.

U. Mortensen

Einführung in die Hauptkomponenten- und Faktorenanalyse

.

WS 2016/2017

Fassung: 09. 01. 2016

Inhaltsverzeichnis

1 Einleitung

Man hat eine Reihe von Variablen $X_1, \ldots, X_n$ gemessen und findet, dass die meisten Korrelationen $r_{jk} = r(X_j, X_k)$ deutlich von Null verschieden sind. Dies legt die Hypothese nahe, dass die Variablen X_j, $j = 1, \ldots, n$ allen gemeinsame "latente" Variable $f_1, \ldots, f_p$, $p \leq n$ erfassen und die Korrelationen auf den gemeinsamen Effekt der L_s, $s = 1, \ldots, p$ zurückzuführen sind. Insbesondere kann man von dem Ansatz

$$X_j = a_{1j} f_1 + \cdots + a_{pj} f_p + e_j \tag{1.1}$$

ausgehen. Demnach gehen die *Faktoren* f_s in die Variable X_j mit den "Gewichten" oder "Ladungen" $a_{1j}, \ldots, a_{pj}$ ein, und e_j ist der übliche und unvermeidbare "Fehler"; e_j repräsentiert Messfehler ebenso wie spezifische latente Variable, die nur auf X_j, nicht aber auf die anderen Variablen einwirken. Dieser Ansatz wird etwas lax als *Faktorenanalyse* bezeichnet, wenn auch die Faktorenanalyse auf weiteren Annahmen beruht, die kurz spezifiziert werden sollen. Dabei wird mit $\mathbb{E}$ der Erwartungswert einer zufälligen Veränderlichen bezeichnet; dies ist das arithmetische Mittel über *alle möglichen* Realisationen der zufälligen Veränderlichen.

Annahmen:
1. $\mathbb{E}(f_k) = 0$ für alle $k = 1, \ldots, p$
2. $\mathbb{E}(\mathbf{e}) = 0$
3. $Kov(\mathbf{e}) = \mathbb{E}(\mathbf{e}\mathbf{e}') = \text{diag}(v_1^2, \ldots, v_p^2)$, d.h. die Effekte der spezifischen Faktoren und Fehler sind unkorreliert,
4. $Kov(\mathbf{f}, \mathbf{e}) = \mathbb{E}(\mathbf{e}'\mathbf{f}) = \mathbf{0}$, d.h. die Kovarianzen zwischen den Faktoren $f_1, \ldots, f_p$ und den spezifischen Faktoren und Fehlern ist gleich Null.

Der Ansatz (1.1) kann in der Form

$$\frac{1}{\sqrt{m}} X = FA' + \mathbf{e} \tag{1.2}$$

geschrieben werden: F ist die $(m \times p)$-Matrix mit den Spalten $f_1, \ldots, f_p$ (Vektoren mit den Faktorwerten als Komponenten), A ist die Matrix der Faktorladungen und $\mathbf{e}$ ist ein Vektor, dessen Komponenten "Fehler" repräsentieren.

Es sei C die Varianz-Kovarianzmatrix der in der spaltenzentrierten Datenmatrix zusammengefassten Variablen. Dann gilt das

Fundamentaltheorem der Faktorenanalyse: (Thurstone 1935)

$$C = (FA' + \mathbf{e})'(FA' + \mathbf{e}) = AA' + V, \tag{1.3}$$

V die Matrix der Stichprobenvarianzen und -kovarianzen der *spezifischen* Faktoren und Fehler. Für die Varianz σ_{jj} der j-ten Variablen folgt dann

$$\sigma_{jj} = a_{j1}^2 + \ldots + a_{jp}^2 + v_j^2, \tag{1.4}$$

v_j^2 die Varianz, die auf den für die j-ten Variable spezifischen Faktor und den Fehler in der j-ten Variablen zurückgeht. Der Anteil

$$h_j^2 = a_{j1}^2 + \ldots + a_{jp}^2 \tag{1.5}$$

an σ_{jj}^2 heißt *Kommunalität* der j-ten Variablen, – dies ist der Anteil der Varianz, der auf die allen Variablen gemeinsamen Faktoren zurückgeht.

Der Ansatz (1.1) ist auch mit Ansätzen kompatibel, die nicht als Faktorenanalyse im engeren Sinne gelten. Insbesondere sind dies die auf der Hauptachsentransformation beruhenden *Hauptkomponentenmethode* bzw. die *Hauptachsenmethnode*. Die Hauptkomponentenmethode wird zunächst vorgestellt, dann die Hauptachsenmethode. Das eigentliche faktorenanalytische Modell wird im zweiten Abschnitt besprochen.

Bei den Darstellungen wird auf das Skriptum *Kurze Einführung in die Vektor- und Matrixrechnung* Bezug genommen, in dem die notwendigen Begriffe aus der Linearen Algebra vorgestellt werden.

2 Die Hauptkomponentenmethode

Die Vorhersage von Merkmalen wie Erfolge im Studium oder Beruf, Erfolg der Teilnahme an einer Therapie etc. wird im Allgemeinen um so besser gelingen, je mehr "Symptome" oder allgemein Prädiktoren man für das zu diagnostiziernde Merkmal berücksichtigt. Der Ausdruck 'Vorhersage' ist natürlich nicht notwendig als zeitliche Vorhersage gemeint, sondern ehr als Ausdruck einer Implikation: zeigt man bestimmte "Symptome", also direkt beobachtbare Merkmale, so wird damit das Vorhandensein oder Nichtvorhandensein eines oder mehrerer anderer Merkmale impliziert. Der generelle Ansatz ist die multiple Regression

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + e, \tag{2.1}$$

wobei die X_j, $j = 1, \ldots, p$ die Prädiktorvariablen sind, $b_0, b_1, \ldots, b_p$ die Regressionsparameter, also die Gewichte, mit denen die X_j in das Urteil eingehen sind und e ist wie üblich der unvermeidbare Fehler. Dieser Ansatz entspricht dem Ansatz (1.1) mit dem Unterschied, dass in (1.1) die Prädiktoren durch die latenten Variablen f_j gegeben sind, die erst aus den Daten herausdestilliert werden müssen, während sie in (2.1) gegeben sind. Wichtig ist, dass beide Ansätze *linear* sind: Faktoren wie Prädiktoren gehen additiv in die Vorhersage von Y (oder X_j im Falle (1.1) ein. Dahinter steckt kein Naturgesetz, sondern die hoffnungsvolle Annahme, dass ein solches lineares Modell bereits hinreichend gute Vorhersagen erlaubt. Das Bemerkenswerte an dieser Annahme ist die Häufigkeit, mit der sie berechtigt ist.

Die f_j in (1.1) werden oft *latente Variable* genannt, weil sie gewissermaßen hinter den beobachteten Variablen X_j wirken und insofern 'latent' sind. In einer

konkreten Untersuchung entsprechen ihnen Vektoren $\mathbf{f}_1, \ldots, \mathbf{f}_p$, ebenso wie den Variablen $Y, X_1, \ldots, X_P$ Vektoren $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_p$ entsprechen. Die Komponenten der $\mathbf{f}_j$ repräsentieren Ausprägungen der entsprechenden latenten Variablen bei den Personen, die Komponenten der $\mathbf{x}_j$ in (2.1) sind Messungen der X_j bei verschiedenen Personen. Aus der Sicht der linearen Algebra sind die $\mathbf{f}_j$ Basisvektoren, d.h. es sind linear unabhängige Vektoren, aus denen die beobachteten Vektoren $\mathbf{x}_j$) in (1.1) als Linearkombinationen bestimmt werden können. Die lineare Algebra zeigt ebenfalls, dass es beliebig viele derartige Basen gibt. Man muß also Kriterien definieren, die auf die Wahl bestimmter Basen führen. Das übliche Kriterium ist, dass die Varianz der Komponenten von $\mathbf{f}_1$ maximal sein soll, die von $\mathbf{f}_2$ soll zweitmaximal sein, etc., die das Postulat der linearen Unabhängigkeit der $\mathbf{f}_j$ wird zu einem Postulat der Orthogonalität verschärft. Diese Forderungen führen, wie im folgenden Abschnitt beschrieben wird, sofort zu einer Lösung des Problems, latente Variable zu finden. Ob es sich um eine gut interpretierbare Lösung handelt, muß dann geprüft werden. Rotationen dieser Startlösung führen u. U. zu besser interpretierbaren Abschätzungen der $\mathbf{f}_j$.

2.1 Der SVD-Ansatz

Es sei X ein $(m \times n)$-Datenmatrix. Für X existiert stets die Singularwertzerlegung (SVD)

$$X = Q\Lambda^{1/2}T', \tag{2.2}$$

Hierin ist
1. Q die orthonormale $(m \times n)$-Matrix der Eigenvektoren von XX',
2. T die orthonormale $(n \times n)$-Matrix der Eigenvektoren von $X'X$,
3. $\Lambda^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$, und $\lambda_1, \ldots, \lambda_n$ sind die Eigenwerte von $X'X$ bzw. XX'.

Anmerkung: Symmetrische Matrizen wie $X'X$ oder XX' haben stets so viele Eigenwerte, wie sie Zeilen bzw. Spalten haben. Es sei $r = \mathrm{rg}(X)$. Dann ist $r = \mathrm{rg}(X'X) = \mathrm{rg}(XX')$ und $r \leq \min(m, n)$ ist gleich der Anzahl von von Null verschiedenen Eigenwerte. Die Aussage 3. bedeutet, dass die von Null verschiedenen Eigenwerte von XX' und $X'X$ identisch sind. Sollten also $n - r$ Eigenwerte gleich Null sein, gehen die dazu korrespondierenden Eigenvektoren aus Q bzw. T nicht in die "Vorhersage" von X ein, die n Spaltenvektoren von X werden dann durch $r < n$ Basisvektoren bestimmt.

Der Ansatz (2.2) entspricht dem Ansatz, die Hauptkomponenten, d.h. die Hauptachsen des durch $X'X$ definierten Ellipsoids als latente Variablen zu wählen. Die Abkürzung ist hierfür oft PCA (von engl. Principal Component Analysis).

Die SVD (2.2) gilt ganz allgemein für irgend eine Matrix X. Wegen der Orthonormalität von T folgt aus (2.2) nach Multiplikation mit T von rechts

$$XT = Q\Lambda^{1/2} = L, \tag{2.3}$$

und da L orthogonal ist folgt

$$L'L = \text{diag}(\|\mathbf{L}_1\|^2, \ldots, \|\mathbf{L}_n\|^2) = \Lambda^{1/2}Q'Q\Lambda^{1/2} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n), \quad (2.4)$$

d.h.

$$\|\mathbf{L}_k\|^2 = \lambda_k, \quad k = 1, \ldots, n \quad (2.5)$$

Ist X spaltenzentriert, so sind die Spaltensummen von X alle gleich Null, d.h. $\vec{1}'X = \vec{0}'$, $\vec{0}$ der n-dimensionale Nullvektor. Dann folgt aber

$$\vec{1}'X = \vec{1}'Q\Lambda^{1/2}T' = \vec{0}'\Lambda^{1/2}T' = \vec{0}', \quad (2.6)$$

weil $\vec{1}'Q = \vec{1}'Q\Lambda^{1/2} = \vec{0}'$, d.h. die Spaltensummen von Q oder L sind ebenfalls gleich Null. Die Komponenten von $\mathbf{L}_k$ sind die Koordinaen der Fälle auf den latenten Dimensionen, so dass $\|\mathbf{L}_k\|^2$ proportional zur Varianz der Koordinaten auf der k-ten latenten Dimension ist (der Proportionalitätsfaktor ist $1/m$ bzw. $1/(m-1)$, m die Anzahl der Fälle ist). (2.5) bedeutet dann, dass die Eigenwerte proportional zu diesen Varianzen sind.

