.

# Einführung in die Korrespondenzanalyse

**U. Mortensen**

.

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

Letzte Korrektur: 20. 10. 2011

# Inhaltsverzeichnis


**1 Einführung** **3**

**2 Definitionen** **5**

**3 Zeilen- und Spaltenprofile** **6**

3.1 $\chi^2$ und Trägheit . . . . . 7

3.2 $\chi^2$-Distanzen und $\chi^2$-Metrik . . . . . 8

**4 Skalenwerte** **10**

4.1 Bedingungen für die Skalenwerte . . . . . 11

4.2 Grundstruktur (SVD) und Skalenwerte . . . . . 12

4.2.1 Bestimmung der Skalenwerte . . . . . 12

4.2.2 Die Verallgemeinerte SVD . . . . . 15

4.2.3 Die Beziehung zwischen den Koordinaten . . . . . 16

4.2.4 Die Koordinaten als Eigenvektoren . . . . . 18

4.2.5 Die Rekonstitutionsformel . . . . . 18

**5 Die Diskussion einer Lösung** **18**

5.1 Die Zerlegung des $\chi^2$ . . . . . 19

5.2 Trägheitsanteil . . . . . 19

5.3 Relative Trägheit . . . . . 19

5.4 Qualität . . . . . 20

5.5 Der $\cos^2$-Anteil . . . . . 21

**6 Beziehungen zu anderen Verfahren und Anwendungen** **21**

6.1 Korrespondenzanalyse und Kanonische Korrelation . . . . . 21

6.2 Multiple Korrespondenzanalyse . . . . . 23

6.2.1 Spezialfall: die bivariate Indikatormatrix ($Q = 2$) . . . . . 23

6.2.2 Multivariate Indikatormatrizen und Burt-Matrizen . . . . . 26

**7 Beispiele** **27**

**8 Anhang: Beweise** **50**

8.1 Die Singularwertzerlegung (Grundstruktur) von $X$ . . . . . 50

8.2 Satz 1 . . . . . 51

8.3 Satz 2 . . . . . 52

8.4 Rekonstitution . . . . . 53

8.5 Zerlegung des $\chi^2$ . . . . . 53

# 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 $S_j$. Im allgemeinen wird es irgendwelche Abhängigkeiten zwischen den Zeilen- und den Spaltenkategorien geben, so daß die $n_{ij}$ in nicht nur zufälliger Weise von den bei Unabhängigkeit von Zeilen- und Spaltenkategorien zu erwartenden Häufigkeiten $n_{i+}n_{+j}/n_{++}$ abweichen. Gesucht ist ein Modell, das eine Deutung der Abhängigkeiten erlaubt. Ein allgemeiner Ansatz für die Konstruktion eines solchen Modells ergibt sich aus der Annahme, daß die Abhängigkeiten darauf zurückzuführen sind, daß Zeilen- und Spaltenkategorien durch bestimmte "latente" Variablen definiert sind. Das Ausmaß, in dem eine Kategorie durch eine latente Variable definiert ist, soll durch einen Skalenwert repräsentiert werden. Dementsprechened sollen sowohl die Zeilen- wie auch die Spaltenkategorien durch Skalenwerte auf ein und derselben Skala repräsentiert werden, und die Relationen zwischen diesen Skalenwerten soll die Struktur der Abhängigkeiten reflektieren.

Es sind von verschiedenen Autoren zum Teil unabhängig voneinander verschiedene Ansätze zur dualen Skalierung gemacht worden, wobei sich aber zeigt, daß sie im wesentlichen stets auf den gleichen Kern hinauslaufen. Einen historischen Überblick findet man in Nishisato (1980). Hier soll nur ein sehr kurzer Überblick gegeben werden, weil es einerseits das Verständnis der Verfahren vertieft, wenn man sieht, wie aus zunächst verschieden formulierten Annahmen das gleiche Verfahren resultiert, und andererseits deutlich wird, daß die unter verschiedenen Namen bekannten Verfahren auf den gleichen Kerngedanken zielen.

1. *Method of Reciprocal Averages:* Dieser Name wurde von Horst (1935) bei der Beschreibung einer Arbeit von Richardson und Kuder (1933) vorgeschlagen, deren Ziel es war, Skalenwerte für Personen einerseits und Items von Fragebögen oder Tests andererseits so zu finden, daß die Variation der Skalenwerte *innerhalb* einer Gruppe oder einer Person so klein wie möglich und die Variation *zwischen* den Gruppen/Personen so groß wie möglich sein sollte, um auf diese Weise möglichst gut zwischen Personen diskriminieren zu können. Es wird vorgeschlagen, den Skalenwert eines Items durch den durschschnittlichen Skalenwert der Personen, die positiv auf das Item reagieren (oder die die Aufgabe lösen, wenn das Item eine Aufgabe ist), zu definieren. Umgekehrt sollte der Skalenwert einer Person als durch den durchschnittlichen Skalenwert der Items definiert sein, die sie positiv beantwortet[1] Dieses Vorgehen erklärt den Ausdruck *Reciprocal Averages*.

2. *Simultane lineare Regression:* Hirschfeld (1935) machte den Ansatz, die Häufigkeiten einer Kontingenztabelle in bezug auf eine 2-dimensionale Verteilung zu interpretieren. Die Skalenwerte der Zeilen- und Spaltenkategorien sollen sich dann durch lineare Regressionen auf diese Dimensionen ergeben.

3. *Diskriminanzanalyse:* Fisher (1940) diskutierte die Verteilung von Haar- und Augenfarben in Caithness (Schottland). Sei Ziel war, den Augenfarben Skalenwerte zuzuordnen derart, daß die Skalenwerte für die Haarfarben so verschieden wie möglich wurden, d.h. man will anhand beobachtbarer Merkmale zwischen Personentypen so gut es geht diskriminieren (daß man versucht, die Haarfarbe "vorherzusagen" scheint trivial zu sein, aber die *beobachtete* Haarfarbe erlaubt eine Überprüfung der Diskriminierung). Fishers Ansatz erweist sich als Spezialfall des Ansatz von Richardson und Kuder (1933) und von Hirschfeld (1935).

[1] "Positiv" kann natürlich durch "Negativ" ersetzt werden. Es kommt nur darauf an, eine bestimmte Reaktionsart festzulegen.

4. *Kanonische Korrelation:* Maung (1941) scheint zuerst auf die Idee gekommen zu sein, daß das Problem, Skalenwerte für die Zeilen- und Spaltenkategorien zu finden, über den Ansatz der Kanonischen Korrelation zu lösen. Der gleiche Ansatz findet sich in Kendall und Stuart (1973), p. 588, die wiederum auf eine Arbeit von Lancaster (1963) verweisen. Der Ansatz der Kanonischen Korrelation soll im Folgenden zuerst besprochen werden, weil er am ehesten die Logik der Skalenzuordnung zu verdeutlichten scheint.

5. *Korrespondenzanalyse:* Hier werden die normierten Differenzen

$$x_{ij} := (n_{ij} - n_{i+}n_{+j})/\sqrt{n_{i+}n_{+j}}$$

eine Hauptachsentransformation unterzogen. Jede Datenmatrix $X$ kann ja in der Form $X = Q\Lambda^{1/2}P'$ dargestellt werden; enthalten die Zeilen von $X$ die Scores von Personen in Tests, die wiederum die Spalten von $X$ definieren, so enthalten die Zeilen von $Q$ die Scores (= Skalenwerte) der Personen und die Spalten von $\Lambda^{1/2}$ die Scores (= Skalenwerte) der Tests auf latenten Dimensionen. Auf die Differenzen $x_{ij}$ angewandt sind diese Scores Skalenwerte für die Zeilen- und Spaltenwerte der Kontingenztabelle. Auch hier besteht eine Beziehung zur Kanonischen Korrelation.

Die Differenz $n_{ij} - n_{i+}n_{+j}$ wird gelegentlich als "chance correction", also als eine "Bereinigung" der $n_{ij}$ von zufälligen Effekten interpretiert derart, dass die $x_{ij}$ nur noch systematische, aber keine zufälligen Effekte mehr enthalten. Diese Interpretation kann in dieser Allgemeinheit sicherlich nicht gelten; in Abschnitt 4 wird dieser Sachverhalt noch einmal aufgegriffen.

Weitere Ansätze findet man in Nishisato (1980).

Im Folgenden wird auf die Korrespondenzanalyse fokussiert. Die Vorgehensweise kann analog zu der bei der Faktorenanalyse bzw. bei der Approximation der Faktorenanalyse durch die Hauptachsentransformation konstruiert werden. Gegeben sei eine $m \times n$-Matrix $X$ von Meßwerten $x_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. $x_{ij}$ ist etwa der Meßwert der $i$-ten Person im $j$-ten Test. Es wird nun angenommen, daß es $r$ "latente" Variablen $L_1, \ldots, L_r$ gibt derart, daß

$$x_{ij} = q_{i1}f_{j1} + \cdots + q_{ir}f_{jr} \tag{1}$$

gilt. Dabei sind die $q_{i1}, \ldots, q_{ir}$ Maße, die die $i$-te Person auf den $L_1, \ldots, L_r$ charakterisieren ("Faktorscores"), und die $f_{j1}, \ldots, f_{jr}$ sind Maße, die die Tests auf den latenten Dimensionen haben ("Faktorladungen"). Die Produkte $q_{ik}f_{jk}$, $k = 1, \ldots, r$ sind additive Komponenten von $x_{ij}$; eine solche Komponente ist einerseits proportional zu $q_{ik}$, d.h. zum Maß, das die $i$-te Zeile (Person) auf der $k$-ten latenten Variablen charakterisiert, und andererseits proportional zum Maß $f_{jk}$, daß die $j$-te Spalte (Test) auf der gleichen latenten Variablen hat. Die $q_{ik}$, $k = 1, \ldots, r$ können als Skalenwerte für die Zeilen der Matrix $X$, und die $f_{jk}$ als Skalenwerte der Spalten aufgefaßt werden. Die $i$-te Zeile (Person) wird dann durch den Vektor $q_i = (q_{i1}, \ldots, q_{ir})'$ repräsentiert, und die $j$-te Spalte (Test) durch den Vektor $f_j = (f_{j1}, \ldots, f_{jr})'$. Damit werden die Zeilen durch Punkte (die Endpunkte der Vektoren $q_i$) in einem $r$-dimensionalen Raum repräsentiert, ebenso werden die Spalten durch die Endpunkte der Vektoren $f_j$ im gleichen Raum abgebildet. Die latenten Variablen werden so bestimmt, daß durch sie eine additive Zerlegung der Gesamtvarianz der Daten in $X$ möglich wird; jede latente Variable "erklärt" einen Anteil dieser Gesamtvarianz. Darüber hinaus werden die Korrelationen, d.h. die Abhängigkeiten, zwischen den Spaltenvariablen ("Tests") oder aber zwischen den Zeilenvariablen ("Personen") durch die latenten Variablen erklärt.

Bei der Analyse der Kontingenztabelle sollen ebenfalls die Abhängigkeiten zwischen Zeilen- und Spaltenkategorien durch latente Variablen erklärt werden. Ein generelles Maß

für die Abhängigkeiten in der Tabelle ist durch das $\chi^2$ gegeben. Wie bei der Hauptachsentransformation einer Matrix $X$ von Meßwwerten $x_{ij}$ für jede latente Variable angegeben werden kann, welchen Anteil der Gesamtvarianz sie erklärt, soll bei der Diskussion einer Kontingenztabelle angegeben werden, welchen Anteil des $\chi^2$ eine gegebene latente Variable erklärt.

Bei der Hauptachsentransformation einer Matrix $X$ von Meßwerten sind die $q_{ik}$ und die $f_{jk}$ Skalenwerte für die gleiche latente Variable $L_k$. Die Punkte, die die Zeilen von $X$ repräsentieren, und die Punkte, die die Spalten repräsentieren, können also im gleichen Achsensystem dargestellt werden: diese simultane Darstellung heißt Biplot. Die Distanzen zwischen den Zeilenpunkten einerseits und den Spaltenpunkten andererseits sind für eine Interpretation der Achsen hilfreich: Cluster von Zeilenpunkten, d.h. Teilmengen von Punkten mit (relativ) kleiner Distanz zwischen ihnen repräsentieren Zeilen (oder Spalten) mit ähnlicher Ausstattung hinsichtlich der latenten Variablen. Bei der Analyse von Kontingenztabellen spielt der Biplot eine analoge Rolle. Man muß allerdings darauf achten, daß nur die Distanzen zwischen den Zeilenpunkten einerseits und den Spaltenpunkten andererseits sinnvoll zu deuten sind.

Die Distanz zwischen einem Zeilen- und einem Spaltenpunkt ist *nicht erklärt*. Dies ist einleuchtend: Eine Zeilenkategorie wird durch die Verteilung der Häufigkeiten in dieser Zeile erklärt, und eine Spaltenkategorie durch die Verteilung der Häufigkeiten in dieser Spalte; man kann sagen, daß eine Zeile durch die Menge der Spalten, eine Spalte durch die Menge der Zeilen charakterisiert wird. Dies macht intuitiv klar, daß die Nähe des Punktes, der eine Zeile repräsentiert, zu einem Punkt, der eine Spalte abbildet, nicht unmittelbar zu deuten ist. Allerdings ist das Skalarprodukt zwischen den Vektoren, die eine Zeilen- und eine Spaltenkategorie in bezug auf die latenten Variablen definieren, erklärt: es entspricht dem Meßwert $x_{ij}$ bei einer Matrix von Meßwerten, und korrespondiert zu den Häufigkeiten $n_{ij}$ bzw. zu den Residuen $n_{ij} - n_{i\cdot}n_{\cdot j}$ in einer Kontingenztabelle ( $n_{i\cdot}n_{\cdot j}$ sind die zu den beobachteten Häufigkeiten $n_{ij}$ korrespondierenden unter der Hypothese, daß keine Abhängigkeiten zwischen Zeilen- und Spaltenkategorien bestehen, erwarteten Häufigkeiten).

