London 1984

[9] Haberman, S.J. Analysis of qualitative data, Vol. I, *National Opinion Research Center*, 1972-1975

[10] Heuer, J.: Selbstmord bei Kindern und Jugendlichen. Ernst Klett Verlag, Stuttgart 1979

[11] Hirshfield, H. O. (1935) A connection between correlation and contingency. *Cambridge Philosophical Society Proceedings* 31, 520-524.

[12] Hofstätter, P.R.: Differentielle Psychologie. Stuttgart 1971

[13] Horst, P. (1935). Measuring complex attitudes. *Journal of Social Psychology*, 6, 369-374.

[14] Kendall, M.G., Stuart, A.: Theadvanced theory of statistics. Vol. 2: Inference and relationship. Griffin, London 1973

[15] Kretschmer, E.: Körperbau und Charakter. 23-24-te Auflage, Berlin 1961

[16] Lancaster, H.O. (1963) Canonical correlations and partitions of χ^2. *Quarterly Journal of Mathematics, Oxford*, 14 (2), 220

[17] Marascuilo, L.A., McSweeny, M.: Non-parametric and distribution-free methods for the social sciences. Monterey, Calif., Brooks/Cole 1977

[18] Maung, L. (1941) Measurement of association in a contingency table with special reference to the pgimentation of hair and eye colours of Scottish schoolchíldren. *Annals of Eugenics*, 11, 189–223

[19] Nishisato, S.: Analysis of categorical data: Dual Scaling and its applications. University of Toronto Press, Toronto 1980

[20] Richardson, M., Kuder, G. F. (1933). Making a rating scale that measures. *Personnel Journal*, 12, 36–40.

[21] Westphal, K. (1931) Körperbau und Charakter des Epileptikers. *Nervenarzt*, 4

[22] Tocher, J.F. (1908) Pigmentation Survey of School Children in Scotland. *Biometrika*, Vol. 6, No. 2/3 (Sep., 1908), pp. 129-235

Index