Die Spaltenvektoren von T definieren die Orientierungen eines Ellipsoids, deren Hauptachsen durch die Vektoren $\sqrt{\lambda_k}\mathbf{t}_k$, $k = 1, \ldots, n$ gegeben sind, $\mathbf{t}_k$ der k-te Spaltenvektor von T. Die Komponenten der Spaltenvektoren von L sind die Koordinaten der Fälle (z.B. Personen) auf den Hauptachsen dieses Ellipsoids.

Nun sei X eine $(m \times n)$-Datenmatrix. Die Spalten repräsentieren gemessene Variablen, die Zeilen die "Objekte" oder "Fälle", an denen die Messungen ausgeführt wurden, also etwa Personen. Die Variablen können verschiedene Maßeinheiten haben, so dass es sinnvoll ist, die Messungen spaltenweise zu standardisieren. Dazu sei $\bar{x}_j$ der Mittelwert für die j-te Variable:

$$\bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} = \frac{1}{m}\vec{1}'\mathbf{x}_j, \quad (2.7)$$

wobei $\vec{1}$ ein m-dimensionaler Vektor ist, dessen Komponenten alle gleich 1 sind. Weiter sei $x_{ij} = X_{ij} - \bar{x}_j$. Dann ist

$$s_j^2 = \frac{1}{m-1}\sum_{i=1}^{m} x_{ij}^2 \quad (2.8)$$

und

$$z_{ij} = \frac{x_{ij}}{s_j}. \quad (2.9)$$

Es sei

$$S = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_n^2 \end{pmatrix}, \quad S^{-1/2} = \begin{pmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{s_n} \end{pmatrix} \quad (2.10)$$

Weiter sei $X^* = (x_{ij})$ die Matrix der $X_{ij} - \bar{x}_j$. Dann ist die Matrix der standardisierten Messwerte durch

$$Z = X^* S^{-1/2}$$

Es ist allerdings sinnvoll, die Matrix Z ein wenig umzudefinieren:

$$Z = \frac{1}{\sqrt{m-1}} X^* S^{-1/2} \tag{2.11}$$

denn die Korrelation zwischen der j-ten und der k-ten Variablen ist ja

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

wenn die z_{ij}, z_{ik} wie in (2.9) definiert sind. Definiert man Z allerdings wie in (2.11), so kann man den Faktor $1/(m-1)$ im Folgenden vernachlässigen und man erhält übersichtlichere Formeln.

Da die SVD für beliebige Matrizen X gilt, kann auch eine SVD für die Matrix Z – definiert wie in (2.11) – gefunden werden.

$$Z = Q\Lambda^{1/2} T', \tag{2.12}$$

mit

$$ZZ' = Q\Lambda Q', \quad Z'Z = T\Lambda T', \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n). \tag{2.13}$$

Hierin ist

$$R = Z'Z \tag{2.14}$$

die Matrix der Korrelationen zwischen den Variablen.

Anmerkung: Die Matrix ZZ' ist *keine* Korrelationsmatrix, da die z_{ij} eben *spaltenstandardisiert* sind. Natürlich ist es im Prinzip möglich, die Korrelationen zwischen den Fällen zu betrachten, – dazu müssen allerdings die Messwerte *zeilenstandardisiert* werden. Führt man diese Analyse aus, so erhält man eine Entsprechung der Cattellschen Q-Analyse, die auf "Typen" von Fällen (Personen) führen soll. Die Frage ist allerdings, ob Korrelationen zwischen Personen auch sinnvoll sind, denn nun übernehmen die Personen die Rolle der Variablen und die Variablen werden wie Fälle behandelt. Es zeigt sich aber, dass die SVD von spaltenzentrierten Matrizen die Frage nach eventuellen Typen automatisch mitbeantwortet, wie im Folgenden deutlich werden wird.

□

2.2 Ladungen und Faktorwerte

Es gibt zwei Möglichkeiten, (2.12) zu interpretieren:

$$Z = Q\Lambda^{1/2} T' = \begin{cases} LT', & L = Q\Lambda^{1/2} \\ QA', & A = T\Lambda^{1/2} \end{cases} \tag{2.15}$$

Im Fall $Z = LT'$ fokussiert man auf eine Repräsentation der Fälle, also etwa der Personen. T ist so definiert, dass $ZT = L$ impliziert, dass $\|\mathbf{L}_1\|^2$ maximal ist, $\|\mathbf{L}_2\|^2$ zweitmaximal, etc. Dies bedeutet, dass $\mathbf{L}_1$ maximal zwischen den Personen diskriminiert, denn die Varianz der Koordinaten L_{i1} auf der ersten Hauptachse ist maximal. Analoge Aussagen gelten für die übrigen $\mathbf{L}_k$. Man kann die Koordinaten auf $\mathbf{L}_1$ etc nun daraufhin inspizieren, ob es z.B. Gruppen von Personen gibt, die auf $\mathbf{L}_1$ voneinander getrennt werden. Bei der Diskriminanzanalyse wird dieser Aspekt im Vordergrund stehen.

Kommentar zu A: Die Spalten von A sind die längenskalierten Spalten von T. Die Spalten von T entsprechen latenten Dimensionen, die Zeilen den Variablen. Dementsprechend stehen die Spalten von A ebenfalls für die latenten Dimensionen und die Zeilen für die Variablen.

Üblicherweise ist man aber an der Struktur der Variablen interessiert, etwa wenn die Variablen Fragen in einem Persönlichkeits- oder Intelligenztest sind. In diesem Fall wird man den Fall $Z = QA'$ betrachten. Es wird zuerst einmal festgestellt, dass

$$Z'Z = AQ'QA' = AA' \tag{2.16}$$

denn Q ist ja orthonormal. A ist eine $(n \times n)$-Matrix, deren Spalten die vermuteten latenten Dimensionen (oder latenten Variablen) repräsentieren und deren Zeilen für die Variablen stehen. Die j-te Zeile von A ist

$$\tilde{\mathbf{a}}_j = (a_{j1}, a_{j2}, \ldots, a_{jn}), \quad a_{jk} = \sqrt{\lambda_k} t_{jk} \tag{2.17}$$

für $k = 1, \ldots, n$.

Definition 2.1 *Das Element a_{jk} von A ist die* Ladung *der j-ten Variablen auf der k-ten latenten Dimension, d.h. auf der k-ten Hauptachse des durch $Z'Z$ definierten Ellipsoids. Die Komponenten der Spaltenvektoren $\boldsymbol{q}_k$ von Q sind die Faktor-Scores (Faktorwerte) der Fälle auf den latenten Dimensionen*[1].

Natürlich ist a_{jk} einfach die Koordinate der j-ten Variablen auf der k-ten Hauptachse, der Ausdruck *Ladung* wurde im Zusammenhang mit der Entwicklung der Faktorenanalyse geprägt. Er soll reflektieren, dass a_{jk} den Anteil – eben die Ladung – repräsentiert, mit dem die j-te Variable die k-te "Dimension" (= latente Variable) erfasst, also das Ausmaß, mit dem die j-te Variable mit der k-ten Dimension "geladen" ist.

Der folgende Satz kann bei dem Versuch, die Ergebnisse der Analyse zu interpretieren, hilfreich sein:

Satz 2.1 *Die Ladung a_{jk} der j-ten Variablen auf der k-ten Dimension entspricht einer Korrelation zwischen der j-ten Variablen und der k-ten latenten Dimension.*

[1] Der Ausdruck 'Faktorwert' oder 'Faktorscore' wird unterschiedlich gebraucht; gelegentlich werden auch die Komponenten der Spaltenvektoren von $L = Q\Lambda^{1/2}$ als Faktorwerte oder Faktorscores bezeichnet.

Beweis: Aus $Z = QA'$ folgt $Q'Z = A'$ oder $A = Z'Q$, d.h. das Element a_{jk} von A ist gleich dem Skalarprodukt der j-ten Zeile von Z' (j-ten Spalte von Z, die die j-te Variable repräsentiert) und der k-ten Spalte $\mathbf{q}_k$ von Q:

$$a_{jk} = \mathbf{z}_j'\mathbf{q}_k = \sum_{i=1}^{m} z_{ij}q_{ik}.$$

Die q_{ik} entsprechen aber einer standardisierten Variablen, weil der Mittelwert $\bar{q}_k = 0$ ist und die Varianz gleich 1 ist, da die $\mathbf{q}_k$ ja normiert sind und der Mittelwert der Komponenten der $\mathbf{q}_k$ gleich Null ist. Also entspricht a_{jk} einem Produkt-Moment-Korrelationskoeffizienten (bis auf den Faktor $1/m$). □

Weiteren Aufschluß über die Bedeutung der Ladungen erhält man durch Diskussion der Produkte AA' und $A'A$:

Das Produkt AA': Aus $Z = QA'$ folgt

$$Z'Z = AQ'QA' = AA' \tag{2.18}$$

wegen der Orthonormalität der Vektoren von Q, so dass

$$R = Z'Z = AA', \tag{2.19}$$

R die Matrix der Korrelationen zwischen den Variablen. Der Korrelation r_{jk} zwischen der j-ten und der k-ten Variablen entspricht nach (2.19) das Skalarprodukt des j-ten Zeilenvektors von A mit dem k-ten Zeilenvektor von A (das ist der k-te Spaltenvektor von A'):

$$r_{jk} = \tilde{\mathbf{a}}_j'\tilde{\mathbf{a}}_k = \sum_{s=1}^{n} a_{js}a_{ks}, \quad j \neq k \tag{2.20}$$

und

$$r_{jj} = 1 = \sum_{s=1}^{n} a_{js}^2 \tag{2.21}$$

Dieses Resultat ist hilfreich, wenn es um die Abschätzung der Anzahl der latenten Dimensionen geht. So sei $\mathbf{v}_j$ der Vektor, der die j-te Variable repräsentiert; wie üblich wird angenommnen, dass der Anfangspunkt von $\mathbf{v}_j$ im Ursprung des Koordinatensystems liegt. (2.21) bedeutet, dass die zu den verschiedenen Variablen korrespondierenden Vektoren alle dieselbe Länge $\|\mathbf{v}_j\| = \|\tilde{\mathbf{a}}_j\| = 1$ haben, d.h. die Endpunkte liegen alle auf einer n-dimensionalen Hyperkugel. Für den Spezialfall, dass die Anzahl der Dimensionen gleich zwei ist, liegen sie alle auf einem Kreis. Dieser Sachverhalt ist hilfreich bei der Abschätzung der Anzahl der Dimensionen.