Im folgenden Abschnitt werden die für die Korrespondenzanalyse zentralen Begriffe eingeführt. Beweise der Aussagen über die Beziehungen zwischen den Begriffen werden im Anhang 8 gegeben und müssen nur gelesen werden, wenn ein tieferes Verständnis der Details gewünscht wird.

# 2 Definitionen

Es sei wieder $K = (n_{ij})$, $i = 1, \cdots, I$, $j = 1, \cdots, J$ eine Kontingenztabelle, d.h. es gebe $I$ Zeilenkategorien $R_i$, und $J$ Spaltenkategorien $S_j$. Eine Zeilenkategorie ist zunächst durch die Verteilung der Häufigkeiten in dieser Zeile definiert, also durch den Vektor $(n_{i1}, \ldots, n_{iJ})'$. Dementsprechend ist der Unterschied zwischen zwei Zeilenkategorien $R_i$ und $R_{i'}$ durch den Unterschied der Vektoren $(n_{i1}, \ldots, n_{iJ})'$ und $(n_{i'1}, \ldots, n_{i'J})'$ charakterisiert.

Existieren keinerlei Abhängigkeiten zwischen Zeilen- und Spaltenkategorien, so weichen die beobachteten Häufigkeiten $n_{ij}$ nur zufällig von den erwarteten Häufigkeiten $n_{i\cdot}n_{\cdot j}/N$ ab. Betrachtet man die Zeilenprofile der Matrix der erwarteten Häufigkeiten, so findet man, daß sie alle parallel zueinander verlaufen (sie sind sogar identisch). Eine analoge Aussage gilt für die Spaltenprofile dieser Matrix. Bei nur zufälligen Abweichungen von den erwarteten Häufigkeiten sind die Zeilen- bzw. Spaltenprofile in guter Näherung parallel zueinander; Abhängigkeiten äußern sich dementsprechend in Abweichungen von

der Parallelität, insbesondere in Gegenläufigkeiten.

Es sei $N = \sum_i \sum_j n_{ij}$ und $p_{ij} = n_{ij}/N$ sei die relative Häufigkeit; die Matrix der relativen Häufigkeiten erhält man aus der Kontingenztabelle $K$ gemäß

$$P = \frac{1}{N}K = (p_{ij}), \quad 1 \le i \le I, \quad 1 \le j \le J \tag{2}$$

Also ist $P = (p_{ij}) = (n_{ij}/N)$ die Matrix der relativen Häufigkeiten:

Tabelle 1: Die Matrix $P$ mit Zeilen- $(R_i)$ und Spaltenkategorien $(S_j)$

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

# 3 Zeilen- und Spaltenprofile

Es sei

$$r_i = \sum_{j=1}^{J} p_{ij} = \frac{1}{N}\sum_{j=1}^{J} n_{ij} = \frac{n_{i.}}{N} \tag{3}$$

$\vec{r} = (r_1, \cdots, r_I)'$ ist der Vektor der Zeilensummen von $P$. Setzt man, wie üblich, $\sum_j n_{ij} = n_{i.}$, so ist offenbar $r_i = n_{i.}/N$. Ebenso sei

$$c_j = \sum_{i=1}^{I} p_{ij} = \frac{1}{N}\sum_{i=1}^{I} n_{ij} = \frac{n_{.j}}{N}, \quad j = 1, \ldots, n \tag{4}$$

$\vec{c} = (c_1, \cdots, c_J)'$ ist der Vektor der Spaltensummen von $P$. Die Beziehung zwischen den $p_{ij}$, $r_i$ und $c_j$ wird in der Tabelle 2 noch einmal veranschaulicht.

**Definition 1** *Die Zeilensummen $r_i$ (Spaltensummen $c_j$) von $P$ heißen* Massen *der Zeilenkategorien (Spaltenkategorien).*

Aus der Definition der $r_i$ und $c_j$ ergibt sich sofort

$$\sum_{i=1}^{I} r_i = \sum_{j=1}^{J} c_j = 1. \tag{5}$$

Die Einführung des Begriffs "Masse" scheint zunächst ein wenig überflüssig zu sein, da man ja einfach von den Zeilen- bzw. Spaltensummen reden kann. In der Begriffswelt der Korrespondenzanalyse ist der Ausdruck aber üblich, zumal er in übereinstimmung mit dem Sprachgebrauch der Physik den Gebrauch anderer Ausdrücke wie z.B. Baryzentrum für den Schwerpunkt einer Punktekonfiguration steht.

**Definition 2** *Der Vektor*

$$\left(\frac{p_{i1}}{r_i}, \ldots, \frac{p_{iJ}}{r_i}\right)'$$

*heißt* $i$-tes Zeilenprofil *der Kontingenztabelle* $K$*. Der Vektor*

$$\left(\frac{p_{1j}}{c_j}, \cdots, \frac{p_{Ij}}{c_j}\right)'$$

*heißt* $j$-tes Spaltenprofil *von* $K$*. Der Vektor* $(r_1, \cdots, r_I)'$ *heißt* mittleres Spaltenprofil*, und* $(c_1, \cdots, c_J)'$ *heißt* mittleres Zeilenprofil *von* $K$*.*

**Anmerkung:** Es ist

$$\frac{p_{ij}}{r_i} = \frac{n_{ij}}{N}\frac{N}{n_{i\cdot}} = \frac{n_{ij}}{n_{i\cdot}} \tag{6}$$

Dementsprechend ist das $i$-te Zeilenprofil auch durch

$$\left(\frac{n_{i1}}{n_{i\cdot}}, \ldots, \frac{n_{iJ}}{n_{i\cdot}}\right)' \tag{7}$$

definiert. Eine analoge Ausssage gilt für die Spaltenprofile.

Ein Zeilenprofil ist also einfach die Verteilung der Häufigkeiten in einer Zeile, relativiert durch die Zeilensumme $n_{i\cdot}$; der Effekt unterschiedlicher Häufigkeiten in den Zeilenkategorien $R_i$ wird also herausgenommen. Damit enthält das Zeilenprofil die *bedingten Häufigkeiten* der Spaltenkategorien, d.h. ein Element eines Zeilenprofils entspricht der bedingten Wahrscheinlichkeit einer bestimmten Spaltenkategorie, gegeben eine bestimmte Zeilenkategorie. Kennt man also die Zeilenkategorie, so gibt die Komponente an, mit welcher (geschätzten) Wahrscheinlichkeit nun eine betimmte Spaltenkategorie zu erwarten ist. Eine analoge Aussage gilt wieder für die Spaltenprofile; ein Spaltenprofil gibt die (Schätzung der) bedingten Wahrscheinlichkeit einer Zeilenkategorie, wenn eine Spaltenkategorie gegeben ist.

## 3.1 $\chi^2$ und Trägheit

Das $\chi^2$ der Tabelle $K$ ist

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/N)^2}{n_{i\cdot}n_{\cdot j}/N} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(N\,p_{ij} - N r_i c_j)^2}{N\,r_i c_j} \\
&= N\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j}
\end{aligned} \tag{8}$$

**Definition 3** *Die Größe*

$$\frac{\chi^2}{N} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{9}$$

*heißt* Gesamt-Inertia *oder* Gesamtträgheit *der Tabelle $K$. Die Teilsumme*

$$\frac{\chi^2_{i\cdot}}{N} = \sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{10}$$

*heißt* $i$-te Zeilen-Inertia*; die Teilsumme*

$$\frac{\chi_{\cdot j}}{N} = \sum_{i=1}^{I} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{11}$$

*heißt* $j$-te Spalten-Inertia.

**Anmerkungen:**

1. Die Inertia $\chi^2/N$ ist offenbar gleich dem Quadrat des Kontigenzkoeffizienten
$$C = \sqrt{\frac{\chi^2}{N}}. \tag{12}$$
2. Der Ausdruck "Inertia" = Trägheit ergibt sich aufgrund formaler Analogien des $\chi^2$-Ausdrucks zum physikalischen Trägheitsbegegriff; eine Diskussion des Trägheitsbegriffs trägt allerdings zum Verständnis der Korrespondenzanalyse nicht weiter bei und wird deshalb hier übergangen.
3. Die Gesamt-Inertia ist offenbar gleich dem Quadrat des Pearsonschen Kontingenzkoeffizienten.
4. Die Summe der Zeilen-Inertiae ist gleich der Summe der Spalten-Inertiae; die Summen sind gleich der Gesamt-Inertia.

□

## 3.2 $\chi^2$-Distanzen und $\chi^2$-Metrik

Je ähnlicher sich nun die Profile sind, desto näher sollten die die Zeilenkategorien repräsentierenden Punkte in dem Koordinatensystem, das die latenten Variablen abbildet, liegen. Der Abstand zwischen diesen Punkten wird als euklidische Distanz berechnet. Sind die Koordinaten dieser Punkte durch $f_{i1}, \ldots, f_{ir}$ bzw. $f_{i'1}, \ldots, f_{i'r}$ gegeben, so ist demnach die Distanz zwischen den Punkten $i$ und $i'$ - die also die $i$-te bzw. die $i'$-te Zeile repräsentieren - durch

$$d(i, i') = \sqrt{\sum_{k=1}^{r} (f_{ik} - f_{i'k})^2} \tag{13}$$

gegeben.

> *Die Koordinaten $f_{ik}$, $f_{i'k}$ sollen so bestimmt werden, daß die gewünschte additive Zerlegung des $\chi^2$-Wertes erreicht wird.*

Dem Abstand, d.h. der Distanz $d(i, i')$ muß nun ein Maß für den Abstand der Verteilungen der Häufigkeiten für die Zeilenkategorien entsprechen. Es zeigt sich, daß das in der folgenden Definition eingeführte Distanzmaß die gewünschten Eigenschaften impliziert:

**Definition 4** *Es werden die $i$-te und die $i'$-te Zeilenkategorie betrachtet. Die durch*

$$\delta^2_{ii'} = \sum_{j=1}^{J} \frac{1}{c_j} \left( \frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2 = \sum_{j=1}^{J} \frac{1}{n_{\cdot j}} \left( \frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}} \right)^2 \tag{14}$$

*definierte Größe $\delta^2_{ii'}$ heißt* $\chi^2$-Distanz *zwischen den Zeilenkategorien $R_i$ und $R_{i'}$. Durch*

$$\delta^2_{jj'} = \sum_{i=1}^{I} \frac{1}{r_i} \left( \frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}} \right)^2 = \sum_{i=1}^{I} \frac{1}{n_{i\cdot}} \left( \frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}} \right)^2 \tag{15}$$

*heißt* $\chi^2$-Distanz *für die Spaltenkategorien $S_j$ und $S_{j'}$.*

**Anmerkungen:**

1. Der Ausdruck $\chi^2$-Distanz weist darauf hin, daß der hier charakterisierte Distanzbegriff in Hinblick auf den $\chi^2$-Wert der Tabelle eingeführt worden ist.

2. Es sei $d(a, b)$ eine Distanz zwischen irgendzwei Punkten $a$ und $b$. Das Distanzmaß $d$ definiert eine *Metrik*, wenn $d$ die Bedingungen

   (i) $d(a, b) \geq 0$,

   (ii) $d(a, b) = d(b, a)$ (Reflexivität),

   (iii) $d(a, c) \leq d(a, b) + d(b, c)$, $c$ ein weiterer Punkt (Dreiecksungleichung)

   genügt. Das Distanzmaß $\delta_{ii''}$, bzw. $\delta_{jj'}$ genügt diesen Bedingungen und definiert damit eine Metrik, hier die sogenannte $\chi^2$-*Metrik*.

Es ist nützlich, sich die Konstruktion der Definition einer $\chi^2$-Distanz klar zu machen. Für die Zeile $R_i$ läßt sich das Profil $n_{i1}/n_{i\cdot}, \ldots, n_{iJ}/n_{i\cdot}$ bzw. $p_{ij}/r_i, \ldots, p_{ij}/r_i$ anschreiben. Die $p_{ij}/r_i$, $j = 1, \ldots, J$ lassen sich als Komponenten eines Vektors auffassen. Die Endpunkte dieser Vektoren für $R_i$ und $R_{i'}$ sind durch die euklidische Distanz

$$d_{ii'} = \sqrt{\sum_j \left( \frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2}$$

voneinander getrennt. Dieses Distanzmaß liefert aber noch nicht die gewünschte Zerlegung des $\chi^2$-Wertes für die Tabelle. Eine solche Zerlegung erreicht man, wenn man die Quadrate der Differenzen $(p_{ij}/r_i - p_{i'j}/r_{i'})^2$ mit dem Faktor $1/c_j$ gewichtet; diese Gewichtung führt zu dem in Definition 4 gegebenen Ausdruck $\delta_{ii'}$, vergl. (14). Die $\chi^2$-Distanzen $\delta$ sind im übrigen euklidische Distanzen, wenn man die Komponenten der Zeilen- bzw. Spaltenprofile gewichtet; für die Zeilen betrachtet man also die Profile

$$\frac{p_{i1}}{r_i\sqrt{c_1}}, \ldots, \frac{p_{iJ}}{r_i\sqrt{c_j}}, \quad i = 1, \ldots, I$$

Die Betrachtung für die Spalten-$\chi^2$-Distanzen ist analog.

