Universitäts-Rechenzentrum Trier

AWS.SPSS.7

Trier, den 28.4.1997

Korrespondenzanalyse mit SPSS

Herausgeber: Universitäts-Rechenzentrum Trier
Universitätsring 15
D-54286 Trier
Tel.: (0651) 201-3417, Fax.: (0651) 3921
Leiter: Prof. Dr.-Ing. Manfred Paul
Autor: Bernhard Baltes-Götz
Mail: baltes@uni-trier.de
Copyright © 1997; URT

Vorwort

Das Manuskript beschreibt die im SPSS-Zusatzmodul **Categories** verfügbare Korrespondenzanalyse. Hier wird versucht, die Zeilen- oder Spaltenprofile einer zweidimensionalen Korrespondenztabelle (z.B. Kontingenztabelle) so als Punkte eines möglichst einfachen geometrischen Raumes (bevorzugt: die Ebene) darzustellen, dass die euklidischen Distanzen zwischen den Punkten annähernd gewissen Profilähnlichkeiten entsprechen.

Als Software kommt SPSS 6.1 für Windows zum Einsatz, jedoch können praktisch alle vorgestellten Verfahren auch mit jüngeren SPSS-Versionen unter Windows, MacOS oder Linux realisiert werden.

Das Manuskript ist als PDF-Dokument zusammen mit den im Kurs benutzen Dateien auf dem Webserver der Universität Trier von der Startseite (http://www.uni-trier.de/) ausgehend folgendermaßen zu finden:

Rechenzentrum > Studierende > EDV-Dokumentationen >
Statistik > Korrespondenzanalyse mit SPSS

Hinweise auf Unzulänglichkeiten im Manuskript werden mit Dank entgegen genommen

1 Einleitung

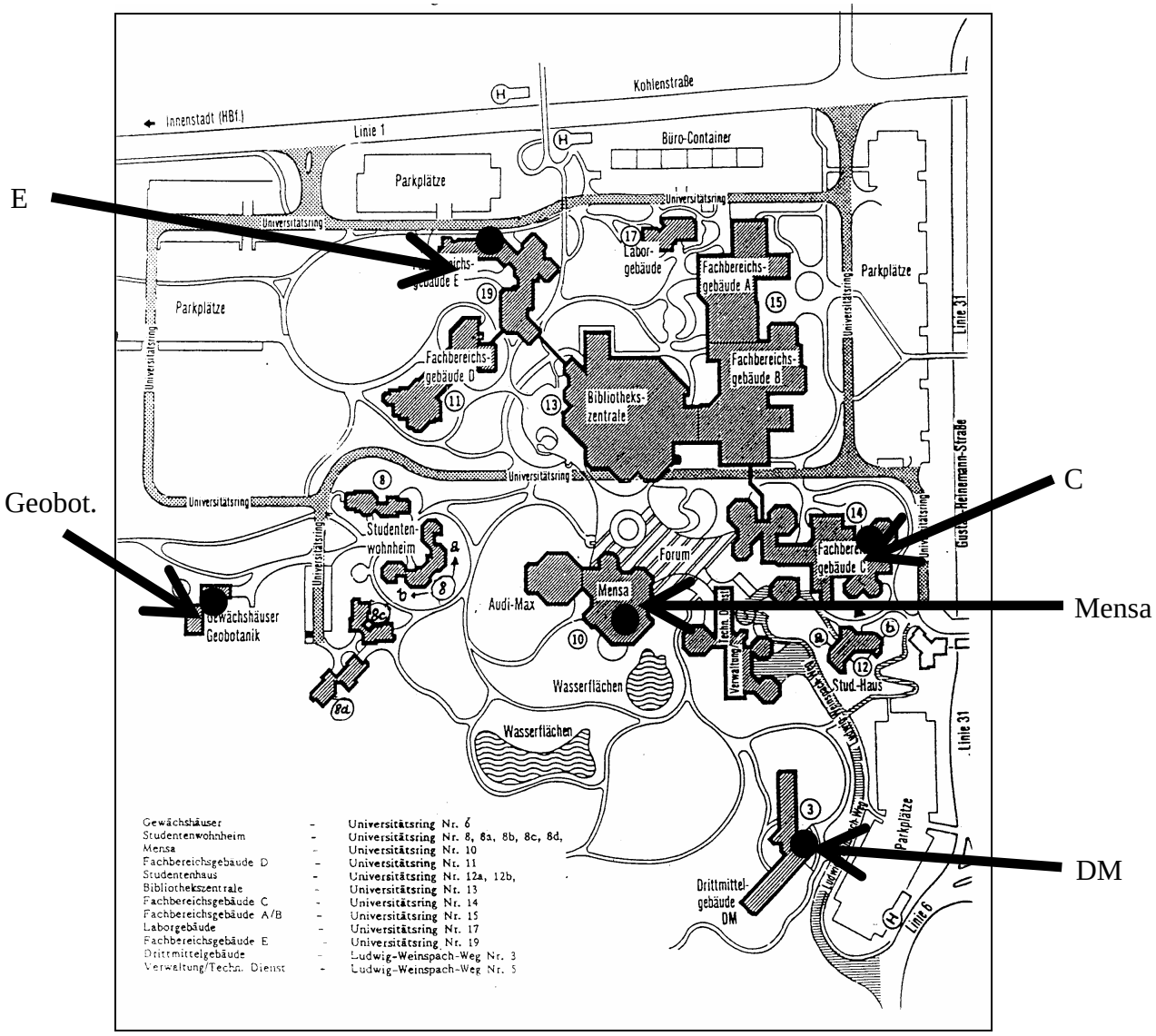
Das Ziel der Korrespondenzanalyse (KA) besteht darin, die Zeilen- oder Spaltenprofile in einer zweidimensionalen Korrespondenztabelle derart als Punkte eines einfachen geometrischen Raumes darzustellen, daß die euklidischen Distanzen (= Entfernungen) zwischen den Punkten annähernd gewissen Profilähnlichkeitsmaßen entsprechen. Wesentliches Endprodukt der Analyse ist die *grafische* Veranschaulichung von Profilähnlichkeiten z.B. anhand der Distanzen zwischen Punkten in der Ebene des $\mathbb{R}^2$. Dort sind dann Korrespondenzen zwischen den Zeilenprofilen bzw. zwischen den Spaltenprofilen oder auch zwischen einer Zeilen- und eine Spaltenkategorie direkt ablesbar.
SPSS bietet im Zusatzmodul CATEGORIES die KA-Prozedur ANACOR an, die an der Holländischen Universität Leiden entwickelt wurde (siehe Gifi 1990).
Häufig geht die Korrespondenzanalyse von einer Häufigkeitstabelle aus (siehe Beispiel 2), doch kann auch eine Tabelle verarbeitet werden, deren Zellen (*i*,*j*) ein beliebiges Maß für die Korrespondenz zwischen der *i*-ten Zeile und der *j*-ten Spalte enthalten.