Das Produkt $A'A$: Die Elemente dieser Matrix sind die Skalarprodukte der Spaltenvektoren von A. Man findet

$$A'A = \Lambda^{1/2}T'T\Lambda^{1/2} = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \tag{2.22}$$

da ja $T'T = I_n$. Die Komponenten der Spaltenvektoren von A repräsentieren die Ladungen einer Variable auf den verschiedenen Dimensionen. Die Orthonormalität von T impliziert, dass die Skalarprodukte verschiedener Spaltenvektoren gleich Null sind. Für die Skalarprodukte dieser Vektoren mit sich selbst findet man

$$\mathbf{a}_k'\mathbf{a}_k = \|\mathbf{a}_k\|^2 = \lambda_k. \tag{2.23}$$

Die Summe der Quadrate der Ladungen der Variablen auf der k-ten Dimension ist gleich λ_k. Andererseits war ja schon $\mathbf{L}_k'\mathbf{L}_k = \|\mathbf{L}_k\|^2 = \lambda_k$, und da die $\mathbf{L}_k$ zentriert sind konnten die λ_k/m als Varianzen der Koordinaten der Fälle auf den latenten Dimensionen interpretiert werden. Die $\mathbf{a}_k$ sind nicht zentriert, λ_k entspricht also nicht der Varianz der Ladungen der Variablen auf der k-ten latenten Dimension. (2.23) bedeutet einfach, dass die Quadratsumme der Ladungen auf einer Dimension gleich der Varianz der Koordinaten der Fälle ist.

Definition 2.2 *Es sei $M = (m_{ij})$ eine $(n \times n)$-Matrix. Die Summe $sp(M) = \sum_{i=1}^n m_{ii}$ der Diagonalelemente von M heißt* Spur *von M.*

Satz 2.2 *Es ist*

$$\frac{1}{m} sp(Z'Z) = sp(AA') = \sum_{j=1}^{n} \lambda_j = n. \tag{2.24}$$

Beweis: Die Aussage $sp(Z'Z)/m = n/m$ ist klar, da ja $r_{jj} = 1$ für alle $j = 1, \ldots, n$. Dass $Z'Z = AA'$, wurde schon in (2.16) gezeigt. Wegen $A = T\Lambda^{1/2}$ hat man

$$AA' = T\Lambda^{1/2}\Lambda^{1/2}T' = T\Lambda T'.$$

Für die Diagonalelemente dieser Matrix gilt

$$\tilde{\mathbf{a}}_j'\tilde{\mathbf{a}}_j = \lambda_j\|\mathbf{t}_j\|^2 = \lambda_j,$$

wegen der Orthonormalität der $\mathbf{t}_j$, d.h. $\|\mathbf{t}_j\|^2 = 1$. Damit folgt $sp(AA') = \sum_{j=1}^n \lambda_j$.

□

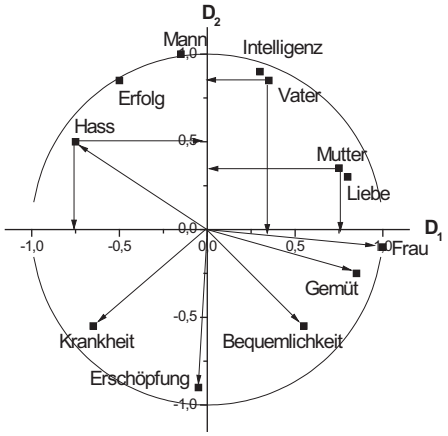
Gesamtvarianz und Varianzanteil: Da die λ_j als Varianzen bzw. als Größen, die proportional zu Varianzen sind interpretiert werden können, kann $\sum_j \lambda_j$ als Gesamtvarianz der durch die latenten Variablen erzeugten Varianzen angesehen werden. Dementsprechend ist

$$\pi_k = \frac{\lambda_k}{\sum_{j=1}^n \lambda_j} = \frac{\lambda_k}{n} \tag{2.25}$$

der Varianzanteil, den die k-te latente Dimension erzeugt oder "erklärt".

Abschätzung der Anzahl latenter Variablen: Hat man n Variablen gemessen und ist n "groß", so ist man daran interessiert, so ist man daran interessiert,

Abbildung 1: Begriffliche Stereotypen in den 50-er Jahren nach P. R. Hofstätter. Die Punkte, die die Begriffe repräsentieren, liegen nahe beim Einheitskreis, so dass die begriffliche Struktur gut durch eine 2-dimensionale "Lösung" beschrieben werden kann. D_1 repräsentiert das "weibliche Prinzip", D_2 das "männliche Prinzip". Entgegen geisteswissenschaftlichen Vorstellungen (Wellek, 1977) sind diese "Prinzipien" nicht polar, also als Gegensätze auf *einer* Dimension, angeordnet, sondern es handelt sich um voneinander unabhängige Prinzipien. In einer Person, gleich ob weiblich oder männlich, können also beide Prinzipien gleichermaßen vorhanden oder nicht vorhanden sein.



sie durch $r < n$ latente, möglichst unkorrelierte Variablen zu ersetzen, (i) um die Kovariation zwischen den beobachteten Variablenzu erklären, (ii) um zu einer möglichst "sparsamen" Beschreibung der Daten (im Sinne von Ockhams Rasierer) zu gelangen. Nun ist es aber so, dass selbst in dem eher unwahrscheinlichen Fall, dass zwei Variablen exakt dasselbe Merkmal messen, die Korrelation zwischen diesen Variablen kleiner als 1 sein wird (vergl. Beispiel 1.5 im Skript "Kurze Einführung - "). Dies bedeutet, dass im Allgemeinen die Datenmatrix "vollen Rang" hat. Die Frage ist dann, ob man die Daten nicht durch eine Matrix mit einem Rang $r < n$ approximieren kann, und wie man den Wert von r schätzen kann.

Aufschuß über den Wert von r können die Eigenwerte der Matrix $X'X$ geben:
1. Der Scree-Test (Cattell (1966)): $\lambda_k = \|\mathbf{L}_k\|^2$, $k = 1, \ldots, n$, d.h. die Eigenwerte sind proportional zur Varianz der Koordinaten $L = Q\Lambda^{1/2}$ der Fälle auf den latenten Dimensionen. Rangordnet man die Eigenwerte der Größe nach, $\lambda_1 > \lambda_2 > \cdots > \lambda_n$, so kann es sei, dass die ersten r Eigenwerte groß sind im Vergleich zu den folgeden $n - r$ Eigenwerten, so dass man argumentieren kann, dass die Unterschiede zwischen der Fällen auf den letzten $n - r$ Dimensionen nur aufgrund zufälliger Effekte zustande gekommen sind und man diese Dimensionen deswegen vernachlässigen kann. Vergl. Abbildung 2.
2. $\lambda_k = \sum_j \alpha_{jk}^2$ (vergl. Gleichung (2.23)), d.h. λ_k ist gleich der Summe der Quadrate der Ladungen der Variablen bei der k-ten latenten Dimension. Ist λ_k klein,

so unterscheiden sich die Variablen in Bezug auf die k-te Dimension nur wenig und die k-te latente Dimension repräsentiert mit großer Wahrscheinlichkeit nur zufällige Effekte.
3. Die SVD $Z = Q\Lambda^{1/2}T'$ kann in der Form

$$Z = \sqrt{\lambda_1}\mathbf{q}_1\mathbf{t}_1' + \sqrt{\lambda_2}\mathbf{q}_2\mathbf{t}_2' + \cdots + \sqrt{\lambda_n}\mathbf{q}_n\mathbf{t}_n' \tag{2.26}$$

geschrieben werden, wobei $\mathbf{t}_k\mathbf{t}_k'$ das dyadische Produkt des k-ten Spaltenvektors von Q und des k-ten Spaltenvektors $\mathbf{t}_k$ von T ist. Man sieht, dass die rechte Seite eine Approximation von Z darstellt, wenn man die letzten $n-r$ Terme der Summe vernachlässigt. Ist Z_r die Summe der ertsten r Terme, so läßt sich zeigen, dass für jeweils gewählten Wert von r die Matrix Z_r eine Approximation von Z im Sinne der Methode der Kleinsten Quadrate ist:

$$\|Z - Z_r\|^2 = \min. \tag{2.27}$$

Es ist möglich, dass die Messwerte für die n Variablen völlig zufällig verteilt sind. Die Eigenwerte $\lambda_1, \ldots, \lambda_n$ würden sich immer noch voneinander unterscheiden und dementsprechend in eine Rangordnung gebracht werden können, aber sie würden sich auch nur wenig, d.h. zufällig voneinanderunterscheiden. Cattell (1968) hat gezeigt, wie sie sich unterscheiden, wenn sie multivariat normalverteilt sind: zeichnet man einen Graphen mit dem größten Eigenwert auf Rangplatz 1, dem zweitgrößten auf Rangplatz 2, etc., so ergibt sich kein linearer, sondern ein eher an eine Exponentialfunktion erinnernder Verlauf der Größen der Eigenwerte.

Die Situation ist anders, wenn einen systematischen Einfluß von einigen, etwa r latenten Variablen gibt, und $n-r$ latente Variable nur zufällige Effekte repräsentieren. Die Abbildung Größe eines Eigenwerts versus Rangplatz ist als *Scree-Test* bekannt. Der Scree-Test ist eine im Prinzip ein verteilungsfreie Methode, die Anzahl der systematisch wirkenden latenten Dimensionen abzuschätzen.
3. Das Kaiser-Kriterium: Es seien $\lambda_1 \geq \cdots \geq \lambda_r \geq 1$, $\lambda_k < 1$ für $k > 3$. Dann ist die Anzahl zu berücksichtigender latenter Variablen gleich r (Kaiser/Dickman (1959)).

3 Die Hauptfaktorenanalyse

3.1 Faktorenmodell und SVD

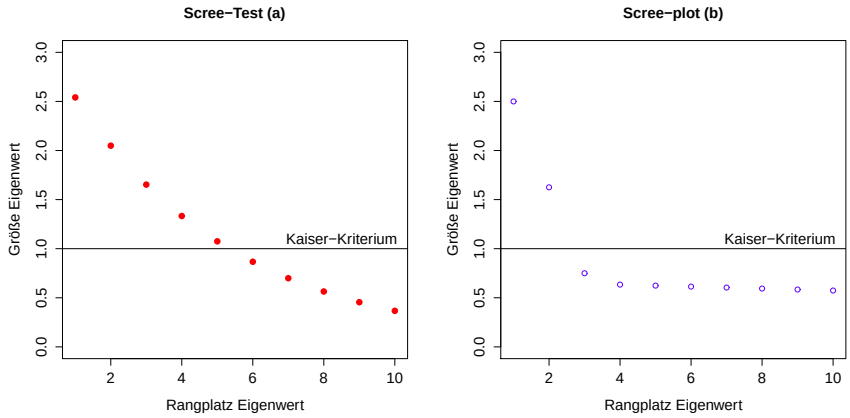
Betrachtet wird das Faktorenmodell

$$\frac{1}{\sqrt{m}}Z = FA' + \mathbf{e}, \tag{3.1}$$

V die Varianz-Kovarianzmatrix für die spezifischen Faktoren und Fehler. Dann ist

$$R = \frac{1}{m}Z'Z = AA' + V, \quad V = \operatorname{diag}(v_1^2, \ldots, v_n^2) \tag{3.2}$$

Abbildung 2: Scree-Tests – (a) zufällige Faktoren, (b) sytematische Faktoren



die Matrix der Korrelationen zwischen den Variablen. Die Kommunalität für die j-te Variable ist $h_j^2 = 1 - v_j^2$; in einer Matrix zusammengefast hat man dann $K = I - V$.