Es soll nun die Beziehung der $\chi^2$-Distanzen zum $\chi^2$ bzw. zur Gesamt-Inertia der Tabelle aufgezeigt werden. Dazu wird zunächst die $\chi^2$-Distanz zum mittleren Zeilenprofil $c_1, \ldots, c_J$ angeschrieben:

$$\delta^2_{i\cdot} = \sum_{j=1}^{J} \frac{1}{c_j} \left( \frac{p_{ij}}{r_i} - c_j \right)^2 \tag{16}$$

Dieser Ausdruck ergibt sich, wenn man die Komponenten $p_{i'j}/r_{i'}$ in (14) durch $c_j$ ersetzt (vergl. Definition 2, Seite 7). Dann läßt sich die folgende Aussage herleiten (der Beweis wird im Anhang, Abschnitt 8.2 gegeben):

**Satz 1**

$$\chi^2/N = \sum_{i=1}^{I} r_i \sum_{j=1}^{J} \frac{1}{c_j} \left( \frac{p_{ij}}{r_i} - c_j \right)^2 = \sum_{i=1}^{I} r_i \delta_{i\cdot}^2 \tag{17}$$

*bzw.*

$$\chi^2 = N \sum_{i=1}^{I} r_i \delta_{i\cdot}^2. \tag{18}$$

*Der Wert des $\chi^2$ der Tabelle K ist also gleich der gewogenen Summe der Quadrate der $\chi^2$-Distanzen zwischen den Zeilenprofilen und dem mittleren Zeilenprofil; die Gewichte sind die Zeilensummen $r_i$. Gleichzeitig wird damit gezeigt, daß für die Zeilen-$\chi^2$ $\chi_i^2$, $i = 1, \ldots, I$ die Aussage*

$$\chi_i^2 = r_i \delta_{i\cdot}^2 \tag{19}$$

*gilt. Für die Spalten gilt eine analoge Aussage.*

> *Weicht also ein Zeilenprofil, etwa das i-te, nicht vom mittleren Profil ab, so gilt $\chi_i^2 = 0$ und damit $\delta_{i\cdot} = 0$, d.h. der Punkt für $R_i$ fällt mit dem Nullpunkt des Koordinatensystems zusammen. Fällt das Profil von $R_i$ nicht mit dem mittleren Profil zusammen, so ist $\delta_{i\cdot}$ proportional zum entsprechenden Zeilen-$\chi^2$ $\chi_i^2$, d.h. $\delta_{i\cdot} = \chi_i^2/r_i$.*

# 4 Skalenwerte

Es sei

$$x_{ij} \stackrel{def}{=} \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} = \frac{1}{\sqrt{N}} \frac{n_{ij} - n_{i\cdot} n_{\cdot j}/N}{\sqrt{n_{i\cdot} n_{\cdot j}/N}} \tag{20}$$

gegeben. Aus (9) folgt dann, daß

$$\sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}^2 = \frac{\chi^2}{N} \tag{21}$$

ist. Die Differenzen $p_{ij} - r_i c_j$ bzw. $n_{ij} - n_{i\cdot} n_{\cdot j}/N$ können als *Residuen* aufgefaßt werden; einer intuitiven Interpretation nach repräsentieren sie den "Rest" an Häufigkeit, der übrig bleibt, wenn man die Häufigkeit zufälligen Zusammentreffens von $R_i$ und $S_j$ aus den Daten entfernt. Eine solche Interpretation muß aber mit Vorsicht akzeptiert werdennn aber durchaus irreführend sein. Denn wenn es systematisch wirkende Abhängigkeitsstrukturen gibt derart, dass $n_{ij} \neq n_{i+} n_{+j}/N$ und diese Ungleichheit nicht nur zufällig ist, so sind die Abhängigkeitsstrukturen auch in den $n_{i+}$ und $n_{+j}$ enthalten und die Differenz $n_{ij} - n_{i+} n_{+j}/N$ ist nicht notwendig von "rein zufälligen" Komponenten "bereinigt". Im Skriptum *Einführung in die Theorie psychometrischer Tests* wird dieser

Sachverhalt etwas ausführlicher am Beispiel der Maße für Urteilerübereinstimmung (z.B. Cohens Kappa, Cohen (1960)) illustriert.

Ob ein Residuum "groß" oder "klein" ist, hängt natürlich von den Zeilen- und Spaltensummen für die entsprechenden Kategorien ab. Die Division durch $\sqrt{r_i c_j}$ bewirkt, daß ein gegebenes Residuum eine *kleine* Auswirkung auf den Wert des $\chi^2$ hat, wenn die Zeilen- bzw. Spaltensumme $r_i$ oder $c_j$ einen *großen* Wert hat. Umgekehrt hat das Residuum eine *große* Auswirkung, wenn die entsprechenden Zeilen- und Spaltensummen einen *kleinen* Wert haben; die Abweichung von $p_{ij}$ von $r_i c_j$ hat dann ja gewissermaßen mehr zu bedeuten; die Division von $p_{ij} - r_i c_j$ durch $\sqrt{r_i c_j}$ entspricht also einer Standardisierung.

Es ist für die folgenden Betrachtungen günstig, die Matrix $X = (x_{ij})$ in Matrixform zu repräsentieren. Es ist $rc' = (r_i c_j)$, und

$$P - rc' = (p_{ij} - r_i c_j), \quad i = 1, \ldots, I; j = 1, \ldots, J.$$

Weiter sei

$$D_r = \begin{pmatrix} r_1 & 0 & 0 & \cdots & 0 \\ 0 & r_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & r_I \end{pmatrix}, \quad D_c = \begin{pmatrix} c_1 & 0 & 0 & \cdots & 0 \\ 0 & c_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & c_J \end{pmatrix}, \tag{22}$$

d.h. $D_r$ sei die Diagonalmatrix der Zeilensummen, $D_c$ sei die Diagonalmatrix der Spaltensummen von $P$, und die unter der Hypothese der Unabhängigkeit von Zeilen- und Spaltenkategorien erwarteten Werte (dividiert durch $N$) seien in $E = (r_i c_j) = rc'$ zusammengefaßt. Dann ist die Matrix $X$ der $x_{ij}$-Werte durch

$$X = D_r^{-1/2}(P - E)D_c^{-1/2} = (x_{ij}), \tag{23}$$

gegeben, wie man durch Nachrechnen verifiziert. Die zu bestimmenden Skalenwerte sollen eine Reihe von Bedingungen erfüllen; diese Bedingungen werden zunächst formuliert:

## 4.1 Bedingungen für die Skalenwerte

Es sollen Skalenwerte $f_{i1} \ldots, f_{ir}$, $i = 1, \ldots, I$ für die Zeilenkategorien und $g_{j1}, \ldots, g_{jr}$, $j = 1, \ldots, J$ für die Spaltenkategorien bezüglich der $r$ latenten Variablen $L_1, \ldots, L_r$ so bestimmt werden, daß die folgenden Bedingungen erfüllt sind:

1. *Den euklidischen Distanzen*

$$d_{ii'} = \sqrt{\sum_{k=1}^{r} (f_{ik} - f_{i'k})^2}, \quad d_{jj'} = \sqrt{\sum_{k=1}^{r} (g_{jk} - g_{j'k})^2}, \tag{24}$$

*wobei $d_{ii'}$ die Distanz zwischen der $i$-ten und der $i'$-ten Zeilenkategorie und $d_{jj'}$ die Distanz zwischen der $j$-ten und der $j'$-ten Spaltenkategorie ist, sollen die $\chi^2$-Distanzen $\delta_{ii'}$ bzw. $\delta_{jj'}$ entsprechen, und*

2. *den latenten Variablen $L_1, \ldots, L_r$ sollen $\chi^2$-Komponenten $\chi^2(L_1), \ldots, \chi^2(L_r)$ entsprechen derart, daß $\chi^2 = \sum_k \chi^2(L_k)$ gilt.*

## 4.2 Grundstruktur (SVD) und Skalenwerte

### 4.2.1 Bestimmung der Skalenwerte

Es sei noch einmal an die Definition der $\chi^2$-Statistik erinnert: Es ist

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{n_{ij}-\hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}\right)^2, \quad \hat{n}_{ij} = n_{i+}n_{+j}/N.$$

Mit

$$x_{ij} = \frac{n_{ij}-\hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}$$

hat man ein Maß für die Beziehung zwischen der $i$-ten Zeilen- und der $j$-ten Spaltenkategorie. Für $x_{ij} = 0$ bedeutet $n_{ij} = \hat{n}_{ij}$, dh die Konjunktion der Kategorien $R_i$ und $C_j$ ist rein zufällig, dh es gibt keine systematische Wirkung latenter Variablen. $x_{ij} \neq 0$ legt die Existenz von Abhängigkeiten nahe, sofern die Abweichung von Null nicht als zufällig bewertet werden kann. Die Frage nach Skalenwerten ist die Frage nach latenten Variablen, über die die Abhängigkeiten in der Tabelle erklärt werden können.

**Ansatz:** Die Definition von $x_{ij}$ weist eine gewisse Ähnlichkeit zur Definition eines standardisierten Messwerts auf: Ist etwa $u_{ij}$ der Messwert für die $j$-te Variable beim $i$-ten Objekt und ist $\bar{u}_j = \sum_i u_{ij}/m$ der Mittelwert der $u_{ij}$ und $s_j = (\sum_i (u_{ij}-\bar{u}_j)^2/m)^{1/2}$ die entsprechende Standardabweichung, so ist $z_{ij} = (u_{ij}-\bar{u}_j)/s_j$ der standardisierte Messwert. Beim PCA-Ansatz wird

$$z_{ij} = p_{j1}\vec{L}_1 + \cdots + p_{jr}\vec{L}_r \tag{25}$$

angenommen, wobei die $\vec{L}_k$ latente und paarweise unabhängige Variable repräsentieren. Diese latenten Variablen erklären im Wesentlichen die Abweichungen $u_{ij} - \bar{u}_j$. Dieser Ansatz kann auf die Erklärung der $x_{ij}$ übertragen werden. Zwar ist $n_{ij}$ kein Messwert im üblichen Sinne, und die unter $H_0$ erwarteten Häufigkeiten $\hat{n}_{ij}$ kann man nicht notwendig als Mittelwert betrachten, ebensowenig wie die $\sqrt{\hat{n}_{ij}}$ eine Standardabweichung sind, aber die Differenz $n_{ij} - \hat{n}_{ij}$ drückt möglicherweise den Effekt systematischer Abhängigkeiten zwischen einer Zeilen- und einer Spaltenkategorie aus, und diesen Effekt gilt es aufzuklären.

Man betrachte die Spalten $\vec{X}_j$ von $X = (x_{ij})$. Sie stehen für die Spaltenkategorien, dh für bestimmte Merkmale, etwa Typen von psychischen Erkrankungen. Die Annahme ist, dass diese Merkmale sich additiv aus bestimmten latenten Merkmalen zusammensetzen. Analog dazu repräsentieren die Zeilenkategorien eine andere Klasse von Merkmalen, etwa Körperbautypen. Die Zeilen von $X$ werden durch Vektoren $\vec{Y}_i$ repräsentiert, die vermutlich ebenfalls als Kombination bestimmter latenter Merkmale aufgefasst werden können.

Analog zur PCA kann nun der Ansatz

$$\vec{X}_j = p_{j1}\vec{U}_1 + \cdots + p_{jr}\vec{U}_r, \tag{26}$$

$$\vec{Y}_i = q_{i1}\vec{V}_1 + \cdots + q_{ir}\vec{V}_s, \tag{27}$$

gemacht werden, wobei die $\vec{U}_k$ ebenso wie die $\vec{V}_k$ paarweise orthogonal sind. Die $\vec{U}_k$ repräsentieren die latenten Merkmale, mit denen die Spaltenkategorien "erklärt" werden können, und die $\vec{V}_k$ die latenten Merkmale, über die die Zeilenkategorien erklärt werden können. Sollten irgendwelche Abhängigkeiten zwischen den Zeilen- und Spaltenkategorien existieren, so liegt die Vermutung nahe, dass diese Abhängigkeiten auf Beziehungen

zwischen den latenten Variablen für die Zeilenkategorien einerseits und denen für die Spaltenkategorien andererseits zurückzuführen sind.

Um diese Beziehungen zu elaborieren ist es nützlich, die Gleichungen (26) und (27) in Matrixform anzuschreiben:

$$X = UP', \quad Y = VQ', \tag{28}$$

wobei die $\vec{U}_k \perp \vec{U}_{k'}$ und $\vec{V}_k \perp \vec{V}_{k'}$ für $k \neq k'$. Die Orthogonalität der Vektoren impliziert, dass $U$ und $V$ orthogonale Matrizen sind, d.h. es müssen die Gleichungen

$$U'U = \Lambda_u, \quad V'V = \Lambda_v \tag{29}$$

gelten, wobei $\Lambda_u$ und $\Lambda_v$ Diagonalmatrizen sind. Es sind

$$U_0 = U\Lambda_u^{-1/2}, \quad V_0 = V\Lambda_v^{-1/2} \tag{30}$$

die auf die Länge 1 normierten Spaltenvektoren von $U$ bzw. $V$.

Der folgende Satz macht die Beziehung zwischen den Repräsentation (26) und (27) der Zeilen- und Spalten von $X$ explizit: die Vektoren $\vec{U}_k$ und die $\vec{V}_k$ repräsentieren dieselben latenten Variablen:

**Satz 2** *Es mögen die Beziehungen in der Gleichung (28) gelten. Dann folgt*

$$P = V_0, \quad Q = U_0 \tag{31}$$

*sowie*

$$X = U_0\Lambda^{1/2}V_0', \tag{32}$$

*wobei $U_0$ und $V_0$ wie in (30) definiert sind, und*

$$\Lambda = \Lambda_u = \Lambda_v. \tag{33}$$

**Beweis:** Wegen der postulierten Orthogonalität von $U$ folgt, dass $U'U = \Lambda_u$ eine Diagonalmatrix ist. Ebenso folgt, dass $V'V = \Lambda_v$ eine Diagonalmatrix ist. Dann hat man $X'X = P\Lambda_u P'$, so dass $P$ die orthonormale Matrix der Eigenvektoren von $X'X$ sein muß, und $\Lambda_u$ enthält die zugehörigen Eigenwerte[2]. Wegen (28) und (30) hat man

$$X = U_0\Lambda_u^{1/2}P', \quad Y = V_0\Lambda_v^{1/2}Q'. \tag{34}$$

Es ist aber $Y = X'$, so dass

$$V_0\Lambda_v^{1/2}Q' = P\Lambda_u^{1/2}U_0',$$

und

$$Y'Y = U_0\Lambda_u U_0' = Q\Lambda_v Q',$$

denn $Y'Y = XX'$, so dass $Q = U_0$ und $\Lambda_u = \Lambda_v$. Weiter ist

$$YY' = V_0\Lambda_v V_0' = P\Lambda_u P',$$

denn $YY' = X'X$, so dass $P = V_0$, weshalb aus (34)

$$X = U_0\Lambda^{1/2}V_0', \text{ und} \tag{35}$$
$$Y = V_0\Lambda^{1/2}U_0' \tag{36}$$

folgt mit $\Lambda = \Lambda_u = \Lambda_v$. Wegen $Y = X'$ ist (36 ) natürlich redundant. □

[2]Die Repräsentation einer symmetrischen Matrix $M$ durch die Eigenvektoren $N$ und -werte $\Lambda_M$ in der Form $M = N\Lambda_M N'$ ist bekanntlich eindeutig.