Beispiel 1: Der Trierer Uni-Campus

Es ist vielleicht lehrreich, die Korrespondenzanalyse einmal auf Zeilenprofile anzuwenden, die tatsächlich einen zweidimensionalen Entstehungshintergrund haben. Dazu wurden aus einer Karte der Trierer Uni-Campusanlage die Entfernungen von fünf Gebäuden zu den Ecken der Karte gemessen. Durch eine Lineartransformation (50 - Entfernung) wurde dafür gesorgt, daß positive Korrespondenzwerte zwischen den Kategorien der Zeilenvariable **Gebäude** und den Kategorien der Spaltenvariable **Ecke** entstanden. Es resultierte die folgende Korrespondenztabelle:

	li. ob.	re. ob.	re. unt.	li. unt.
E	35,50	33,80	28,00	29,30
Geobot.	30,10	23,30	29,30	39,60
Mensa	27,00	28,70	36,50	33,80
C	24,10	31,10	38,50	28,90
DM	20,00	24,50	43,60	32,80

Sie finden diese Tabelle in der SPSS-Datendatei **CAMPUS.SAV** an der im Vorwort vereinbarten Stelle.
Es ist zu hoffen, daß die Korrespondenzanalyse eine zufriedenstellende zweidimensionale Repräsentation der vierdimensionalen Zeilenprofile ermittelt, welche die räumlichen Verhältnisse näherungsweise wiedergibt. Hier betreiben wir keine empirische Forschung, sondern studieren in einer wohlbekannten Situation, welche Informationen wir von einer Korrespondenzanalyse erwarten können.
Diese Anwendung der Korrespondenzanalyse kann übrigens auch als mehrdimensionale Skalierung (MDS) eingeordnet werden (vgl. Gifi 1990, S. 280).