Achätzung: Die Schätzung des Modells folgt in zwei Schritten:

1. Bestimmung der Einzelvarianzen v_j^2 und damit der Kommunalitäten,
2. Es wird die *reduzierte Korrelationsmatrix* bestimmt:

$$R_h = R - V, \tag{3.3}$$

und eine Hauptachsentransformation für R_h bestimmt, d.h. es wird die Matrix $\Lambda = \text{diag}\lambda_1, \ldots, \lambda_p)$ mit der zugehörigen Matrix $T = [\mathbf{t}_1, \ldots, \mathbf{t}_p]$ der Eigenvektoren berechnet. Dann ist $A = T\Lambda^{1/2}$ und

$$AA' = T\Lambda T' = R_h, \quad R = AA' + V.$$

Die Faktorwerte sind dann wegen $ZT = F\Lambda^{1/2}$

$$F = ZT\Lambda^{-1/2} = ZR_h^{-1}A, \quad F = [\mathbf{F}_1, \ldots, \mathbf{F}_p] \tag{3.4}$$

Die $\mathbf{F}_k$, $k = 1, \ldots, p$ heißen *Hauptfaktoren*. Da die Hauptachsentransformation auf die Matrix R_h angewendet wurde, hat man nun

$$\lambda_1 + \cdots + \lambda_p = h_1^2 + \cdots + h_p^2, \tag{3.5}$$

die Summe der Eigenwerte ist nun also gleich der Summe der Kommunalitäten. Dabei werden nur Eigenwerte $\lambda_k > 0$ betrachtet. Wie bei der Hauptkomponentenmethode werden nur die ersten r Eigenwerte und damit Faktoren berücksichtigt, wobei r ebenfalls wie bei der Hauptkomponentenmethode bestimmt wird.

3.2 Interpretation und Rotation

Die SVD-Lösung, die Hauptachsen der durch die Kovarianz- oder Korrelationsmatrix definierten Ellipsoide als latente Variable zu wählen, hat zwar den Vorteil, dass die Varianz der Projektionen der Fälle auf die erste Dimension maximal ist, auf die zweite dann zweitmaximal etc, aber diese Lösung kann suboptimal sein, wenn man an einer Interpretation der Variablen-Dimensionen interessiert ist. Die Faktorenanalyse läßt für diesen Fall eine Rotation der Hauptachsen zu und stellt dafür verschiedene Kriterien bereit, nur bedeutet eine Rotation i.A., dass die neuen Koordinatenachsen nicht mehr unkorreliert sind. Die Wahl der Hauptachsen als latente Variablen oder Dimensionen bedeutet ja, dass in Bezug auf diese Dimensionen die Punktekonfiguration der Fälle achsenparallel ist, und dieser Sachverhalt ist eben gleichbedeutend mit Unkorreliertheit. So wird gelegentlich argumentiert, dass bei der Hauptkomponentenmethode nicht rotiert werden dürfe.

Diese Aussage muß nicht unwidersprochen akzeptiert werden. Denn zum einen ist die Unkorreliertheit nur ein mögliches, wenn auch wichtiges Kriterium, und zum anderen trifft das Argument, eine Rotation zerstöre die Unkorreliertheit, nur zum Teil zu.

Eine Rotation wird durch eine orthonormale Matrix bewirkt; sie werde mit S bezeichnet. Der Ausdruck $Z = Q\Lambda^{1/2}T'$ kann dann wie folgt erweitert werden:

$$Z = QSS'\Lambda^{1/2}T' \text{ bzw. } Z = Q\Lambda^{1/2}SS'T' \tag{3.6}$$

geschrieben werden. Da S orthonormal ist, gilt natürlich $S^{-1} = S'$. Mit $L = Q\Lambda^{1/2}$, $A = T\Lambda^{1/2}$ hat man dem entsprechend

$$Z = LSS'T' \text{ bzw.} \tag{3.7}$$

$$Z = QSS'A'. \tag{3.8}$$

Es gilt der folgende

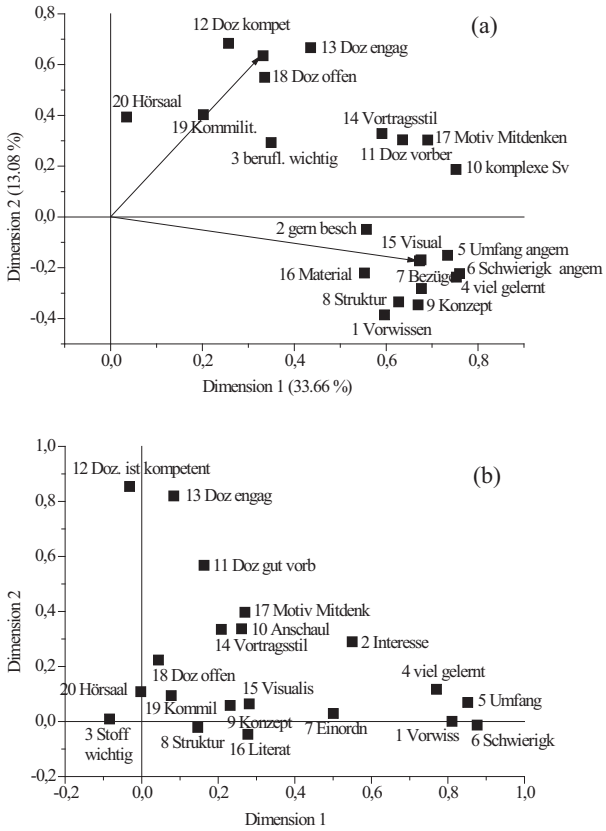
Satz 3.1 *Es gelte (3.7), und es sei $U = LS$, $V = TS$. Dann ist V orthonormal, U nicht. Gilt (3.8), so sei $U = QS$, $V = AS$. Dann ist U orthonormal, V nicht.*

Beweis: Es gelte (3.7). Es ist $U'U = S'L'LS = S'\Lambda S \neq I$. Es ist $V'V = S'T'TS = S'S = I$. Nun gelte (3.8); dann erhält man $U'U = S'Q'QS = I$, d.h. U it orthonormal, $V'V = S'A'AS = S'\Lambda S$, d.h. V ist nicht orthonormal. □.

Man kann also je nach Fokus die Koordinaten des Merkmals- oder Variablenraums beliebig orthogonal rotieren, aber die Hauptkomponenten des jeweils anderen Raumes sind nach dieser Rotation nicht mehr orthogonal.

Diese Aussagen gelten für alle orthogonalen Rotationen. Man muß sich aber für eine bestimmte entscheiden, und dafür benötigt man Kriterien. Hier soll nur eines dieser Kriterien vorgestellt werden, das

Abbildung 3: Fragen zur Evaluation einer Vorlesung

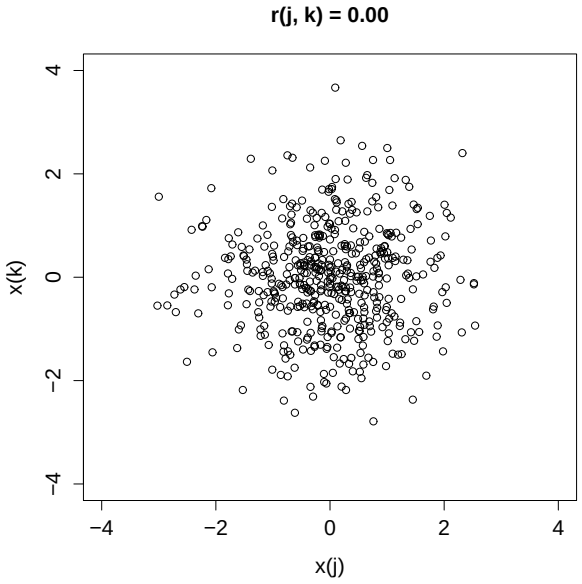


Varimax-Kriterium: Dieses Kriterium wurde von Kaiser (1958) eingeführt. Die Idee ist, die Faktoren so zu rotieren, dass auf jedem Faktor bestimmte Variablen hoch "laden" und der Rest niedrig. Kann dieses Ziel erreicht werden, so hat man eine *Einfachstruktur* (simple structure) erreicht. Um dieses Ziel zu erreichen, muß man einen diesem Begriff entsprechenden Algorithmus entwicklen, der wiederum eine formale Definition des Kriteriums erzwingt. Diese ergibt sich allerdings sofort aus dem Kriterium: für einen gegeben Faktor sollen die Ladungsquadrate bestimmter Variablen hoch, für den Rest sollen sie niedrig sein. Man kann sagen, dass die Varianz der Ladungsquadrate maximiert werden soll. Dies erklärt den namen *Varimax* für dieses Rotationskriterium.

Eine Rotation kann stets durch eine Matrix (Rotationsmatrix) definiert werden; es sei M die Matrix, die eine Varimax-Rotation bewirkt. M wird auf die Ladungsmatrix A angewendet; dementsprechend sei

$$\tilde{A} = AM \tag{3.9}$$

Abbildung 4: Stochastisch unabhängige Gauss-Variablen



die durch die Rotation erzeugte Matrix der Ladungen. Dann

$$e_1 = \sum_{k=1}^{r}\sum_{j=1}^{n}(\tilde{a}_{kj}^2 - \bar{\tilde{a}}_k)^2 \to \max_M, \quad \bar{\tilde{a}} = \frac{1}{n}\sum_{j=1}^{n}\tilde{a}_{kj} \tag{3.10}$$

Für eine gegebene Matrix M liegt der Wert von e_1 fest; die Suche nach dem Maximim bedeutet also die Suche nach der "richtigen" Matrix M. Dies geschieht iterativ; es stehen geeignete Programme zur Verfügung, die diese Suche bewerkstelligen.