**Anmerkungen:** Geht man von den Repräsentationen (26) und (27) aus ohne zunächst eine Beziehung zwischen ihnen zu postulieren, so wird man gleichwohl auf die Aussagen (31) und damit auf die SVD (32) geführt. Die Beziehungen zwischen den Zeilen- und Spaltenkategorien werden also durch *einen* Satz von latenten Variablen konstituiert, die sich allerdings in verschiedenen Skalen für die Zeilenkategorien einerseits und die Spaltenkategorien andererseits äußern. Andererseits kann man die Zerlegung (32) für *jede* Matrix $X$ finden, – also auch für $X \approx 0$, wenn also $x_{ij} \approx 0$ für alle $i, j$. Dieser Fall bedeutet $n_{ij} \approx \hat{n}_{ij}$, d.h. die Abwesenheit irgendeiner systematischen Beziehung zwischen den Zeilen- und den Spaltenkategorien. □

Bei der üblichen Hauptachsentransformation (PCA[3]) werden die Achsen so gewählt, daß die $D_k$ jeweils maximale Varianzanteile der Daten erklären. Damit hat man eine bestimmte euklidische Metrik gewählt. Man kann zu einer anderen euklidischen Metrik übergehen. wenn man die Projektionen (d.h. die Koordinaten) der Punkte (Personen oder Tests) auf eine Achse mit einer Zahl multipliziert; die Punktekonfiguration wird dann entlang dieser Achse gedehnt, wenn die Zahl größer als 1 ist, und ist sie kleiner als 1, so wird sie gestaucht.

Die spezielle Metrik, die eine Maximierung der Varianz pro Dimension bedeutet, macht dann Sinn, wenn man - wie bei Meßwerten üblich - annehmen kann, daß die Meßwerte in der Form $x_{ij} = \mu_{ij} + \xi_{ij}$ geschrieben werden können. Dabei ist $\xi_{ij}$ eine zufällige Veränderliche, die einen "Meßfehler" repräsentiert. Bei Häufigkeiten $n_{ij}$ einer Kontingenztabelle ist eine solche Darstellung aber nicht adäquat; die Häufigkeiten in einer Zeile sind z.B. multinomialverteilt und eine additive Zerlegung in einen "wahren" Wert $\mu_{ij}$ und einen Fehlerterm $\xi_{ij}$ ist nicht möglich. Deshalb muß man fragen, ob eine Metrik gefunden werden kann, die die Variation ("Varianz") in den Daten in sinnvoller Weise abbildet.

Für eine Kontingenztabelle lassen sich die Abhängigkeiten bzw. Zufälligkeiten durch das $\chi^2$ ausdrücken. Dementsprechend kann man versuchen, die Metrik so zu wählen, daß die latenten Dimensionen jeweils bestimmte $\chi^2$-Komponenten des Gesamt-$\chi^2$ der Tabelle repräsentieren; die Summe der Komponenten ergibt das Gesamt-$\chi^2$. Diese $\chi^2$-Komponenten entsprechen dann den Varianzanteilen bei einer Hauptachsentransformation. Dementsprechend wird man fordern, daß die Koordinaten (d.h. die Skalenwerte) der Zeilen- bzw. der Spaltenkategorien so gewählt werden sollen, daß die *euklidischen Distanzen* zwischen zwei Zeilen- bzw. zwei Spaltenpunkten den entsprechenden $\chi^2$-Distanzen entsprechen. Dies erfordert eine entsprechende Skalierung der Vektoren bzw. Komponenten von $U_0$ und $V_0$ in (32). Es läßt sich nun zeigen (Anhang, Abschnitt 8.3), daß die folgende Aussage gilt:

**Satz 3** *Die in (24) definierte euklidische Distanz $d_{ii'}$ entspricht der $\chi^2$-Distanz $\delta_{ii'}$, d.h. es gilt*

$$d_{ii'} = \delta_{ii'} \tag{37}$$

*genau dann, wenn die $f_{ik}$, $f_{i'k}$ in (24) wie in*

$$f_{ik} = u_{ik} \frac{\sqrt{\lambda_k}}{\sqrt{r_i}}, \tag{38}$$

*mit $i = 1, \ldots, I$, $k = 1, \ldots, s$ definiert sind, d.h. wenn die Elemente $u_{ik}$ der Matrix $U_0$ mit den Faktoren $\sqrt{\lambda_k}/\sqrt{r_i}$ gewichtet werden. Analog gilt für die Spaltenkategorien, daß*

$$d_{jj'} = \delta_{jj'} \tag{39}$$

[3]Principal Component Analysis

*gilt, wenn die Koordinaten der $S_j$ durch*

$$g_{jk} = v_{jk}\frac{\sqrt{\lambda_k}}{\sqrt{c_j}} \tag{40}$$

*mit $j = 1, \ldots, J$ gegeben sind; die Elemente $v_{jk}$ der Matrix $V_0$ werden also mit den Faktoren $\sqrt{\lambda_k}/\sqrt{c_j}$ gewichtet.*

Definiert man

$$D_r^{-1/2} = \begin{pmatrix} 1/\sqrt{r_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{r_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{r_I} \end{pmatrix} \tag{41}$$

$$D_c^{-1/2} = \begin{pmatrix} 1/\sqrt{c_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{c_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/\sqrt{c_J} \end{pmatrix} \tag{42}$$

so lassen sich die die $f_{ik}$ als Elemente der Matrix $F$ und die $g_{jk}$ als Elemente der Matrix $G$ angeben, die gemäß

$$F \stackrel{def}{=} D_r^{-1/2} U \Lambda^{1/2} \tag{43}$$

$$G \stackrel{def}{=} D_c^{-1/2} V \Lambda^{1/2} \tag{44}$$

bestimmt werden.

### 4.2.2 Die Verallgemeinerte SVD

Im Anhang wird die *Single Value Decomposition* (SVD) eingeführt. Nach (106), Seite 50 im Anhang gilt für eine reelle Matrix $X$ stets die Zerlegung $X = U\Lambda^{1/2}V'$, wobei $U_0$ und $V_0$ in (32) 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 10 sind in Gleichung (20) 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

$$X = D_r^{-1/2}(P - rc')D_c^{-1/2} = U\Lambda^{1/2}V'. \tag{45}$$

Dann folgt weiter

$$P - rc' = D_r^{1/2}U\Lambda^{1/2}V'D_c^{1/2}. \tag{46}$$

Es seien nun

$$A = D_r^{1/2}U, \quad B = D_c^{1/2}V. \tag{47}$$

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

$$U'U = A'D_r^{-1}A = I, \quad V'V = B'D_c^{-1}B = I, \tag{48}$$

und statt (46) läßt sich

$$P - rc' = A\Lambda^{1/2}B' \tag{49}$$

schreiben. (49) 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 5** *Die Gleichung (49) heißt, zusammen mit (48), die* generalisierte SVD *der Matrix $P - rc'$. (vergl. Greenacre (1984), p. 87).*

Die Koordinaten $F$ und $G$ lassen sich dann in der Form

$$\begin{aligned} F &= D_r^{-1}A\Lambda^{1/2} && (50)\\ G &= D_c^{-1}B\Lambda^{1/2} && (51) \end{aligned}$$

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

### 4.2.3 Die Beziehung zwischen den Koordinaten

Es wird zunächst der folgende Hilfssatz bewiesen:

**Hilfssatz:** Es gelten die Gleichungen

$$\begin{aligned} r'F &= 0 && (52)\\ c'G &= 0. && (53) \end{aligned}$$

**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

$$D_r^{-1}r = \vec{\mathbf{1}}, \quad D_c^{-1}c = \vec{\mathbf{1}}, \tag{54}$$

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. (5))

$$r'\vec{\mathbf{1}} = \sum_i r_i = 1, \quad c'\vec{\mathbf{1}} = \sum_j c_j = 1. \tag{55}$$

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

Zum Beweis von (52) und (53) 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

$$F = (D_r^{-1}P - \mathbf{1}c')D_c^{-1}B \tag{56}$$
$$G = (D_c^{-1}P' - \mathbf{1}r')D_r^{-1}A, \tag{57}$$

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 (55). Damit folgt $r'F = 0$. Auf analoge Weise zeigt man $c'G = 0$.

□

**Die Beziehung zwischen den Koordinaten: Übergangsgleichungen.** Es gilt

$$G = D_c^{-1}P'F\Lambda^{-1/2}, \tag{58}$$
$$F = D_r^{-1}PG\Lambda^{-1/2} \tag{59}$$

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

$$D_r^{-1}(P - rc')D_c^{-1} = D_r^{-1}A\Lambda^{1/2}B'D_c^{-1}, \tag{60}$$

bzw. in transponierter Form

$$D_c^{1/2}(P' - cr')D_r^{1/2} = D_c^{-1}B\Lambda^{1/2}A'D_r^{-1}. \tag{61}$$

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

$$D_r^{-1}PD_c^{-1} - D_r^{-1}rc'D_c^{-1} = FB'D_c^{-1}, \tag{62}$$

und andererseits

$$D_c^{-1}P'D_r^{-1} - D_c^{-1}cr'D_r^{-1} = GA'D_r^{-1}. \tag{63}$$

Multipliziert man (62) von rechts mit $B\Lambda^{1/2}$ und (63) mit $A\Lambda^{1/2}$, so erhält man wegen (48)

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

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

$$D_r^{-1}PG = F\Lambda^{1/2} \tag{64}$$
$$D_c^{-1}P'F = G\Lambda^{1/2} \tag{65}$$

folgt. Multiplikation von rechts mit $\Lambda^{-1/2}$ liefert dann gerade die Gleichungen (58) und (59). □

### 4.2.4 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 (64) 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 (59) 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

$$(D_r^{-1}PD_c^{-1}P')F = F\Lambda \tag{66}$$

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

$$(D_c^{-1}P'D_r^{-1}P)G = G\Lambda \tag{67}$$

gilt.

### 4.2.5 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 (**??**) dann den Ausdruck

$$x_{ij} = u_{i1}\sqrt{\lambda_1}v_{j1} + u_{i2}\sqrt{\lambda_2}v_{j2} + \cdots + u_{ir}\sqrt{\lambda_s}v_{jr} \tag{68}$$

Das normierte Residuum $x_{ij}$ ergibt sich also als Skalarprodukt des $i$-ten Zeilenvektors $(u_{i1}, \ldots, u_{ir})'$ von $U$ und des $j$-ten Zeilenvektors $(\sqrt{\lambda_1}v_{j1}, \ldots, \sqrt{\lambda_r}v_{jr})'$ von $\Lambda^{1/2}V$. 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 $e_{ij} = n_{i.}n_{.j}/N$. Dann gilt (vergl. (23))

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

gilt. Wie im Anhang, Abschn. 8.4 gezeigt wird, ergibt sich hieraus unter Berücksichtigung von (43) und (44) die *Rekonstitutionsformel*

$$P = D_rF\Lambda^{-1/2}G'D_c + E. \tag{69}$$

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.

# 5 Die 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 $\chi^2$ führen.

## 5.1 Die Zerlegung des $\chi^2$

Für die Zerlegung des $\chi^2$ gelten die folgenden Aussagen (vergl. Anhang, Abschn. 8.5), wobei $F_k$ der $k$-te Spaltenvektor von $F$ ist:

$$F_k^{'} D_r F_k = \lambda_k \tag{70}$$

$$In(i) = \frac{\chi_{i\cdot}^2}{N} = \sum_{k=1}^{s} \lambda_k u_{ik}^2 \tag{71}$$

$$In(j) = \frac{\chi_{\cdot j}^2}{N} = \sum_{k=1}^{s} \lambda_k v_{jk}^2 \tag{72}$$

$$In(K) = \frac{\chi^2}{N} = \sum_{k=1}^{s} \lambda_k \tag{73}$$

Die Gesamt-Inertia und damit das Gesamt-$\chi^2$ ist also nach (73) durch die Summe der Eigenwerte $\lambda_k$ 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.

## 5.2 Trägheitsanteil

**Definition 6** *Der Quotient*

$$\pi_k \stackrel{def}{=} \frac{\lambda_k}{\sum_k \lambda_k} = \frac{\lambda_k}{In(K)} \tag{74}$$

*heißt* Trägheits- *oder* Inertia-Anteil*; $\pi_k$ 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 $100\pi_k$.

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

## 5.3 Relative Trägheit

**Definition 7** *Die Quotienten*

$$\rho_{i\cdot} \stackrel{def}{=} \frac{In(i)}{In(K)} \tag{75}$$

$$\rho_{\cdot j} \stackrel{def}{=} \frac{In(j)}{In(K)} \tag{76}$$

*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

$$\rho_{i\cdot} = \frac{\chi^2_{i\cdot}}{\chi^2}, \quad \rho_{\cdot j} = \frac{\chi^2_{\cdot j}}{\chi^2}, \tag{77}$$

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

Nach (70) der Eigenwert $\lambda_k$, $k = 1, \ldots, s$, 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 (70)

$$\lambda_k = f_{k1}^2 r_1 + f_{k2}^2 r_2 + \cdots + f_{kI}^2 r_I \tag{78}$$

Setzt man diesen Ausdruck in (74) 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 $\pi_k$ setzt sich demnach additiv wiederum aus den Anteilen

$$\pi_{i\cdot k} \stackrel{def}{=} f_{ik}^2 r_i / In(K), \quad i = 1, \ldots, I \tag{79}$$

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 8** *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.