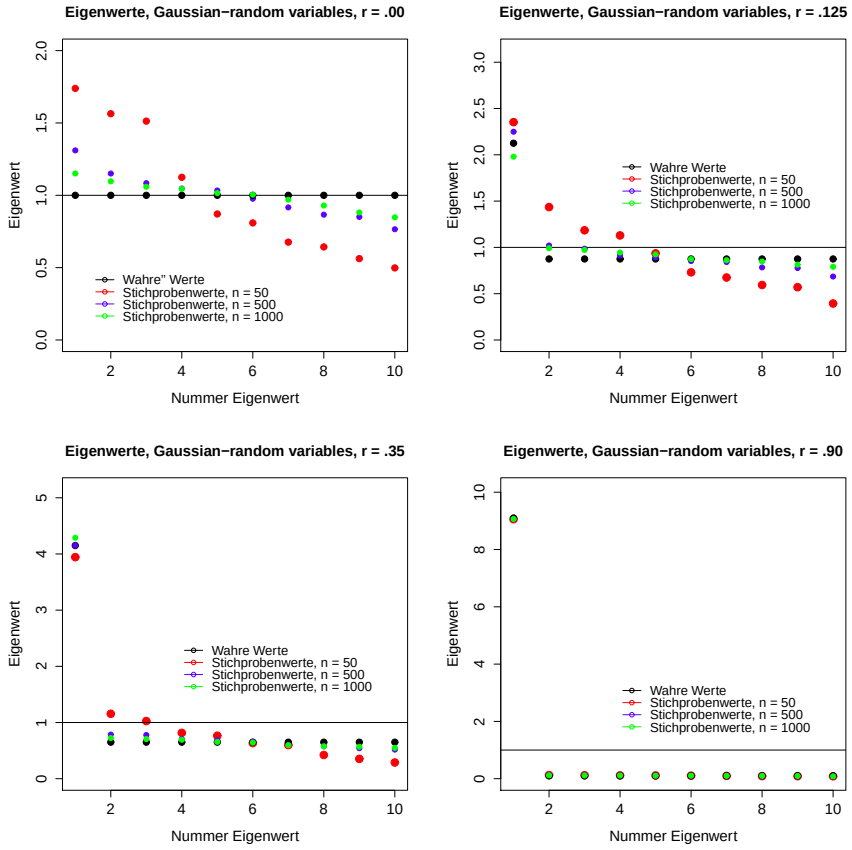
4 Statistische Aspekte

4.1 Tests der Unabhängigkeit der Variablen

Es ist von Interesse, zu untersuchen, unter welchen Bedingungen Eigenwerte größer als 1 auftreten können. So ist es möglich, dass die gemessenen Variablen statistisch unabhängig sind (s. a. Abbildung 4), so dass der wahre (im Unterschied zum Stichprobenwert der Korrelationen) Korrelationskoeffizient $\rho_{ij} = 0$ ist für alle $i \neq j$. Dies ist ein Spezialfall eines allgemeineren Falles:

$$\rho_{jk} = \begin{cases} 1, & j = k \\ \rho & j \neq k \end{cases} \tag{4.1}$$

Abbildung 5: Verlauf der Eigenwerte bei identischen Korrelationen zwischen den Variablen



d.h. ρ ist eine Konstante für für alle $i \neq j$. Es sei p die Anzahl der Variablen, so dass R eine $(n \times n)$-Matrix ist. Für die Eigenwerte von R lassen sich die folgenden Beziehungen herleiten:

$$\lambda_1 = 1 + (n-1)\rho \tag{4.2}$$

$$\lambda_2 = \lambda_3 = \cdots = \lambda_n = 1 - \rho \tag{4.3}$$

(Basilevsky (1994), p. 120). Abbildung 5 zeigt die Verteilung der Eigenwerte für verschiedene Werte von ρ. Für $\rho = 0$ (alle Variablen sind stochastisch unabhängig) sind alle Eigenwerte gleich 1, $\lambda_1 = \cdots = \lambda_n = 1$ (schwarze Punkte). Man kann dann Gauß-verteilte Zufallszahlen zur Simulation von Messwerten bestimmen, die dazu korrespondierende Korrelationsmatrix ausrechnen und die Eigenwerte berechnen. Das Ergebnis ist abhängig vom Stichprobenumfang. Wir man der Abbildung für den Fall $r = .00$ zeigt ergeben sich für alle betrachteten Stich-

probenumfänge Eigenwerte größer als 1, – nach dem Kaiserkriterium indizieren diese Eigenwerte die Existenz von latenten Variablen. Diese Schätzungen sind das Resultat von zufällig von Null abweichenden Korrelationen in den "Daten". Es sollte klar sein, dass die alleinige Betrachtung der Eigenwerte (Scree-Test) zur Abschätzung der Anzahl relevanter latenter Variablen keine valide Auskunft gibt; man wird statistische Tests der Nullhypothese "R ist eine Diagonalmatrix" suchen.

Für $\rho > 0$ ergibt sich nach (4.2) und (4.3) eine Eigenwertverteilung, für die $\lambda_1 > 1$ und $\lambda_j < 1$ für $j = 2, \ldots, n$. Aber auch für Korrelationen größer als Null ergeben für kleinere Stichprobenumfänge (wie sie in der Psychologie typisch sind) sich mehr als nur ein Eigenwert, die größer als 1 sind. Erst für größere Korrelationen entsprechen die Simulationen auch für kleinere Stichprobenumfänge Eigenwertverteilungen, die dem fehlerfreien Fall entsprechen.

Eine erste Frage in empirischen Untersuchungen ist, ob die gemessenen Variablen stochastisch unabhängig voneinander sind, d.h. ob die *wahren* Korrelationskoeffizienten für Paare gemessener Variablen gleich Null sind. Die berechneten Korrelationskoeffizienten werden gleichwohl mehr oder weniger von Null abweichen. Für eine gegebene Korrelationsmatirx stellt sich demnach die Frage, ob sie nurzufällig von einer Identitätsmatrix abweicht oder nicht. Unter der Bedingung, dass die Daten multivariat normalverteilt ($N(0, \Sigma)$-verteilt) sind, läßt sich die Verteilung der Korrelationskoeffizienten herleiten: es ist die Wishart-Verteilung, die hier aber nicht im Detail diskutiert werden soll (vergl. Johnson & Wichart (2002), p. 174), da u.a. der Begriff der Determinante vorausgesetzt wird, der in dieser Vorlesung aus Zeitgründen nicht behandelt wurde. Generell haben Tests die Form

$$H_0 : R = D, \text{ versus } H_1 : R \neq D \tag{4.4}$$

wobei D eine Diagonalmatrix ist; $R = D$ heißt dann ja, dass die Korrelationen zwischen verschiedenen Variablen gleich Null ist. Unter der Annahme, dass die Daten nultivariat normalverteilt sind wird der Likelihood-Quotient $\lambda = L(\Omega_0)/L(\Omega_1)$ betrachtet, wobei $L(\Omega_0)$ die Likelihood der Daten fr den Fall $R = D$ bezeichnet und $L(\Omega_1)$ die Likelihood der Daten für den Fall $R \neq D$. Die genaue Verteilung von λ ist nicht bekannt, aber es kann gezeigt werden, dass[2]

$$\chi^2 = -2 \log \lambda = -n \left[\log |\hat{\Sigma}| - \log \left(\prod_{i=1}^{p} \hat{\sigma}^2 \right) \right], \quad n \to \infty, \; df = \frac{p(p-1)}{2} \tag{4.5}$$

wobei $|\hat{\Sigma}|$ die Determinante der geschätzten Varianz-Kovarianz-Matrix und $\hat{\sigma}_i$ die geschätzte Varianz der der i-ten Variablen ist. Wird die Korrelationsmatrix R betrachtet, so wird λ zu

$$\lambda = |R|^{n/2}, \quad \chi^2 = -n \log |R| \tag{4.6}$$

[2]log bezeichnet den natürlichen Logarithmus (zur Basis e).

Test der Sphärizität: Dieser Test bezieht sich auf Kovarianzmatrizen und enthält den Fall von Korrelationsmatrizen als Spezialfall. Man betrachtet

$$H_0 : \Sigma = \sigma^2 I \tag{4.7}$$

Betrachtet wird der Likelihood-Ratio

$$\lambda = \frac{|\hat{\Sigma}|^{n/2}}{(\hat{\sigma}^2)^{np/2}}, \tag{4.8}$$

n die Anzahl der Fälle, p die Anzahl der Variablen. Für $n \to \infty$ hat man unter H_0 $-2 \log \lambda \to \chi^2$ mit $df = (p+2)(p-1)/2$. Keiner der in Abbildung 4.8 betrachteten Fälle widerspricht der Nullhypothese; man sieht, der alleinige Fokus auf das Kaiser-Kriterium kann zu Fehleinschätzungen der Struktur in der Datenmatrix führen.

4.2 Die Rolle der Skalen

Die stillschweigende Annahme war bisher, dass die Variablen *metrisch* sind, d.h. dass sie mindestens Intervallskalenniveau haben. Hat man Ratings erhoben, so kann es sein, dass nur ein ordinales Skalenniveau vorliegt; PCAs werden gleichwohl unter der Annahme, dass zumindest aproximativ ein Intervallskalenniveau vorliegt durchgeführt.

Generell mag man fragen, welche Restriktionen bezüglich der Skaleneigentschaft der zu analysierenden Daten gefordert werden muß. Wie Jolliffe ((1986, p. 200) ausführt, gibt es keinen Grund, zu fordern, dass Variablen von einem bestimmten Typ sein müssten, so lange die Untersuchung rein explorativ ist[3].

Was geschieht aber, wenn die die x_{ij} nur die Werte 0 oder 1 annehmen können? Dies ist z.B. dann der Fall, wenn in einem Intelligenztest die Antworten nur als "richtig" oder "falsch" beurteilt werden. Natürlich kann man auch von $(0, 1)$-Daten Korrelationen berechnen: der Produkt-Moment-Koeffizient geht dann ja über in den ϕ- oder Vierfelderkoeffizienten:

Der ϕ-Koeffizient: Es seien X und Y dichotome Variable, d.h. es gelte $X = \{0, 1\}$ und $Y = \{0, 1\}$. Es werden N Messungen gemacht, d.h es wird bei N Probanden der Wert sowohl von X als auch von Y bestimmt. Gesucht ist die Korrelation zwischen X und Y. Das Ergebnis der Messungen kann in einer Vier-

[3]When PCA is used as a desriptive Technique, there is no reason for the variables in the analysis ot be of any particular type.

feldertafel zusammengefasst werden:

	Y		
X	1	0	Σ
1	a	b	a + b
0	c	d	c + d
Σ	a + c	b + d	N

(4.9)

Für die folgenden Betrachtungen ist es nützlich, relative Häufigkeiten zu betrachten. Man hat dann

	Y		
X	1	0	Σ
1	a/N	b/N	$(a+b)/N = p_x$
0	c/N	d/N	$(c+d)/N = q_x$
Σ	$(a+c)/N = p_y$	$(b+d)/N = q_y$	1

(4.10)

wobei natürlich $q_x = 1 - p_x$, $q_y = 1 - p_y$. Wendet man die gewöhnliche Formel für den Produkt-Moment-Korrelationskoeffizienten auf die Daten in (4.9) an, so ergibt sich nach einigen Vereinfachungen der Ausdruck

$$r_{xy} = \phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{4.11}$$

Für die relativen Häufigkeiten findet man

$$\phi = \frac{(ad - bc)/N^2}{\sqrt{p_x p_y q_x q_y}} \tag{4.12}$$

Natürlich sind die Ausdrücke (4.11) und (4.12) für ϕ numerisch identisch.

Der Produkt-Moment-Korrelationskoeffizient r_{xy} für Messwerte liegt bekanntlich zwischen den Grenzen -1 und +1, d.h. $-1 \leq r_{xy} \leq 1$. Für ϕ ist dies allerdings nur unter bestimmten Randbedingungen der Fall: allgemein gilt $-1 \leq \phi_{\min} \leq \phi \leq \phi_{\max} \leq 1$:

Satz 4.1 *Für gegebene Randverteilungen p_x und p_y gilt*

$$\phi_{\min} = \max\left(-\sqrt{\frac{p_x p_y}{q_x q_y}}, -\sqrt{\frac{q_x q_y}{p_x p_y}}\right) \leq \phi \leq \min\left(\sqrt{\frac{p_x q_y}{q_x p_y}}, \sqrt{\frac{p_y q_x}{p_x q_y}}\right) = \phi_{\max} \tag{4.13}$$

und

$$\phi_{\max} = 1 \iff p_x = p_y \tag{4.14}$$

$$\phi_{\min} = -1 \iff p_x = 1 - p_y \tag{4.15}$$

$$-1 \leq \phi \leq 1 \iff p_x = p_y = .5. \tag{4.16}$$

Beweis:

Die Bestimunng von $\phi_{\max}$: Aus (4.12) folgt, dass *für gegebene Randverteilungen* p_x und p_y der Koeffizient ϕ maximal wird, wenn $bc = 0$ ist.

Der Fall $b = 0$, $c > 0$: Aus der Tabelle 4.10 liest man ab, dass nun $a/N = p_x$ und $d/N = q_y$ gilt, so dass

$$\phi_{b=0,c>0} = \frac{p_x q_y}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = \sqrt{\frac{p_x q_y}{q_x p_y}}. \tag{4.17}$$

Der Fall $b > 0$, $c = 0$: aus der Tabelle 4.10 ergeben sich $a/N = p_y$ und $d/N = q_y$, so dass

$$\phi_{b>0,c=0} = \frac{p_y q_x}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = \sqrt{\frac{p_y q_x}{p_x q_y}}. \tag{4.18}$$

Die Inspektion dieser Ausdrücke zeigt, dass $\phi_{b=0,c>0} = 1/\phi_{b>0,c=0}$. ϕ kann keinen Wert größer als 1 annehmen. Gilt also $\phi_{b>0,c=0} \leq 1$, so folgt $\phi_{b=0,c>0} \geq 1$, und umgekehrt. Also folgt

$$\phi_{\max} = \min(\phi_{b=0,c>0}, \phi_{b>0,c=0}) = \min\left(\sqrt{\frac{p_x q_y}{q_x p_y}}, \sqrt{\frac{p_y q_x}{p_x q_y}}\right) \tag{4.19}$$

Der Fall $b = c = 0$: Der Tabelle 4.10 entnimmt man, dass nun $a/N = p_x = p_y$ und $d/N = q_y = q_x$, so dass

$$\phi_{\max} = \sqrt{\frac{p_x q_y}{q_x p_y}} = \sqrt{\frac{p_x q_x}{q_x p_x}} = 1. \tag{4.20}$$

Umgekehrt impliziert $\phi_{\max} = 1$, dass $b = c = 0$, denn sonst wäre $\phi_{\max} < 1$. Dann aber folgt $a/N = p_x = p_y$ (und natürlich $d/N = q_x = q_y$).

Die Bestimmung von $\phi_{\min}$: ϕ wird minimal, wenn $ad = 0$ gilt.

Der Fall $a = 0$, $d > 0$: Aus der Tabelle 4.10 ergibt sich $b/N = p_x$ und $c/N = p_y$ und

$$\phi_{a=0,\,d>0} = -\frac{p_x p_y}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = -\sqrt{\frac{p_x p_y}{q_x q_y}}. \tag{4.21}$$

Der Fall $a > 0$, $d = 0$:, Tabelle 4.10 liefert $b/N = q_y$, $c/N = q_x$, so dass

$$\phi_{a>0,\,d=0} = -\frac{q_x q_y}{\sqrt{p_x p_y (1-p_x)(1-p_y)}} = -\sqrt{\frac{q_x q_y}{p_x p_y}}. \tag{4.22}$$

Wieder gilt $\phi_{a=0,\,d>0} = 1/\phi_{a>0,\,d=0}$, und $\phi_{a>0,\,d=0} \leq 1$ impliziert $\phi_{a=0,\,d>0} \geq 1$, und umgekehrt. Daraus folgt

$$\phi_{\min} = \max(\phi_{a=0,\,d>0},\, \phi_{a>0,\,d=0}) = \max\left(-\sqrt{\frac{p_x p_y}{q_x q_y}}, -\sqrt{\frac{q_x q_y}{p_x p_y}}\right). \tag{4.23}$$

Der Fall $a = d = 0$: Der Tabelle 4.10 entnimmt man, dass nun $b/N = p_x = q_y = 1 - p_y$, $c/N = p_y = q_x = 1 - p_x$ und

$$\phi_{a=d=0} = \phi_{\min} = -\frac{p_x p_y}{\sqrt{p_x p_y q_x q_y}} = -\frac{p_x q_y}{(1-p_x)(1-p_y)} = -\frac{p_x p_y}{p_y p_x} = -1. \quad (4.24)$$

Umgekehrt sei $\phi_{\min} = -1$. Dann muß $a = d = 0$ sein, da sonst $\phi_{\min} > -1$. Dann folgt $b/N = p_x = q_y$ und $c/N = p_y = q_x$ und

$$\frac{p_x p_y}{q_x q_y} = 1 \Rightarrow p_x p_y = (1-p_x)(1-p_y) \Rightarrow 1 = p_x + p_y.$$

$\phi_{\max} = 1$ und $\phi_{min} = -1$ bedeuten dann $p_x = p_y$ und $p_x = 1 - p_y$, woraus $p_x = p_y = .5$ folgt. □

Implikationen: Es seien $X = \{0, 1\}$, $Y = \{0, 1\}$, d.h. es sei $P(X = 1) = P_x(+) = p_x$, $P(Y = 1) = P_y(+)$, $q_x = 1 - p_x = P(X = 0)$, $q_y = 1 - p_y = P(Y = 0)$. Repräsentieren die Werte $X = 1$ und $Y = 1$ korrekte oder zustimmende Antworten bei bei Testitems, so sind p_x und p_y die Schwierigkeiten der korrespondierenden Items. Ungleiche Randverteilungen (also $p_x \neq p_y$) bedeuten, dass $|\phi| < 1$, d.h. die Schätzung der Korrelation durch ϕ ist verzerrt. Da dann $|\phi|$ kleiner ausfällt als möglich, bedeutet $p_x \neq p_y$ in faktorenanalytischen Untersuchungen eine Erhöhung der Anzahl der Faktoren. Die zusätzlichen Faktoren reflektieren nur die Ungleichheit der Grundquoten bzw. die unterschiedlichen Schwierigkeiten der Aufgaben. Den vollen Wertebereich $[-1, 1]$ hat man nur, wenn $p_x = p_y = 1/2$. Für dichotome Items bedeutet dies, dass nur Items mit einer Schwierigkeit von .5 eine unverzerrte Abschätzung der latenten Dimensionen ermöglichen.

Für den Fall $p_x \neq p_y$ ist vorgeschlagen worden, die Größe

$$\phi_0 = \frac{\phi}{\phi_{\max}} \quad (4.25)$$

zu betrachten: für ϕ_0 gilt stets $-1 \leq \phi_0 \leq 1$. Der Vorschlag ist kritisch diskutiert worden (Guilford, 1965; Davenport & El-Sanhurry, 1991). Die korrespondenzanalytische Behandlung von binären Daten könnte allerdings die sauberste Lösung dieses Problems darstellen.

4.3 Korrelierte Daten

Bisher ist angenommen worden, dass die Fälle in einer Datenmatrix stochastisch unabhängig voneinander sind: die Messungen bei einer Person sind häufig unabhängig von denen bei einer anderen Person. So kann man die Zeilen (sofern sie Fälle repräsentieren) einer Datenmatrix X vertauschen, ohne dass sich die Ergebnisse der PCA verändern.

Die Situation ist eine andere, wenn z.B. bei einer Person verschiedene Variablen als Funktionen der Zeit gemessen werden. Die $\tilde{\mathbf{x}}_i$ sind dann Vektoren, deren

Komponenten Werte zu Zeitpunkten sind: $x_{ij} = x_j(t_i)$, d.h. der i-te Messwert der j-ten Variablen ist ein Messwert zum Zeitpunkt t_i. Man kann dann die Faktorwerte $L_{ik} = L_k(t_i)$ als Werte einer latenten Funktion zur Zeit t_i interpretieren; die gemessenen Verläufe der j-ten Variablen werden dann als gewogene Summen von latenten Funktionen dargestellt. Dem entspricht dem aus der Mathematik bekannten Sachverhalt, dass viele Funktionen als additive Überlagerung von *Basisfunktionen* repräsentiert werden können.

Insbesondere kann es dabei sinnvoll sein, die Korrelationen der Variablenwerte zwischen verschiedenen Zeitpunkten in Rechnung zu stellen. Im Prinzip betrachtet man dann die n Variablen als multivariaten stochastischen Prozess, dessen Grundstruktur beschrieben werden soll; man denke etwa an fMRI-Untersuchungen, bei denen Aktivitäten an verschiedenen Positionen über die Zeit betrachtet werden. Ein wichtiger Typ von PCA-Analyse ist die Karhunen-Loève-Spektralzerlegung (Basilevsky(1994), p. 445), auf die hier nur hingewiesen werden kann.

5 Korrespondenzanalyse

Kontingenztabellen kommen in der empirischen Praxis häufig vor, sind bei größeren Zeilen- und Spaltenzahlen unübersichtlich und schwer zu intrepretieren, und das χ^2 bzw. der Kontingenzkoeffizient $\sqrt{\chi^2/N}$, N die Gesamtzahl der Beobachtungen, ist häufig die einzige Statistik, die auf die Daten angewendet wird; sie erlaubt eine Aussage über die Wahrscheinlichkeit der Existenz von Abhängigkeiten zwischen Zeilen- uns Spaltenkategorien. Ein Beispiel ist die berühmte Tabelle von Westphal zur Kretschmerschen Typentheorie: Im Prinzip ist die Ta-

Tabelle 1: Typen nach Kretschmer

	Erkrankung			
Typ	man-dep	Epilepsie	Schizophrenie	Σ
pyknisch	879	83	717	1679
athletisch	91	435	884	1410
leptosom	201	378	2636	3271
dysplastisch	15	444	550	1009
atypisch	115	165	450	73
Σ	1361	1505	5233	8099

belle eine Datenmatrix X mit m Zeilen und n Spalten und man könnte auf die Idee kommen, eine SVD auf X anzuwenden, d.h. eine PCA durchzuführen. Man erhielte dann latente Dimensionen, die die eventuellen Abhängigkeiten zwischen den Zeilen und Spalten "erklären" könnten. Man hätte etwa

$$n_{ij} = a_{j1}L_{i1} + \cdots + a_{jp}L_{ip} \tag{5.1}$$

n_{ij} die Häufigkeit in der (i, j)-ten Zelle der Tabelle, p die Anzahl der latenten Variablen. Rechnen läßt sich eine derartige Analyse, die Frage ist, was die Ergebnisse bedeuten. Die rechte Seite beschreibt ein additives Modell für die Zusammensetzung der Häufigkeiten n_{ij}, aber Häufigkeiten lassen sich zum Beispiel nicht sinnvoll in eine "wahre" Häufigkeit $\tilde{n}_{ij}$ plus einem "Fehler" zerlegen, etwa in der Art $n_{ij} = \tilde{n}_{ij} + e_{ij}$. Die Häufigkeiten geben an, wie häufig eine bestimmte Kategorienkombination in einer Stichprobe aufgetreten ist. Die rechte Seite von (5.1) sollte sich dann auf eine Zerlegung der Kategorien in additive Komponenten beziehen, aber wie diese Zerlegung aussehen soll, ist unklar.