## 5.4 Qualität

In (16) ist die $\chi^2$-Distanz zwischen der $i$-ten Zeilenkategorie und dem mittleren Zeilenprofil angegeben worden. Nach (37) bzw. (39) entsprechen die $\chi^2$-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 9** *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*

$$q_{i\cdot} = \frac{\hat{d}_{i\cdot}}{d_{i\cdot}} \tag{80}$$

*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.

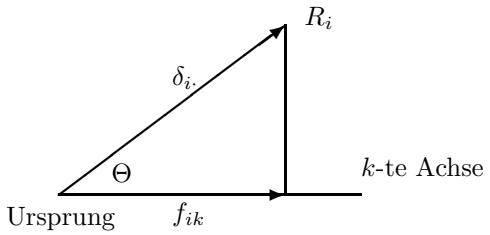
### 5.5 Der $\cos^2$-Anteil

Nach (17) ist

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

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

Abbildung 1: Projektion von $R_i$ auf die $k$-te Achse




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 $\Theta$ 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

$$\cos^2(\theta) = \frac{f_{ik}^2}{\delta_{i\cdot}^2} = \frac{f_{ik}^2 r_i}{\delta_{i\cdot}^2 r_i} \tag{81}$$

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 $\chi^2$-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.

# 6 Beziehungen zu anderen Verfahren und Anwendungen

## 6.1 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

$$U = XA', \quad V = YB', \tag{82}$$

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

$$R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx} A = A\Lambda^2 \tag{83}$$

$$R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy} B = B\Lambda^2 \tag{84}$$

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 $\Lambda$.

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 Kategorisierun-

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

gen für alle Personen 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

$$K = Z_1' Z_2 \tag{85}$$

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 $P$:

$$P = \frac{1}{m} Z_1' Z_2. \tag{86}$$

Weiter findet man, dass

$$D_r = \frac{1}{m} Z_1' Z_1, \quad D_c = \frac{1}{m} Z_2' Z_2, \tag{87}$$

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 4.2.4 wurde gezeigt, dass die Skalenwerte $F$ und $G$ für die Zeilen- bzw. Spaltenkategorien die Lösungen der Eigenvektorgleichungen

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

sind. Wgen (86) und (87) sieht man, dass diese Gleichungen den Gleichungen (83) und (84) äqivalent sind. Die Spalten von $F$ und $G$ entsprechen also kanonischen Variablen, und die Inertiaanteile in $\Lambda$ entsprechen kanonischen Korrelationskoeffizienten.

## 6.2 Multiple Korrespondenzanalyse

Die multiple Korrespondenzanalyse wird auf Indikatormatrizen angewendet. Solche Matrizen sind im Prinzip bereits in Abschnitt 6.1 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 2;

Tabelle 2: $Q$-variate Indikatormatrix

| | $\overbrace{\quad}^{J_1}$ | $\overbrace{\quad}^{J_2}$ | | $\overbrace{\quad}^{J_Q}$ |
|---|---|---|---|---|
| 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 |

### 6.2.1 Spezialfall: die bivariate Indikatormatrix ($Q = 2$)

Dieser Fall repräsentiert die Rohdaten für eine $(J_1 \times J_2)$-Kontingenztabelle. Ist $Z_1$ 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

$$Z = [Z_1, Z_2] \tag{88}$$

darstellbar; Tabelle 3 erläutert die Konstruktion einer solchen Indikatormatrix. Die Kontingenztabelle $K$ ergibt sich durch das Produkt

$$K = Z_1'Z_2. \tag{89}$$

Die Tabelle 4 ist die der Tabelle 3 entsprechende Kontingenztabelle. Man beachte, dass nach Tabelle 4 die Gesamtzahl der Fälle gleich 193 ist, dass aber in Tabelle 3 ein Gesamtwert von 386 auftritt. Das liegt daran, dass in der Tabelle 3 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

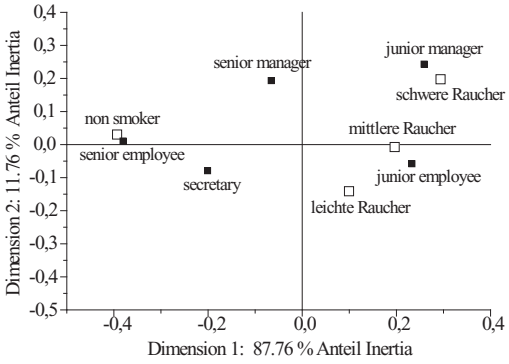
Tabelle 3: Status und Rauchgewohnheit; SM = Senior Manager, JM = Junior Manager, SE = senior employee, JE = junior employee, SC = secretary, no = Nichtraucher, li = light, med = medium, hv = heavy. (1, 1): Person ist SM und li-Raucher, (5,4): Person ist Sekretär(in) und schwere(r) Raucher(in).

| Klassif. | SM | JM | SE | JE | SC | no | li | med | hv | $\Sigma$ |
|---|---|---|---|---|---|---|---|---|---|---|
| (1, 1) | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| (1, 1) | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| ⋮ | | | | | | | | | | |
| (1,2) | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 |
| ⋮ | | | | | | | | | | |
| (1, 3) | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 |
| ⋮ | | | | | | | | | | |
| (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 |

Tabelle 4: 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 |

Abbildung 2: Typ der Anstellung und Rauchverhalten



der Tabelle 3 als Spaltensummen für $Z_1$ auf, und die Spaltensummen von Tabelle 4 treten als Spaltensummen von $Z_2$ in Tabelle 3 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

$$P^Z = \frac{1}{2N} Z \tag{90}$$

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

$$D_r^Z = \frac{1}{N} I, \quad D_c^Z = \frac{1}{2} \begin{pmatrix} D_r & 0 \\ 0 & D_c \end{pmatrix}, \tag{91}$$

mit $D_r$ und $D_c$ die Diagonalmatrizen der relativen Zeilen- bzw. Spaltenhäufigkeiten von $K$. Nach (66) 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 $\Gamma^Z$ von $Z$ in analoger Form durch die Beziehung

$$\left[ 2 \begin{pmatrix} D_r^{-1} & 0 \\ 0 & D_c^{-1} \end{pmatrix} \frac{1}{2} Z' N \frac{1}{2N} Z \right] \Gamma^Z = \Gamma^Z D_\lambda^Z \tag{92}$$

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

$$B^* = \begin{pmatrix} Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2 \end{pmatrix} \tag{93}$$

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. (88), und $Z_1'Z_1$ und $Z_2'Z_2$ sind die Diagonalmatrizen der Zeilen- bzw. Spaltenhäufigkeiten von $K$. In entsprechender Weise kann man $\Gamma^Z$ aufteilen:

$$\Gamma^Z = \begin{pmatrix} \Gamma_1^Z \\ \Gamma_2^Z \end{pmatrix}, \tag{94}$$

wobei $\Gamma_1^Z$ $J_1$ Zeilen und $\Gamma_2^Z$ $J_2$ Zeilen hat. (92) läßt sich dann in der kompakten Form

$$\frac{1}{2N} \begin{pmatrix} D_r^{-1} & 0 \\ 0 & D_c^{-1} \end{pmatrix} \begin{pmatrix} Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2 \end{pmatrix} \begin{pmatrix} \Gamma_1^Z \\ \Gamma_2^Z \end{pmatrix} = \begin{pmatrix} \Gamma_1^Z \\ \Gamma_2^Z \end{pmatrix} D_\lambda^Z \tag{95}$$

schreiben. Multipliziert man diese Gleichungen aus, so erhält man

$$\frac{1}{2N}(D_r^{-1}Z_1'Z_1\Gamma_1 + D_r^{-1}Z_1'Z_2\Gamma_2) = \Gamma_1 D_\lambda \quad (96)$$

$$\frac{1}{2N}(D_c^{-1}Z_2'Z_1\Gamma_1 + D_c^{-1}Z_2'Z_2\Gamma_2) = \Gamma_2 D_\lambda. \quad (97)$$

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

$$\Gamma_1^Z + D_r^{-1}P\Gamma_2^Z = 2\Gamma_1^Z D_\lambda^Z \quad (98)$$

$$D_c^{-1}P'\Gamma_1^Z + \Gamma_2^Z = 2\Gamma_2^Z D_\lambda^Z \quad (99)$$

folgt. Multipliziert man (98) von links mit $D_c^{-1}P'$ und setzt man den Ausdruck für $D_c^{-1}P'\Gamma_1^Z$ aus (99) ein, so erhält man

$$D_c^{-1}P'D_r^{-1}P\Gamma_2^Z = \Gamma_2^Z(2D_\lambda^Z - I)(2D_\lambda^Z - I). \quad (100)$$

Auf analoge Weise erhält man

$$D_r^{-1}PD_c^{-1}P'\Gamma_1^Z = \Gamma_1^Z(2D_\lambda^Z - I)(2D_\lambda^Z - I). \quad (101)$$

Die Gleichungen (100) und (101) sind Gleichungen für die Eigenvektoren $\Gamma_1^Z$ und $\Gamma_2^Z$; gleichzeitig entsprechen diese Gleichungen den Gleichungen (66) und (67), so dass die Lösungen dafür auch Lösungen für (100) und (101) 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

$$\lambda = (2\lambda^Z - 1)^2, \text{ oder } \lambda^Z = (1 \pm \lambda^{1/2})/2. \quad (102)$$

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.

### 6.2.2 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

$$B = Z'Z = \begin{pmatrix} Z_1'Z_1 & Z_1'Z_2 & \cdots & Z_1'Z_Q \\ Z_2'Z_1 & Z_2'Z_2 & \cdots & Z_2'Z_Q \\ & & \vdots & \\ Z_Q'Z_1 & Z_Q'Z_2 & \cdots & Z_Q'Z_Q \end{pmatrix}. \quad (103)$$

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 $Z_q$.

Tabelle 5: Körperbau und psychische Erkrankung: beobachtete und erwartete Häufigkeiten

| | | Erkrankung | | | |
|---|---|---|---|---|---|
| Typ | | man./dep. | Epilepsie | Schizophr. | $\Sigma$ |
| 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 | |
| | $\Sigma$ | 1361 | 1505 | 5233 | $N = 8099$ |

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

$$\lambda^B = (\lambda^Z)^2. \tag{104}$$

Eine Illustration der Konstruktion und Anwendung einer Burt-Matrix wird in Beispiel 4, Seite 36, gegeben.

# 7 Beispiele

**Beispiel 1** Westphal (1931) erstellte eine Tabelle zum Zusammenhang zwischen Körperbau und Charakter, wie er etwa von Kretschmer (1961) diskutiert wurde. Westphal klassifizierte insgesamt 8099 Patienten in Psychiatrischen Landeskrankenhäusern (i) nach ihrem Körperbautyp, und (ii) nach der bei ihnen diagnostizierten Störung. Die Tabelle wird in Hofstätter (1971), p. 330, wiedergegeben; Hofstätter diskutiert die Daten, indem er Vierfelderkorrelationen zwischen bestimmten Störungen und bestimmten Körperbautypen berechnet. Die höchste Korrelation ist die zwischen pyknischem Körperbau und manisch-depressiver Störung; ein Zusammenhang, der auch in der folgenden korrespondenzanalytischen Analyse der Daten als besonders prominent hervorsticht. Diese Analyse liefert aber insgesamt einen deutlicheren Einblick in die unterliegende Struktur der Daten (vergl. Abb. 4, Seite 29) als die Betrachtung einzelner Korrelationskoeffizienten, zumal Vierfelderkoeffizienten bei ungleichen Randverteilungen eine systematische Unterschätzung des Zusammenhanges suggerieren können. Die Zeile "erw." enthält dabei die Häufigkeiten, die man gemäß

$$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}, \quad n_{i\cdot} = \sum_j n_{ij}, \quad n_{\cdot j} = \sum_i n_{ij} \tag{105}$$

erwarten würde, wären Körperbau und psychische Erkrankung unabhängig voneinander; $N = 8099$ ist die Gesamtzahl der Beobachtungen in der Tabelle.