Ein alternativer Ansatz geht von der Formel für den χ^2-Test aus:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(n_{ij} - n_{i+}n_{+j}/N)^2}{n_{i+}n_{+j}/N} \tag{5.2}$$

n_{i+} die Zeilensummen und n_{+j} die Spaltensummen der Tabelle; $n_{i+}n_{+j}/N$ sind die unter H_0 erwarteten Häufigkeiten.

Es sei nun

$$x_{ij} = \frac{n_{ij} - n_{i+}n_{+j}/N}{\sqrt{n_{i+}n_{+j}/N}} \tag{5.3}$$

x_{ij} entspricht einer standardisierten Variablen. Die Differenz $n_{ij} - n_{i+}n_{+j}/N$ repräsentiert gewissermaßen den von Zufälligkeiten (wie sie von den unter H_0 erwarteten Häufigkeiten $n_{i+}n_{+j}/N$ abgebildet werden) bereinigten systematischen Beziehungen zwischen der i-ten Zeilenkategorie und er j-ten Spaltenkategorie. Es läßt sich nun zeigen[4], dass die SVD der Matrix $X)(x_{ij})$ auf latente Variablen führt, die statistisch unabhängige Komponenten der Kategorien repräsentieren, die wiederum einer additiven Zerlegung des χ^2 entsprechen, – analog zu der additiven Zerlegung der Gesamtvarianz der Daten bei einer PCA einer Matrix von Messwerten.

Die SVD von X sei durch

$$X = U_0 \Lambda^{1/2} V_0' \tag{5.4}$$

gegeben: U_0 ist die $(m \times p)$-Matrix der Eigenvektoren von XX', V_0 ist die $(n \times p$-Matrix der Eigenvektoren von $X'X$, und $\Lambda^{1/2}$ ist die Diagonalmatrix der von Null verschiedenen Eigenwerte von $X'X$ bzw. XX', und p ist die Anzahl der latenten Dimensionen.

Man könnte nun, wie bei einer PCA üblich, sich entweder auf die Zeilen- oder auf die Spaltenkategorien fokussieren, d.h. für die Zeilenkategorien $L = U_0\Lambda^{1/2}$ oder für die Spaltenkategorien $A = V_0\Lambda^{1/2}$ zu berechnen. Ein derartiges Vorgehen beruht aber auf der Asymmetrie zwischen Fällen und Variablen bei einer gewöhnlichen Datenmatrix X, bei der die x_{ij} Messwerte von Variablen bei verschiedenen Fällen sind: die Fälle sind eben Personen oder Objekte, an denen

[4] http://www.uwe-mortensen.de/caneu1c.pdf

MEssungen der Variablen vorgenommen werden. Eine Kontigenztabelle dagegen ist symmetrisch: sowohl die Zeilen wie auch die Spalten stehen für Kategorien, und welche Kategorien man als Zeilen- und welche man als Spaltenkategorien anschreibt ist im Prinzip beliebig. Die Skalierung der Zeilen- und Spaltenkategorien sollte dementspreched in einer symmetrischen Art und Weise geschehen. Betrachtet man bei einer gewöhnlichen, messwertebasierten PCA die Koordinaten L der Fälle, so werden die Koordinaten so gewählt, dass die Skalenwerte auf der ersten latenten Dimension eine maximale Varianz haben, etc., die verschiedenen latenten Dimensionen erklären jeweils Anteile an der Gesamtvarianz. Bei der SVD einer Häufigkeitstabelle sollen die latenten Dimensionen Anteile am Gesamt-χ^2 erklären, und auch hier sollte die Symmetrie von Spalten- und Zeilenkategorien erhalten bleiben. Die in der folgenden Gleichung eingeführten Koordinaten für die Zeilen- und Spaltenkategorien erfüllen diese Bedingung

$$f_{ik} = u_{ik}\sqrt{\lambda_k / r_i} \tag{5.5}$$

$$g_{jk} = v_{jk}\sqrt{\lambda_k / c_j}. \tag{5.6}$$

Hierin sind die u_{ik} das Element in der i-ten Zeile und k-ten Spalte von U_0, wobei $k = 1, \ldots, p$ die latenten Dimensionen indiziert, und v_{jk} ist das ELement in der j-ten Zeile und k-ten Spalte von V_0, j indiziert die j-te Spaltenkategorie und k wieder die k-te Dimension. $r_i = n_{i+}/N$ die relative i-te Zeilensumme (r_i von englisch *row*), und $c_j = n_{+j}/N$ die relative j-te Spaltensumme ($c_j$von englisch *column*). Offenbar geht $\lambda_k^{1/2}$ sowohl in die Koordinaten der Zeilen- wie auch der Spaltenkategorien ein: dies reflektiert die oben angesprochene Symmetrie von Zeilen- und Spaltenkategorien.

Die f_{ik} und g_{jk} sind Skalenwerte der Kategorien auf den $k = 1, \ldots, p$ latenten Dimensionen. Ihre Definition ist so gewählt, dass die euklidischen Distanzen zwischen den Kategorien dem Beitrag entsprechen, den ihr jeweiliger Unterschied zum Gesamt-χ^2 ausmacht: die Distanzen

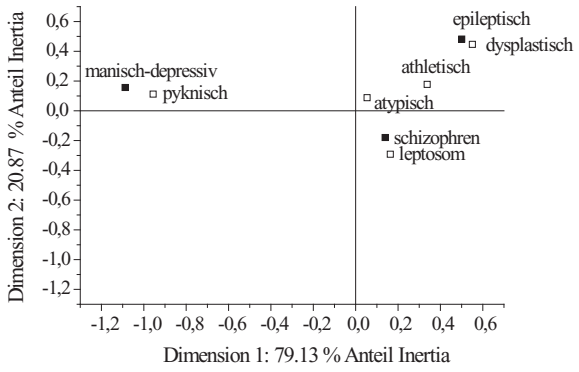
$$d_{ii'} = \sqrt{\sum_{k=1}^{p} (f_{ik} - f_{i'k})^2} \tag{5.7}$$

$$d_{jj'} = \sqrt{\sum_{k=1}^{p} (g_{jk} - g_{j'k})^2} \tag{5.8}$$

zwischen den Zeilenkategorien einerseits und den Spaltenkategorien andererseits heißen dementsprechend χ^2-*Distanzen*. Distanzen d_{ij} zwischen den Zeilenkategorien und den Spaltenkategorien sind *nicht* erklärt; bei der Interpretation der graphischen Repräsentation der Kategorien im *Biplot* (s. unten) muß man sich also vor Fehlinterpretationen hüten. Die Beziehung zwischen einer Zeilen- und dien Spaltenkategorie läßt sich durch das Skalarprodukt der Vektoren $\mathbf{f}_i = (f_{i1}, \ldots, f_{ip})'$ und $\mathbf{g}_j = (g_{j1}, \ldots, g_{jp})'$ ausdrücken:

$$\mathbf{f}_i'\mathbf{g} = \|\mathbf{f}_i\| \|\mathbf{g}_j\| \cos\theta_{ij}. \tag{5.9}$$

Abbildung 6: Kretschmers Typen



Für gegebene $\mathbf{f}_i$ und $\mathbf{g}_j$ wird $\mathbf{f}_i'\mathbf{g}_j$ maximal, wenn $\theta_{ij} = 0$; dief_{ik} und g_{jk} sind dann proportional zueinander, d.h. die beiden Kategorien sind bis auf einen Proporitonalitätsfaktor in der gleichen Weise aus den latenten Dimensionen zusammengesetzt.

Biplot: Die Ergebnisse einer Korrespondenzanalyse werden am besten graphisch repräsentiert. Man erhält zunächst zwei Abbildungen; in einer werden die Zeilenkategorien durch Punkte mit den Koordinaten f_{ik} abgebildet, in der anderen werden die Spaltenkategorien mit den Koordinaten g_{jk} abgebildet. Interessanterweise werden im Allgemeinen nur zwei latente Dimensionen benötigt, also $k = 1, 2$. Da der Skalierungsfaktor $\lambda_k^{1/2}$ in die Definition sowohl der f_{ik} wie auch der g_{jk} eingeht, sind die beiden Abbildungen miteinander kompatibel und können zu einer einzigen Abbildung zusammengefasst werden: dies ist der *Biplot*. Für jede Dimension (Koordinatenachse) kann angegeben, wieviel des Gesamt-χ^2 der Tabelle durch die Dimension erklärt wird.

Beispiele: Zunächst die Analyse von Westphals (1931) Tabelle: Die beiden latenten Dimensionen erklären das Gesamt-χ^2 der Tabelle zu 100%. Die Position der Kategore "atypisch" nahe beim Ursprung des Koordinatensystems entpsricht dem Sachverhalt, dass 'atypisch' gewissermaßen den Durchschnitt aller Körperbautypen repräsentiert. Im Übrigen zeigt der Biplot eine nahzu perfekte Entsprechung von Daten und Theorie, Kretschmer hätte sich sicher gefreut, hätte es damals schon die Korrespondenzanalyse gegeben. Die Möglichkeit, dass Westphal seinerzeit die Klassifikation von Körperbau und Erkrankung nach Maßgabe der Kretschmerschen Theorie vorgenommen haben könnte, muß aber dem Betgrachter bewußt sein: ein als Epileptiker diagnostizierter Kranker wird möglicherweise eher als Dysplastiker oder Athlet beurteilt, als wenn er als 'manisch-depressiv' beurteilt worden wäre, ein als 'schizophren' diagnostizierter Patient kann eher als 'leptosom' klassifiziert werden – zumal diese Erkrankung in eher jüngeren Jahren auftritt – als wenn er als 'epileptisch' diagnostiziert worden wäre, etc. Selbst wenn

es so ist: die Korrespondenzanalyse extrahiert auch in diesem Fall die Strukturen sehr schön aus den Daten, denen man eine so gute Übereinstimmung mit der Theorie nicht so ohne Weiteres ansehen würde. Einer der Gründe dafür ist die implizite Berücksichtigung der unterschiedlichen Häufigkeiten, mit denen die Erkrankungen bzw. Körperbautypen in der Population auftreten.