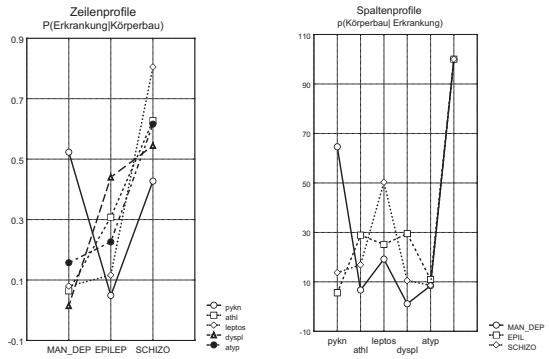
Tabelle 6 zeigt die relativen Häufigkeiten der Kombinationen von Körperbau und Erkrankung, d.h. die Werte $p_{ij} = n_{ij}/N$, zusammen mit den *Massen* der Kategorien; $r_i$ die Massen der Zeilen-, $c_j$ die Massen der Spaltenkategorien. Wie bereits der Tabelle 5 zu entnehmen ist, bilden die Leptosomen einerseits die größte Teilstichprobe, wenn man nach Körperbau klassifiziert, und die Schizophrenen andererseits, wenn man nach Erkrankung klassifiziert. Die Inspektion der Tabellen 5 und 6 zeigt außerdem, daß die Kombinationen pyknisch-manisch/depressiv, leptosom-Schizophrenie und dyplastisch-Epilepsie am häufigsten auftreten. Dementsprechend liegt es nahe, zu vermuten, daß es eine Assoziation zwischen Körperbau und Erkrankung gibt. Um die Art der Assoziationen genauer zu erfassen, wird man aber die Profile untersuchen müssen. Abb. 3 zeigt die Zeilen-

Tabelle 6: Körperbau und psychische Erkrankung: relative Häufigkeiten $p_{ij}$

| | Erkrankung | | | |
|---|---|---|---|---|
| Typ | man.-dep. | Epilepsie | Schizophr. | Massen $r_i$ ($\Sigma$) |
| 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 |
| Massen $c_j$ ($\Sigma$) | .1680 | .1858 | .6461 | 1.000 |

und Spaltenprofile der Datenmatrix. Die Zeilenprofile entsprechen den bedingten Wahr-

Abbildung 3: Zeilen- und Spaltenprofile



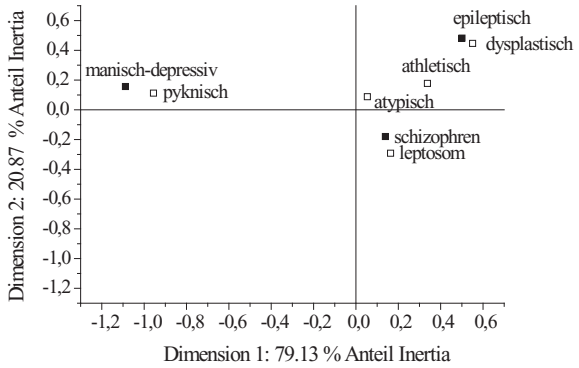

scheinlichkeiten $P$(Erkrankung|Körperbautyp), die Spaltenprofile den bedingten Wahrscheinlichkeiten $P$(Körperbautyp|Erkrankung). So sieht man, daß die Wahrscheinlichkeit, manisch-depressiv zu sein unter der Bedingung, einen pyknischen Körperbau zu haben, ca. 5-mal so groß ist wie die Wahrscheinlichkeit, unter dieser Bedingung an Epilepsie zu leiden. Die Wahrscheinlichkeit, an Schizophrenie zu leiden, ist für den Pykniker fast so groß wie die, manisch-depressiv zu sein.

Für einen Leptosomen ist allerdings die Wahrscheinlichkeit, an Schizophrenie zu leiden, wesentlich höher als für einen Pykniker. Dafür ist die Wahrscheinlichkeit, manisch-

depressiv zu sein, sehr viel geringer. Die Leptosomen werden also keinen Gegenpol zu den Pyknikern auf einer gemeinsamen Skala bilden; ein solcher Gegenpol könnte allenfalls von den Dyplastikern[4] dargestellt werden, da deren Profil eine gewisse Gegenläufigkeit zu dem der Pykniker darstellt. Athleten und Atypische werden irgendwo im Mittelfeld liegen, da ihre Profile dem durchschnittlichen Profil ähnlich sind.

Die Betrachtung der Spaltenprofile führt zu ähnlichen Hypothesen. Das Profil der manisch-depressiven Patienten ist gegenläufig zu dem der Epileptiker, so daß man diese beiden Krankheitstypen als Gegenpole einer Skale denken kann. Manisch-Depressive und Schizophrene haben einen qualitativ, nicht quantitativ ähnlichen Verlauf, unterscheiden sich aber bei den Pyknikern im Sinne einer Gegenläufigkeit. Man kann vermuten, daß Schizophrene und Manisch-Depressive nicht Gegenpole auf *einer* Skala sind, sondern aufgrund der teilweisen Gemeinsamkeit und der teilweisen Unterschiedlichkeit *verschiedene* Dimensionen definieren. Der Biplot (vergl. Abb. 4) schließlich liefert eine quantitative

Abbildung 4: Biplot: Kretschmertypen nach (1931)




Präzisierung dieser relativ lose und qualitativ formulierten Hypothesen. Man findet eine 2-dimensionale Lösung, wobei die erste Dimension ca 80 % der Inertia (bzw. des $\chi^2$) der Tabelle erklärt, und die zweite Dimension die restlichen 20 %. Die Projektionen von pyknisch und dysplastisch auf die erste Achse zeigen, daß diese beiden Körperbautypen in der Tat als Gegenpole auf einer Skala aufgefaßt werden können. Die Rangreihe der Projektionen auf die erste Achse ist, von links nach rechts, pyknisch, atypisch, leptosom, athletisch und schließlich dysplastisch. Auf der einen Seite hat man also den rundlichen Typ - nach Meinung vieler Philosophen ist die Kugel der vollkommene Körper - , auf der anderen den unharmonisch knochigen Typ. Alle nicht-pyknischen Typen liegen im "positiven" Bereich der Skala (die Wahl des Vorzeichens ist natürlich irrelevant), d.h. die Profile der anderen Typen sind einander relativ (im Vergleich zu dem des Pyknikers) ähnlich, unterscheiden sich relativ stark von dem des Pyknikers. Die Tatsache, daß der atypische Typ nahe am Ursprung des Koordinatensystems liegt entspricht der Tatsache, daß das Profil dieses Typs dem durchschnittlichen Profil sehr ähnlich ist. Betrachtet man die Projektionen der Körperbautypen auf die zweite Achse, so erhält man die Rangreihe - von unten nach oben - leptosom, atypisch, athletisch und schließlich dysplastisch. Die Leptosomen weichen auf der ersten Dimension kaum vom Mittelpunkt ab, sind also nahezu vollständig durch die zweite Dimension definiert; die bestätigt die Vermutung, daß sie eine von der Polarität pyknisch-dysplastisch unabhängige Dimension repräsentieren.

[4]Nach Kretschmer (1977) ist der dysplastische Typ ein Sammelbegriff für unharmonisch geformte Menschen.

Der Gegenpol der Leptosomen auf der zweiten Achse sind wiederum die Dysplastiker. Der Dysplastiker repräsentiert gewissermaßen die Grobknochigkeit, der leptosome Typ die Feingliedrigkeit.

Es bleibt noch, die Lage der Körperbautypen relativ zu der der Erkrankungen im 2-dimensionalen Koordinatensystem zu diskutieren. Wie aus Abb. 4 ersichtlich ist, existieren klare Korrespondenzen: die Punkte für pyknisch und manisch-dpressiv, leptosom und schizophren sowie dysplastisch und epileptisch liegen jeweils sehr nahe beieinander. Wie schon ausgeführt ist es aber nicht sinnvoll, die euklidische (bzw. pythagoräische) Distanz zwischen den Punkten die Erkrankungen einerseits und Körperbautypen andererseits repräsentieren, zu interpretieren. Den euklidischen Distanzen zwischen Körperbautypen einerseits und Erkrankungsarten andererseits entsprechen ja $\chi^2$-Distanzen zwischen den entsprechenden Zeilen bzw. Spalten, aber eine Distanz zwischen einer Zeilen- und einer Spaltenkategorie ist nicht definiert, so daß die Distanz zwischen einem Zeilen- und einem Spaltenpunkt keine Entsprechung in der ursprünglichen Tabelle der Häufigkeiten hat. Die Beziehung zwischen den Koordinaten $F$ und $G$ und damit zwischen den Zeilen- und Spaltenpunkten ist durch die Gleichungen (58) und (59) gegeben; diese Beziehung ist sicherlich zu unanschaulich, um bei der Interpretation eines Biplots von unmittelbarer Nützlichkeit zu sein. Man weiß aber andererseits, daß die Achsen für die Zeilen- und die Spaltenkategorien die gleiche Bedeutung haben müssen. Die Lage der Erkrankungen ist analog zu der der Körperbautypen; *deswegen* kann man folgern, daß es eine Korrespondenz zwischen bestimmten Körperbautypen und bestimmten Erkrankungen geben muß.

Die Betrachtung der Werte der normierten Residuen $x_{ij} = (n_{ij} - n_{i\cdot}n_{\cdot j}/N)/\sqrt{n_{i\cdot}n_{\cdot j}/N}$ der Matrix $X$ stüzt diese Vermutung, denn die Residuen lassen sich ja nach der Gleichung (112) aus den Koordinaten $F$ und $G$ rückrechnen. Die Tabelle 1 zeigt die $x_{ij}$. Die Werte für die Kombinationen pyknisch-manisch/depressiv, leptosom-Schizophrenie

Tabelle 7: Körperbau und psychische Erkrankung: skalierte Werte $x_{ij}$

| 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 |

und dysplastisch-Epilepsie sind sind die jeweils größten positiven Werte, der Wert für die Kombination athletisch-Epilepsie ist ebenfalls noch positiv. Positive $x_{ij}$-Werte bedeuten, daß die beobachteten Häufigkeiten größer als die bei nur zufälliger Assoziation von Körperbau und Erkrankung erwarteten Häufigkeiten sind. Die betragsmäßig größten negativen Werte sind die für die Kombination pyknisch-Epilepsie, gefolgt von pyknisch-Schiziphrenie, leptosom-manisch/depressiv, gefolgt von leptosom-Epilepsie, und dyplastisch-manisch/depressiv. Die Abweichungen der beobachteten von den erwarteten Werten für den atypischen Körperbau sind alle gering: das Profil für den atypischen Körperbau entspricht einigermaßen genau dem des durchschnittlichen Profils der Körperbautypen.

Die Tabelle 8 gibt einige "globale" Statistiken an: Da die Tabelle nur drei Spalten hat, gibt es zwei von Null verschiedene Eigenwerte. Der erste Eigenwert $\lambda_1$ ist ca 3.8-mal so groß wie der zweite $\lambda_2$, und dementsprechend sind die Inertia und das $\chi^2_1$ für die erste

Tabelle 8: Globale Statistiken

| Dim. | $\chi^2$ | Anteil Inertia ($\pi_k$) | Eigenwert |
|---|---|---|---|
| 1 | $\chi_1^2 =$ 2090.152 | .791 | .258 |
| 2 | $\chi_2^2 =$ 551.407 | .021 | .068 |
| $\sum$ | $\chi_g^2 =$ 2641.559 | 1.000 | In(K) = .326 |

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.*

Es sollen nun die Qualitätsmerkmale der Repräsentation betrachtet werden. Zunächst ist dabei an die Qualität $q_{i\cdot}$ (vergl. Gleichung (80)) 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.

In der Tabelle 9 sind alle relevanten Statistiken zusammengefaßt worden. Die Spalte $f_{i1}$ enthält die Koordinaten für die erste Achse, $f_{i2}$ die für die zweite Achse. Pykniker und

Tabelle 9: Koordinaten und Statistiken: Körperbau und Erkrankung

| Körperbau | $f_{i1}$ | $f_{i2}$ | $\pi_{i\cdot 1}$ | $\cos^2\Theta_{11}$ | $\pi_{i\cdot 2}$ | $\cos^2\Theta_{21}$ | $\rho_{i\cdot}$ |
|---|---|---|---|---|---|---|---|
| pyknisch | -.9559 | .1125 | .7340 | .9863 | .0385 | .0136 | .5888 |
| athletisch | .3373 | .1775 | .0776 | .7831 | .0806 | .2169 | .0776 |
| leptosom | .1632 | -.2918 | .0417 | .2384 | .5052 | .7616 | .1384 |
| dysplast. | .5509 | .4469 | .1465 | .6032 | .3654 | .3968 | .1922 |
| atypisch | .0540 | .0883 | .0010 | .2720 | .0103 | .7280 | .0030 |
| **Erkrankung** | $g_{j1}$ | $g_{j2}$ | $\pi_{\cdot j1}$ | $\cos^2\Theta_{12}$ | $\pi_{\cdot j2}$ | $\cos^2\Theta_{22}$ | $\rho_{\cdot j}$ |
| man.-dep. | -1.0883 | .1569 | .7712 | .9796 | .0607 | .0203 | .6229 |
| Epilepsie | .5005 | .4819 | .1803 | .5189 | .6338 | .4811 | .2751 |
| Schizophr. | .1391 | -.1794 | .0484 | .3775 | .3054 | .6245 | .1021 |

| | |
|---|---|
| $f_{ik}$, $g_{jk}$ | 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 $In(K)$ |

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 erzeugen 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-$\chi^2$ bzw. zur Geamt-Inertia beiträgt. Der Beitrag zum Gesamt-$\chi^2$ wird durch die Länge eines Vektors, dessen Endpunkt eine Kategorie repräsentiert, bestimmt, denn diese Länge entspricht der $\chi^2$-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.

Zwei grundsätzliche Bedenken sollten angemerkt werden: (i) Westphal hat das Alter der Patienten nicht mit erhoben, (ii) es ist nicht klar, ob Westphal nicht bereits die Patienten nach Maßgabe der Theorie Kretschmers kategorisiert hat; die im Biplot aufscheinenden Beziehungen zwischen den Kategorien würden dann eine Form einer *Erschleichung des Beweises* repräsentieren. □

**Beispiel 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 10 enthält die Häufigkeiten der Antworten, aufgeschlüsselt nach dem religiösen Bekenntnis der Befragten, sowie die Zeilen- und Spalten-$\chi^2$ und das Gesamt-$\chi^2$ ( = 40.175).

Tabelle 10: Religiöse Präferenz und Einstellung zur Abtreibung (Marscuilo und McSweeny 1977)

| Response | Catholic | Protestant | Jewish | Others | $\sum$ | Zeilen-$\chi^2$ |
|---|---|---|---|---|---|---|
| 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-$\chi^2$ | 8.627 | 5.932 | 7.683 | 18.132 | | 40.175 |

Tabelle 11: 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 11 angegeben. Abbildung 5 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 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

Abbildung 5: Zeilen- und Spaltenprofile

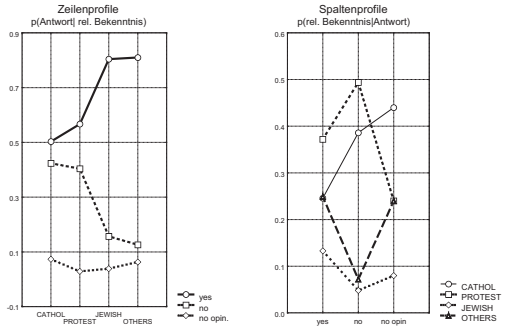
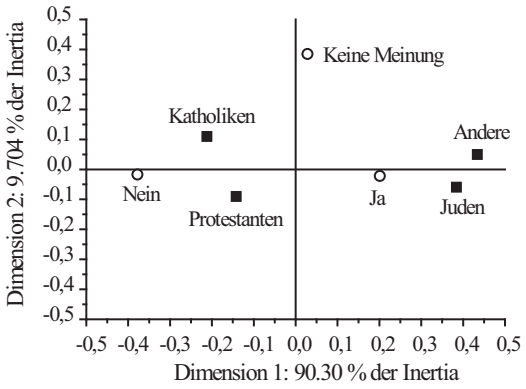



Abbildung 6: Religiöses Bekenntnis und Einstellung zur Abtreibung: Biplot




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 6 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 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.

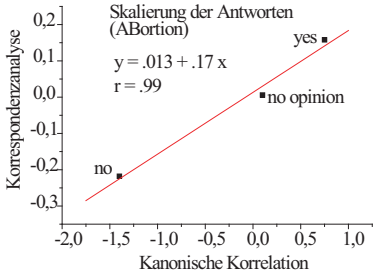
Schaut man im Biplot auf die Antworten, so wird deutlich, daß die erste Achse durch die Antworten "nein" und "ja" bestimmt zu sein scheint. Die logische Beziehung zwischen "yes", "no" und "no opinion" verlangt, daß "no opinion" zwischen "yes" und "no" liegt, und für die erste Dimension gilt dies auch. 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 $\chi^2$- 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" 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. 6 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 $\chi^2$-Distanzen der Profile zum entsprechenden mittleren Profil, und diesen Distanzen entsprechen die $\chi^2$-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 7 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 7: Religiöse Präferenz und Einstellung zur Abtreibung




**Beispiel 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 $\chi^2$ für diese Tabelle ist hochsignifikant: $\chi^2 = 2466.14$ bei $df = 12$ Freiheitsgraden

Tabelle 12: 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 |

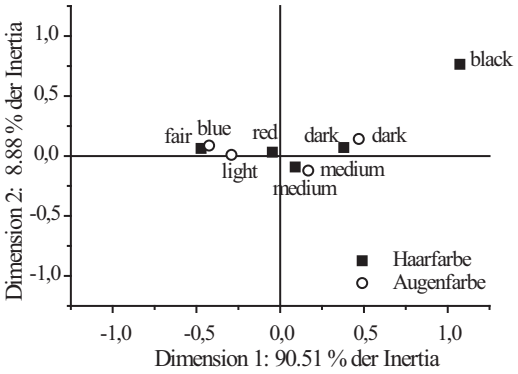
hat unter $H_0$ (kein Zusammenhang) eine Wahrscheinlichkeit von $p = .000$. Der statistische Zusammenhang zwischen Merkmalen läßt sich oft durch Korrelationen ausdrücken, – aber wie will man im Falle einer solchen Tabelle die Korrelation zwischen Augen- und Haarfarbe berechnen? Die Korrespondenzanalyse liefert Skalen, in Bezug auf die die Merkmale verglichen werden können. Das Resultat der Korrespondenzanalyse dieser Daten wird in Abbildung 8 präsentiert. Offenbar erklärt die erste Dimension gut 90 % des Gesamt-$\chi^2$. Nur das sehr dunkle Haar ("black") scheint eine systematische Abweichung von der ersten Dimension zu erzeugen.

Die Häufigkeitsverteilungen in der Tabelle 12 weisen bereits auf eine enge Kopplung zwischen bestimmten Augen- und Haarfarben hin. Im Biplot (8) 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 $\chi^2$-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 $\chi^2$-Distanzen reflektieren also gewissermaßen genetische Nachbarschaften. □

**Beispiel 4** Burt (1950) bestimmte für 100 zufällig in Liverpool ausgewählte Personen

Abbildung 8: Beziehung zwischen Augen- und Haarfarbe (Maung (1941), nach Daten von Tocher (1908)



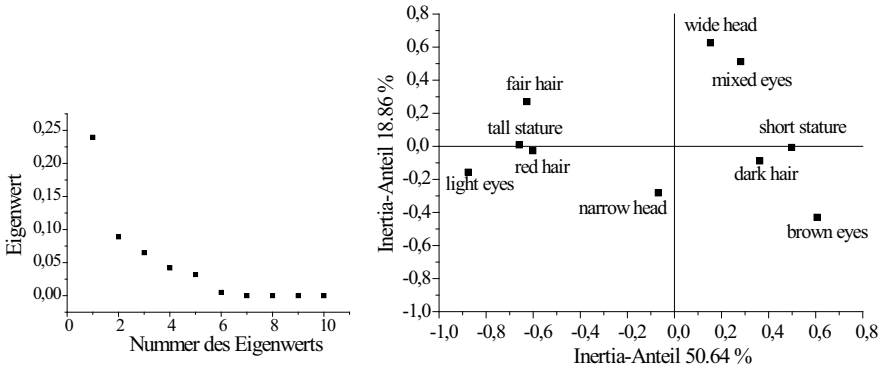

(i) die Haarfarbe (fair, red, dark), (ii) die Augenfarbe (light, medium, brown), (iii) Kopfform (narrow, wide), und (iv) die Statur (tall, short). Es ergibt sich zunächst eine 4-dimensionale Kontingenztabelle mit jeweils 3, 3, 2 und 2 Kategorien, die die Burt-Matrix 13 findet man in Basilevsky (1994), p.532. Die Tabelle hat die Eigenwerte .2395, .0892,

Tabelle 13: 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 |

.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 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. □

Abbildung 9: Eigenwerte und (Bi)Plot der Merkmale für die Daten aus Tabelle 13; Daten: vergl. Basilevsky (1994)




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

**Beispiel 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. 10 zeigt die Zeilen- und Spaltenprofile der Häufigkeiten.

Tabelle 14: Straftaten dänischer Jugendlicher

| | Alter | | | | | | |
|---|---|---|---|---|---|---|---|
| Jahr | 15 | 16 | 17 | 18 | 19 | $\sum$ | Zeilen-$\chi^2$ |
| 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-$\chi^2$ | 7.088 | 11.967 | 2.884 | 9.020 | 7.450 | | $\chi^2$ =38.410 |

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. (9), p. 7). 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

Abbildung 10: Dänische Jugendkriminalität und Polizeiverhalten: Profile

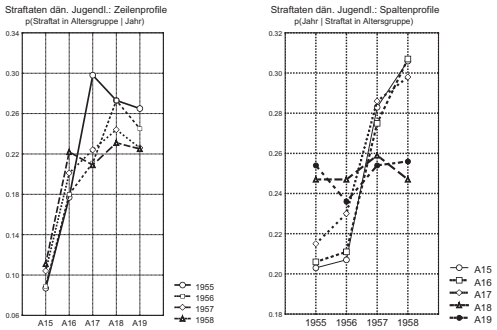



Abbildung 11: Dänische Jugendkriminalität und Polizeiverhalten: Biplot

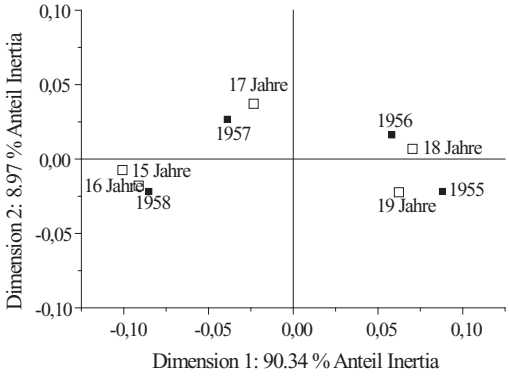


Tabelle 15: Koordinaten und Statistiken: Straftaten dän. Jugendlicher

| 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 | | | | | | | |

Straftaten der betrachteten Art[5] aufweist. Für die Jahre 1955 und 1956 fallen die 17-jährigen auf: 1955 werden von dieser 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 Ko-

[5]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.

ordinatensystem, 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 (69), 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 $\chi^2$-Metrik andererseits berücksichtigt werden. Schaut man nun in die Datentabelle 14, 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 10 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. □

**Beispiel 6** Die Daten der Tabelle 16 reflektieren die Verteilung von Ansichten über die Behandlung Krimineller in den USA in verschiedenen Jahren[6]. Das $\chi^2$ 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 $\chi^2/N$ erklärt werden, durch die zweite nur noch .067. Es sollen zuerst die Profile betrachtet werden (vergl. Abb. 12). Die Zeilenprofile - d.h. die (relativen) Häufigkeiten der Einstellungen für ein gegebenes Jahr - sind

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

Tabelle 16: 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 $= 12$, $p = .0000$, $In(K) = \chi^2/N = .016298$

Abbildung 12: Ansichten über die Behandlung von Kriminellen in US-Gefängnissen: Profile

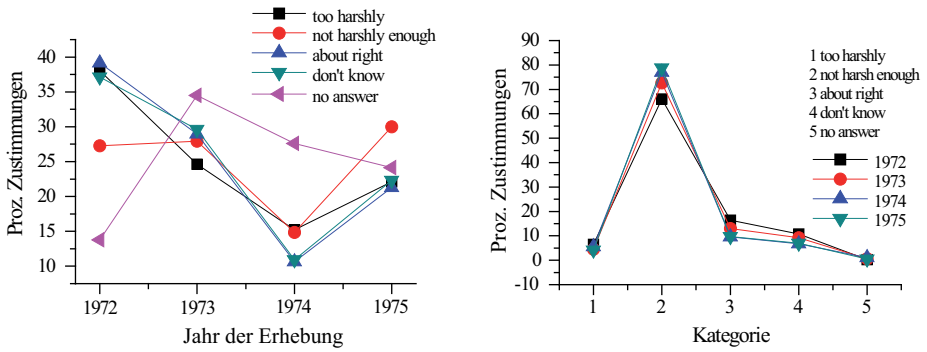


Tabelle 17: Koordinaten und Statistiken: Einstellung gegenüber Kriminellen in den USA

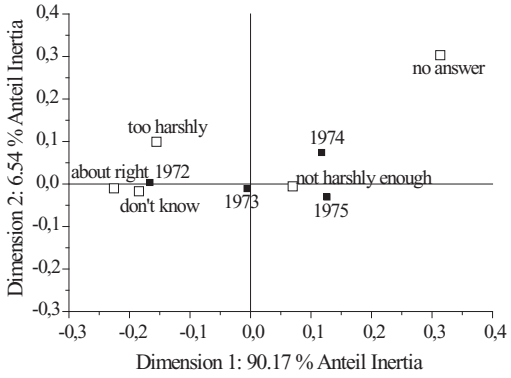
| Einstellung | $f_{i1}$ | $f_{i2}$ | $\pi_{i\cdot 1}$ | $\cos^2\Theta_{11}$ | $\pi_{i\cdot 2}$ | $\cos^2\Theta_{21}$ | $\rho_{i\cdot}$ | $q_i$ |
|---|---|---|---|---|---|---|---|---|
| too harshly | -.155 | .099 | .085 | .638 | .477 | .261 | .119 | .899 |
| not h. enough | .069 | -.005 | .241 | .993 | .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 |

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 17 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 richtig" (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-$\chi^2$ so gut wie vollständig; es ist also nicht notwendig, weitere Dimensionen zu betrachten. Schlüsselt man das $\chi^2$ hinsichtlich der Zeilenkategorien (Einstellungen) auf, so sieht man (Spalte $\rho_i$), daß "ungefähr richtig" (a.r.) den größten Anteil erklärt, während "keine Ant-

Abbildung 13: Ansichten über die Behandlung von Kriminellen in US-Gefängnissen: Biplot

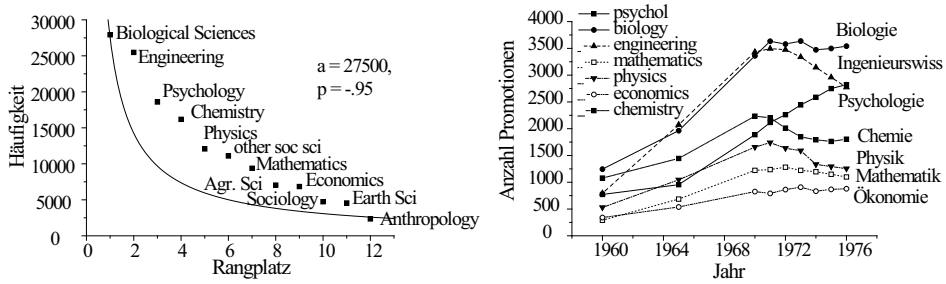



wort" nur einen vernachlässigbaren Anteil erzeugt. Dem entspricht, daß sich die Werte der Zeilenprofile für diese Einstellung über die Jahre so gut wie nicht unterscheiden. Bei den Jahreskategorien ist es das Jahr 1972, das die größten Unterschiede generiert, gefolgt von 1975. Schließlich kann man sich noch einen überblick darüber verschaffen, welche Kategorien durch welche Dimension am besten "erklärt" werden. Für die erste Dimension ist der $\chi^2$-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

Abbildung 14: Verteilung der Häufigkeiten der Fächer




Einstellung darüber, ob man seine Ansicht bekannt geben soll. □

**Beispiel 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 18. Den Zeilensummen der Tabelle entnimmt man

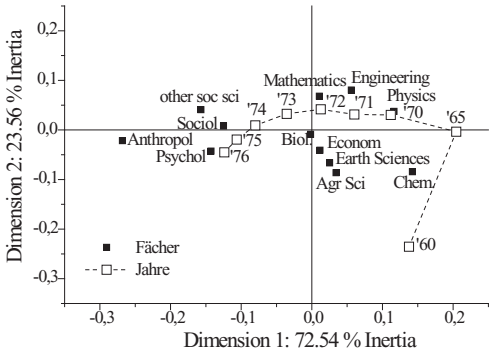
Tabelle 18: Trends bei Doktorgraden in den Jahren 1960 - 1976

| | 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 |

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. 14 zeigt die Verteilung der Häufigkeiten für die einzelnen Fächer. 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 "gemittelten" (d.h. aggregierten) Häufigkeiten für die Fächer. Die Kurve ist die Pareto-Kurve (vergl. Beispiel 8); diese Daten sind ein eher seltenes Beispiel für den *mangelnden* Fit des Pareto-Modells.