Zeitliche Trends: Die Korrespondenzanalyse kann auch auf die Entschlüsselung von zeitlichen Trends eingesetzt werden. Die Tabelle 2 enthält die Häufigkeiten,

Tabelle 2: 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

mit denen in den verschiedenen Fächern zu verschiedenen Jahren promoviert wurde. Man entnimmt der Tabelle, dass in den verschiedenen Fächern unterschielich häufig promoviert wurde, und dass sich die Häufigkeiten mit den Jahreszahlen verändern. Der Biplot liefert eine gute Übersicht über die Art der Entwicklungen in den einzelnen Fächern. Der Biplot zeigt einen klaren Trend von der Physik über Ingenieursstudiengänge zu den Humanwissenschaften. Wieder erklären die beiden latenten Dimensionen nahezu 100% des Gesamt-χ^2 der Tabelle.

Selbstmorde: Diese Tabelle ist 3-dimensional, und insofern für eine Korrespondenzanalyse nicht geeignet, – es sei denn, man verwandelt sie in eine 2-dimensionale Tabelle, indem man die Tabellen für männliche und weibliche Suicid-Fälle nebeneinander schreibt. Die so entstehende Tabelle hat die Altersgruppen als Zeilenkategorien, und die Methoden – einmal für männliche, einmal für weibliche Fälle – als Spaltenkategorien.

Die Abbildung 8 zeigt die Häufigkeiten von Selbstmorden, nach Geschlecht getrennt, für weibliche und männliche Fälle. Offenbar sind insbesondre die Jahre zwischen dem 20-ten und dem 50-ten Lebensjahr für Männer schwerer zu ertragen als für Frauen, die sich dem inspizierenden Studium der Tabellen nicht so leicht enthüllen. Der Biplot verweist suicfreqgesauch auf weitere qualitative Unterschie-

Abbildung 7: Biplot Doktorate - Jahre

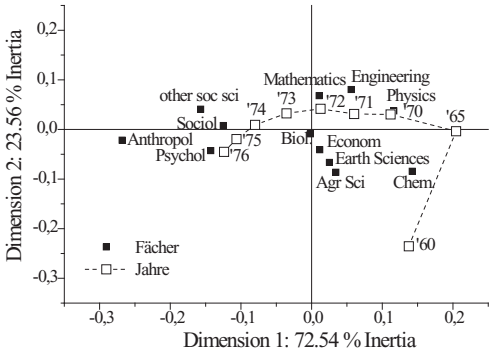
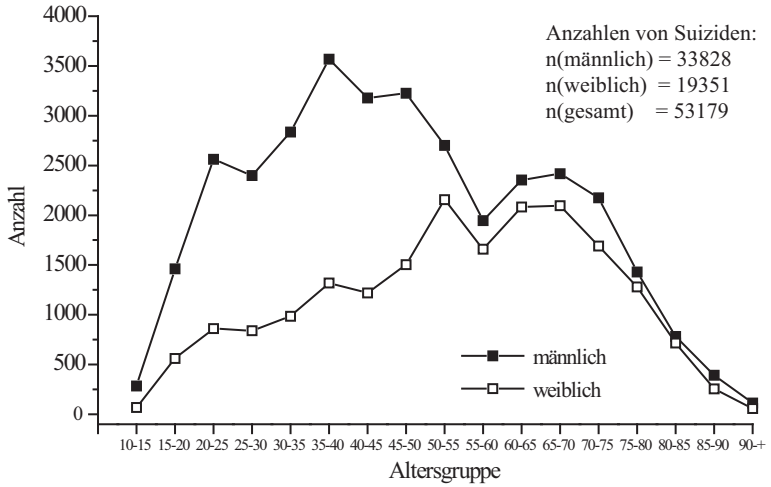


Abbildung 8: Häufigkeitsverteilung der Selbstmorde, aggregiert über Methoden



de zwischen Mnnern und Frauen: Offenbar sind die Abhängigkeiten zwischen Alter und Methode bei den Frauen nicht so klar strukturiert wie bei den Männern. Das Resultat ist kein Artefakt des Aufbaus der Tabelle, in der die Daten für Männer und Frauen nebeneinander auftreten. Gäbe es keinen Unterschied, so würden die Punkte für die Methoden (Männer) und die Methoden (Frauen) nahe beieinander positioniert (die Korrespondenzanalyse "wei?? ja nicht, dass z.B. 'Erhängen' bei den Männern und den Frauen dieselbe Methode ist, – für die Korrespondenzanalye sind es zwei verschiedene Methoden. Analoge Betrachtungen gelten für die Altersgruppen.

Die Korrespondenzanalyse ist hier im Wesentlichen nur illustriert worden, – eine ausgiebige Darstellung der Methode und der Analyse der Daten der Beispiele

Tabelle 3: Selbstmorde in Westdeutschland 1974-1977 (Heuer, 1979

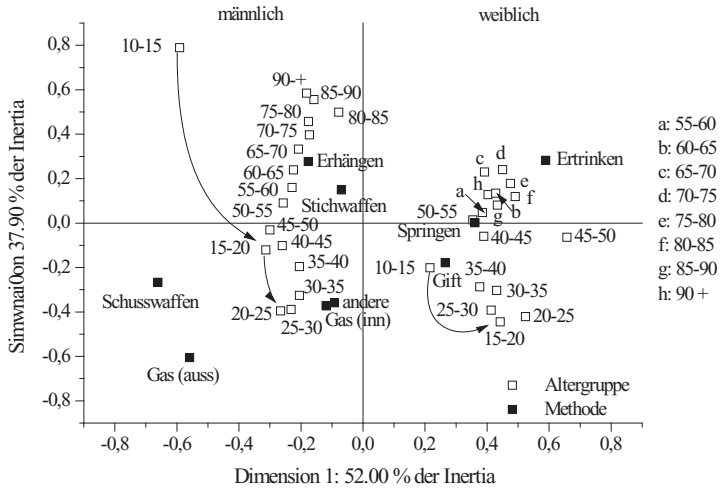
Alter/männl	Materie	Gas (h)	Gas (a)	Hängen	Ertrinken	Schußw.	Stichw.	Springen	Andere
10-15	4	0	0	247	1	17	1	6	9
15-20	348	7	67	578	22	179	11	74	175
20-25	808	32	229	699	44	316	35	109	289
25-30	789	26	243	648	52	268	38	109	226
30-35	916	17	257	825	74	291	52	123	281
35-40	1118	27	313	1278	87	293	49	134	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

wird im Skript 'Einführung in die Korrespondenzanalyse'[5] gegeben.

xxx

[5]http://www.uwe-mortensen.de/caneu1c.pdf

Abbildung 9: Biplot Selbstmorde: Methode, Altersgruppen und Geschlecht



6 Anhang: Der ϕ-Koeffizient

6.1 Herleitung des ϕ-Koeffizienten

Die Produkt-Moment-Korrelation ist durch

$$r_{xy} = \frac{Kov(X,Y)}{s_x s_y} = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{\frac{1}{n}\sum_i x_i y_i - \bar{x}\cdot\bar{y}}{s_x s_y}. \tag{6.1}$$

gegeben. Zunächst gilt

$$N = a + b + c + d. \tag{6.2}$$

Die x_i, y_i nehmen nun nur die Werte 1 oder 0 an; dann ist

$$\sum_i x_i y_i = a(1\cdot 1) + b(1\cdot 0) + c(0\cdot 1) + d(0\cdot 0) = a.$$

Weiter ist

$$\bar{x} = \frac{a+b}{N}, \quad \bar{y} = \frac{a+c}{N},$$

so dass

$$\begin{aligned} Kov(x,y) &= \frac{a}{N} - \frac{(a+b)}{N}\frac{(a+c)}{N} = \frac{Na - (a+b)(a+c)}{N^2} \\ &= \frac{Na - a^2 - ac - ab - bc}{N^2} = \frac{a(N-a) - ac - ab - bc}{N^2} \\ &= \frac{a(N-a-b-c) - bc}{N^2} = \frac{ad - bc}{N^2} \end{aligned} \tag{6.3}$$

wegen (6.2). Für s_x^2 findet man

$$s_x^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 = \frac{1}{N}\sum_i x_i^2 - \bar{x}^2,$$

so dass

$$s_x^2 = \frac{(a+b)1^2}{N} - \frac{(a+b)^2}{N^2} = \frac{a+b}{N} - \frac{(a+b)^2}{N^2} = \frac{a+b}{N}\left(1 - \frac{a+b}{N}\right) = p_x(1-p_x)$$

Analog

$$s_y^2 = \frac{(a+c)}{N}\left(1 - \frac{a+c}{N}\right) = p_y(1-p_y).$$

Dann ist einerseits

$$s_x s_y = \sqrt{p_x(1-p_x)p_y(1-p_y)}, \tag{6.4}$$

so dass wegen (6.3)

$$\phi = \frac{ad-bc}{N^2\sqrt{p_x(1-p_x)p_y(1-p_y)}}, \tag{6.5}$$

und andererseits

$$\begin{aligned} s_x s_y &= \sqrt{\frac{(a+b)}{N}\left(1 - \frac{(a+b)}{N}\right)\frac{(a+c)}{N}\left(1 - \frac{(a+c)}{N}\right)} \\ &= \sqrt{\frac{(a+b)}{N}\left(\frac{N-a-b}{N}\right)\frac{(a+c)}{N}\left(\frac{N-a-c}{N}\right)} \end{aligned}$$

Da $N = a+b+c+d$, folgt $N-a-b = a+b+c+d-a-b = c+d$, und $N-a-c = a+b+c+d-a-c = b+d$, so dass

$$s_x s_y = \sqrt{\frac{(a+b)(a+c)(b+c)(c+s)}{N^4}} = \frac{1}{N^2}\sqrt{(a+b)(a+c)(b+d)(c+d)},$$

unn man erhält (s. (6.3))

$$\phi = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}. \tag{6.6}$$

Literatur

[1] Basilevsky, A.: Statistical Factor Analysis and related methods. John Wiley & Sons, New York 1994

[2] Davenport E.C., El-Sunhurry, N.A. (1991) *Educational and Psychological Measurement*, 51, 821 – 828

[3] Eckart, C., Young, G. (1936), The approximation of one matrix by another of lower rank. *Psychometrika*, 1 (3): 211–8. doi:10.1007/BF02288367.

[4] Dollard, J., Doob, L., Miller, N., Mowrer, O., & Sears, R. (1939). Frustration and aggression. New Haven, CT: Yale University Press.

[5] Ferguson, G. A. (1941) The factorial interpretation of test difficulty. *Psychometrika*, (6(5), 323 – 325

[6] Fischer, G.: Lineare Algebra. Friedr. Vieweg & Sohn Verlagsgesellschaft, Braunschweig Wiesbaden 1997

[7] Guilford, J.P. (1965) The minimal phi coefficient and the maximal phi. *Educational and Psychological Measurement*, Vol. XXV, 1, 3 – 8

[8] Haken, H.: Synergetics. Springer Verlag, Berline, Heidelberg

[9] Jolliffe, I. T.: Principal Component Analysis. Springer-Verlag, New York, Berlin, Heidelberg Tokyo 1983

[10] Kaiser, H. F. (1958) The Varimax Criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187 – 200

[11] Lorenz, F.: Lineare Algebra I, II. Mannheim, 1988

Index