Die eigentlich bemerkenswerten Strukturen, die die Tabelle 18 enthält, zeigt der Biplot. Es wird ein klarer zeitlicher Verlauf sichtbar, der eine Veränderung der Präferenzen anzeigt. □

Abbildung 15: Biplot Doktorate - Jahre




**Beispiel 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 19 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. 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. 16 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,

Tabelle 19: 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 16: Häufigkeitsverteilung der Selbstmorde, aggregiert über Methoden

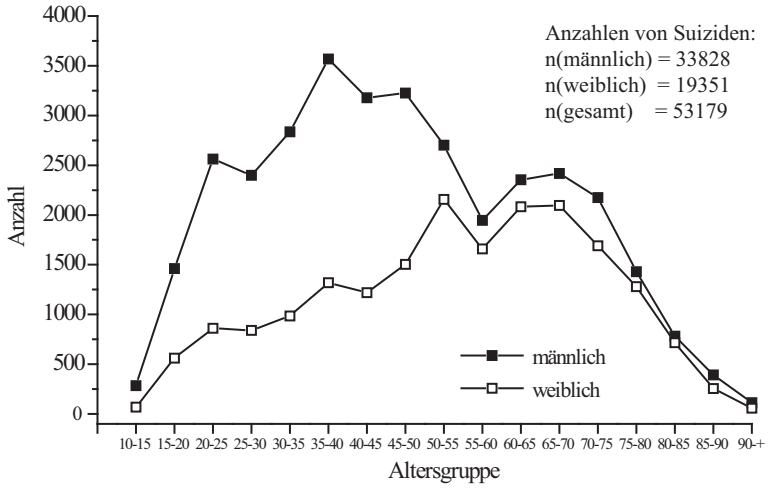
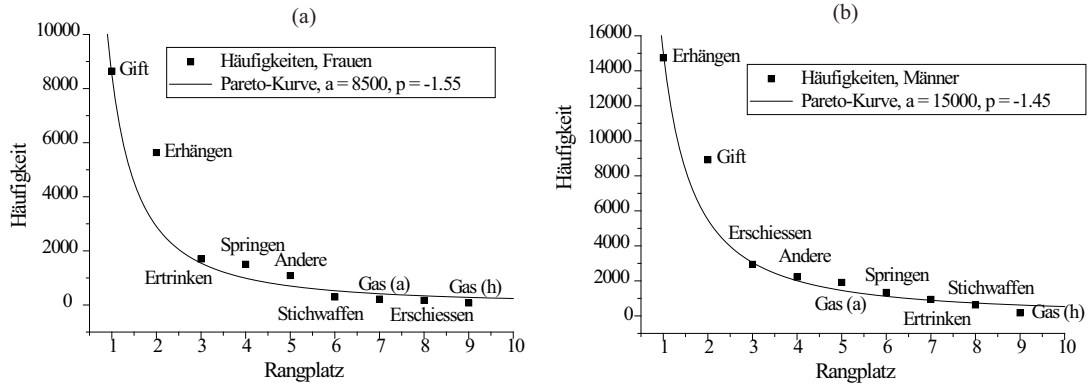
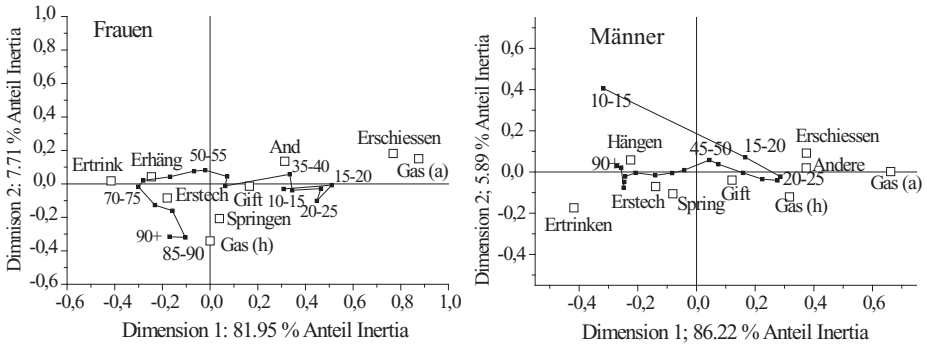



Abbildung 17: Ranggeordnete Häufigkeit der Methoden, (a) Frauen, (b) Männer.Zur Definition der Pareto-Kurve siehe Text.




daß die Verteilungen für die Frauen und die Männer von der Gruppe der 55-60-jährigen an nahezu parallel verläuft. Abb. 17 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. Pareto 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

Abbildung 18: Biplots Methode × Altersgruppen, für Frauen und Männer getrennt.



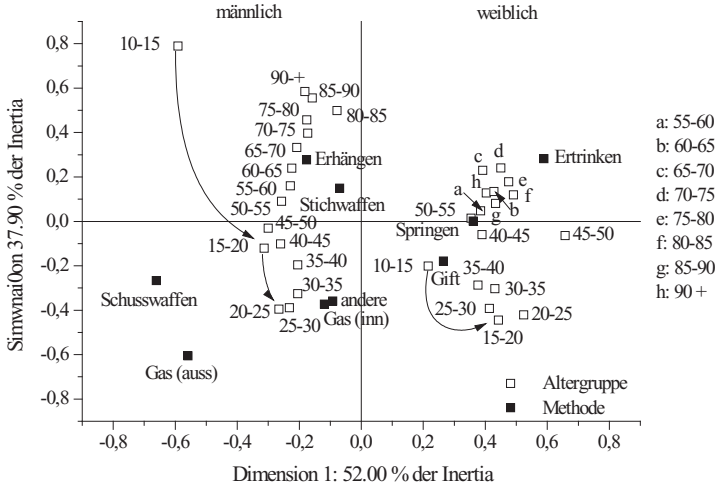

die Kurve als Zipfsches Gesetz bekannt: trägt man die Häufigkeit, mit der Worte in einer Sprache verwendet werden, gegen ihren entsprechenden Rangplatz auf, so entspricht die entstehende Kurve bei den meisten Sprachen einer Pareto-Formel. Die Formel gilt in vielen anderen Fällen ebenfalls in guter Näherung; 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. 17 aufscheinenden Wechselwirkung zwischen Methode 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. 18 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

Abbildung 19: Biplot: Selbstmorde: Methode, Altersgruppen und Geschlecht




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. 19 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. 18 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. 19 ist, dass die Frauen und Männer nach Methoden getrennt werden; die Analyse zeigt im Unterschied zu den Einzelanalysen für Frauen und Männer in Abb. 18, 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 19 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.

□

# 8 Anhang: Beweise

## 8.1 Die Singularwertzerlegung (Grundstruktur) von $X$

Grundlage der Korrespondenzanalyse ist die *Singular Value Decomposition* (SVD) der Matrix $X$. Generell gilt für eine beliebige reelle $m \times n$-Matrix $X$ mit dem Rang $r \leq \min(m, n)$ die Beziehung

$$X = U\Lambda^{1/2}V', \tag{106}$$

wobei $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ die Diagonalmatrix der Eigenwerte von $X'X$ bzw. $XX'$ ist. $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$; die $\sqrt{\lambda_k}$, $1 \leq k \leq r$ heißen auch *Singularwerte* (singular values). $U$ ist die $m \times r$-Matrix der Eigenvektoren von $XX'$, und $V$ ist die $n \times r$-Matrix der Eigenvektoren von $X'X$.

Die Beziehung (106) ist leicht einzusehen. Die Matrix $X$ habe den Rang $r$; dann kann man stets eine orthogonale (Teil-)Basis $L_r$ finden derart, daß

$$X = L_r V' \tag{107}$$

wobei $V$ eine $n \times r$-Matrix und $L_r$ eine $m \times r$-Matrix ist, deren Spalten die Basisvektoren sind; auf diese Weise werden die Spaltenvektoren von $X$ als Linearkombinationen der Spaltenvektoren von $L_r$ dargestellt. Der Index $r$ soll anzeigen, daß die Matrix $L_r$ den Rang $r$ hat. Dann folgt jedenfalls

$$X'X = VL_r^{'}L_rV'. \tag{108}$$

Da die Spaltenvektoren von $L_r$ als orthogonal vorausgesetzt wurden, muß $L_r^{'}L_r = \Lambda$ eine Diagonalmatrix sein, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$. Dann besagt (108) aber, daß $V$ die Matrix der Eigenvektoren von $X'X$ ist, und die $\lambda_k$, $1 \leq k \leq r$ müssen die zugehörigen Eigenwerte sein. Da $X$ den Rang $r$ hat, muß auch $X'X$ den Rang $r$ haben, und $V$ enthält gerade $r$ Eigenvektoren von $X'X$ die zu den $r$ von Null verschiedenen Eigenwerten $\lambda_k$ $k = 1, \ldots, r$ korrespondieren. Wegen der Symmetrie von $X'X$ sind die Eigenvektoren in $V$ orthogonal, darüber hinaus können sie ohne Einschränkung der Allgemeinheit als normiert vorausgesetzt werden, so daß $V'V = I$, $I$ die Einheitsmatrix. Dann folgt aus (107) aber

$$XV = L_r$$

Es sei $l_k$ der $k$-te Vektor in $L_r$, $1 \leq k \leq r$. Wegen $l_k^{'}l_k = \lambda_k$, , folgt, daß der Vektor $l_k$ die Länge $\sqrt{\lambda_k}$ hat. Die Vektoren $l_k$ in $L_r$ werden mithin durch Multiplikation von $L_r$ von rechts mit $\Lambda^{-1/2}$ normiert, so daß $XV\Lambda^{-1/2} = L_r\Lambda^{-1/2} = U$, und die Spaltenvektoren in $U$ haben die Länge 1. Also kann $L_r = U\Lambda^{-1/2}$ geschrieben werden. Daß $U$ die Matrix der Eigenvektoren von $XX'$ ist, sieht man sofort: es ist ja $XX' = U\Lambda^{1/2}V'V\Lambda^{1/2}U' = U\Lambda U'$, wegen $V'V = I$; dabei wurde vorausgesetzt, daß nur die zu Eigenwerten ungleich Null korrespondierenden Eigenvektoren von $XX'$ betrachtet werden. □

## 8.2 Satz 1

**Beweis:** Es ist

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

gemäß (21), 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 (20). Also folgt

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

□

## 8.3 Satz 2

**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 (106). 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

$$\begin{aligned}
\frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{r_k}} &= \frac{1}{\sqrt{r_i}} \frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}} - \frac{1}{\sqrt{r_k}} \frac{p_{kj} - r_k c_j}{\sqrt{r_k c_j}} \\
&= \frac{1}{\sqrt{c_j}} \left( \frac{p_{ij}}{r_i} - c_j \right) - \frac{1}{\sqrt{c_j}} \left( \frac{p_{kj}}{r_k} - c_j \right) \\
&= \frac{1}{\sqrt{c_j}} \left( \frac{p_{ij}}{r_i} - \frac{p_{ik}}{r_k} \right). \qquad (109)
\end{aligned}$$

Dann ist aber

$$\sum_{j=1}^{J} \left( \frac{x_{ij}}{\sqrt{r_i}} - \frac{x_{kj}}{\sqrt{r_k}} \right)^2 = \sum_{j=1}^{J} \frac{1}{c_j} \left( \frac{p_{ij}}{r_i} - \frac{p_{kj}}{r_k} \right)^2 = \delta_{ik}^2 \qquad (110)$$

und also

$$\begin{aligned}
\delta_{ik}^2 &= \sum_{j=1}^{J} \left( \sum_{s=1}^{r} (f_{is} - f_{ks}) v_{js} \right)^2 \\
&= \sum_{s=1}^{r} (f_{is} - f_{ks})^2 \sum_{j=1}^{J} v_{js}^2 \\
&+ 2 \sum_{j=1}^{J} \sum_{s<s'} (f_{is} - f_{ks})(f_{is'} - f_{ks'}) v_{js} v_{js'}. \qquad (111)
\end{aligned}$$

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 (111) die $\chi^2$-Distanz zwischen der $i$-ten und der $k$-ten Zeile, und rechts die entsprechende euklidische Distanz bezüglich der Koordinaten $F$. □

## 8.4 Rekonstitution

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 $D_r^{1/2}$

$$X = D_r^{1/2}FV'$$

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

$$X = D_r^{1/2}F\Lambda^{-1/2}G'D_c^{1/2} \tag{112}$$

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

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

## 8.5 Zerlegung des $\chi^2$

**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

$$y_{jj} = \frac{1}{N}\sum_{i=1}^{I}\frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N} = \chi^2_{.j}/N \tag{113}$$

Andererseits ist nach (106) $X'X = V\Lambda V'$, 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 (72) folgt.

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

$$\sum_j x_{ij}x_{il} = \sum_j\left(\frac{n_{ij} - n_{i.}n_{.j}/N}{\sqrt{n_{i.}n_{.j}/N}}\right)\left(\frac{n_{il} - n_{l.}n_{.j}/N}{\sqrt{n_{l.}n_{.j}/N}}\right) \tag{114}$$

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 (106) 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. □

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

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

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

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

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

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

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

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

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

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

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

[15] 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

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

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

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

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

# Index

Baryzentrum, 6

Chi-Quadrat-Distanz, 9

Gesamt-Inertia, 8

Massen, 6

Profil,
    Spalten-, 7
    Zeilen-, 7

Residuen, 10

Spaltenprofil, 5
Spaltenprofil, mittleres, 7

Trägheit, 8

Zeilenprofil, 5
Zeilenprofil, mittleres, 7