Die Korrespondenzanalyse im Vergleich zum klassischen Homogenitätstest

Im Falle einer zweidimensionalen Häufigkeitstabelle beschränkt sich die klassische, konfirmatorisch orientierte Statistik auf den χ^2-Test zur Nullhypothese homogener Zeilen- bzw. Spaltenprofile.[1] Es wird also nur die Frage untersucht, ob die Profile identisch oder verschieden sind.[2] Detaillierte Interpretationshilfen werden nicht gegeben. Während bei kleinen Tabellen (z.B. 2×2) das Effektmuster noch leicht überblickt werden kann, ist es bei großen Tabellen schwierig, die wesentlichen Strukturen durch "optische Inspektion" zu eruieren. Im Unterschied zum klassischen χ^2-Test versucht die KA, möglichst reichhaltige deskriptive Informationen über die Beziehungen zwischen den Profilen zu gewinnen. Diese Methode ist also der explorativen Statistik zuzurechnen.

[1] Äquivalent zur Profilhomogenität ist die Unabhängigkeit der Zeilenvariablen von der Spaltenvariablen.

[2] In den meisten Anwendungsfällen kann die Nullypothese perfekt homogener Profile a priori als falsch verworfen werden. Ob es zu einem signifikanten Testergebnis kommt, hängt ausschließlich von der Stichprobengröße bzw. von den Randverteilungen ab.

Beispiel 2: Verteilung von Forschungsgeldern auf verschiedene Wissenschaftsdisziplinen

In der folgenden Tabelle aus Greenacre (1993, S.75) ist für verschiedene Wissenschaftsdisziplinen angegeben, wie ihre Anträge auf Forschungsförderung in die Unterstützungsklassen A (= stärkste Förderung), .., D (geringe Förderung) und E (Antrag abgelehnt) eingestuft wurden:

	A	B	C	D	E	Summe
Geologie	3	19	39	14	10	85
Biochemie	1	2	13	1	12	29
Chemie	6	25	49	21	29	130
Zoologie	3	15	41	35	26	120
Physik	10	22	47	9	26	114
Ingenieurwesen	3	11	25	15	34	88
Mikrobiologie	1	6	14	5	11	37
Botanik	0	12	34	17	23	86
Statistik	2	5	11	4	7	29
Mathematik	2	11	37	8	20	78
Summe	31	128	310	129	198	796

Sie finden die Daten in der SPSS-Datendatei **FOUND.SAV** an der im Vorwort vereinbarten Stelle.
Informativer als die absoluten Zellhäufigkeiten sind die *relativen*, bezogen auf Zeilen oder Spaltensummen. So erhalten wir in unserem Beispiel für jede Zeile (Wissenschaftsdisziplin) ein Erfolgsprofil mit den relativen Häufigkeiten der einzelnen Förderklassen bei ihren Anträgen.

SPSS-ANACOR liefert uns die folgenden Zeilenprofile:

```
The Rowprofiles:

                  1         2         3         4         5
                     A         B         C         D         E   Margin

   1 Geologie     ,035      ,224      ,459      ,165      ,118    1,000
   2 Biochemi     ,034      ,069      ,448      ,034      ,414    1,000
   3 Chemie       ,046      ,192      ,377      ,162      ,223    1,000
   4 Zoologie     ,025      ,125      ,342      ,292      ,217    1,000
   5 Physik       ,088      ,193      ,412      ,079      ,228    1,000
   6 Ingenieu     ,034      ,125      ,284      ,170      ,386    1,000
   7 Mikrobio     ,027      ,162      ,378      ,135      ,297    1,000
   8 Botanik      ,000      ,140      ,395      ,198      ,267    1,000
   9 Statisti     ,069      ,172      ,379      ,138      ,241    1,000
  10 Mathemat     ,026      ,141      ,474      ,103      ,256    1,000
              -------- -------- -------- -------- --------
     Margin       ,039      ,161      ,389      ,162      ,249
```

Bei fast jeder Analyse einer zweidimensionalen Tabelle wird sich eine *asymmetrische* Betrachtung aufdrängen, im Sinne der Unterscheidung zwischen einer unabhängigen und einer abhängigen Variablen, wobei die Rollenverteilung von der Fragestellung abhängt. In unserem Beispiel sind wohl vor allem die Förder- bzw. Zeilenprofile in Abhängigkeit von der Wissenschaftsdisziplin interessant, möglich ist aber auch die Betrachtung der Spaltenprofile, also der förderstufenbedingten Fächerverteilungen:

The Columnprofiles:

	1 A	2 B	3 C	4 D	5 E	Margin
1 Geologie	,097	,148	,126	,109	,051	,107
2 Biochemi	,032	,016	,042	,008	,061	,036
3 Chemie	,194	,195	,158	,163	,146	,163
4 Zoologie	,097	,117	,132	,271	,131	,151
5 Physik	,323	,172	,152	,070	,131	,143
6 Ingenieu	,097	,086	,081	,116	,172	,111
7 Mikrobio	,032	,047	,045	,039	,056	,046
8 Botanik	,000	,094	,110	,132	,116	,108
9 Statisti	,065	,039	,035	,031	,035	,036
10 Mathemat	,065	,086	,119	,062	,101	,098
Margin	1,000	1,000	1,000	1,000	1,000	

2 Gesamtvariation einer Tabelle (Inertia) und χ^2-Distanzen

Zur Beurteilung der Frage, ob ein Vergleich der Zeilenprofile interessant ist, kann bei Häufigkeitstabellen, nicht jedoch bei sonstigen Korrespondenztabellen, die Pearson- χ^2-Statistik zur Unabhängigkeits- bzw. Homogenitätshypothese herangezogen werden. Die Homogenitätshypothese zu einer zweidimensionalen Häufigkeitstabelle behauptet gerade die Identität der Zeilenprofile bzw. der bedingten Verteilungen in den Zeilen. Die Pearson χ^2-Statistik summiert quadrierte und gewichtete Abeichungen der beobachteten Zellenhäufigkeiten n_{ij} von den unter der Homogenitätsannahme zu erwartenden Häufigkeiten e_{ij} auf:

$$\chi^2 := \sum_{i=1}^{z}\sum_{j=1}^{s}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}, \quad \text{mit } e_{ij} := \frac{n_{i.}\cdot n_{.j}}{n} \tag{1}$$

Darin bedeuten:

z, s	=	Anzahl der Zeilen bzw. Spalten
n_{ij}	=	Beobachtete Häufigkeit in Zelle ij
e_{ij}	=	Unter der Homogenitätshypothese erwartete Häufigkeit in Zelle ij
$n_{i.}$	=	Beobachtete Häufigkeit in Zeile i
$n_{.j}$	=	Beobachtete Häufigkeit in Spalte j
n	=	Größe der Gesamtstichprobe

Ein kleiner χ^2-Wert signalisiert eine weitgehende Homogenität, und in dieser Situation kann natürlich auch die KA kaum interessante Strukturen entdecken. Wurde die betrachtete Häufigkeitstabelle aus einer Zufallsstichprobe gewonnen, dann kann trotz der oben geäußerten Skepsis gegenüber der Inferenzstatistik der χ^2-Test zur Homogenitätshypothese als Interpretationshilfe herangezogen werden.
Er ist in SPSS über folgenden Menübefehl verfügbar:

Statistik > deskriptive Statistik > Kreuztabellen...

In der Subdialogbox **Kreuztabellen, Statistiken** muß das Kontrollkästchen **Chi-Quadrat** angekreuzt werden. Für unser Beispiel erhalten wir:

Chi-Square	Value	DF	Significance
Pearson	65,97151	36	,00168

Die χ^2-Statistik kann als Summe der quadrierten und gewichteten χ^2-Abstände der Zeilenprofile bzw. -verteilungen vom Zeilen-Randprofil bzw. von der zugehörigen Randverteilung ausgedrückt werden. Dieser wichtige Satz bedarf einiger Erläuterungen, vor allem muß der Begriff eines χ^2-Abstandes noch definiert werden.
Wir wollen im folgenden das i-te Zeilenprofil mit den relativen Häufigkeiten der Zellen in der i-ten Zeile, bezogen auf die Gesamthäufigkeit $n_{i.}$ der i-ten Zeile, mit "p_i" bezeichnen:

$$p_i := \left(p_{i_1} \quad . \quad . \quad p_{i_s}\right) := \left(\frac{n_{i1}}{n_{i.}} \quad . \quad . \quad \frac{n_{is}}{n_{i.}}\right),\ i = 1,..,z$$

Unter dem Zeilen-Randprofil ist der folgende Vektor c zu verstehen (in der obigen SPSS-ANACOR-Ausgabe: "Margin"):

$$c := (c_1 \quad . \quad . \quad c_s) := \left(\frac{n_{.1}}{n} \quad . \quad . \quad \frac{n_{.s}}{n} \right)$$

Die Komponente c_j im Zeilen-Randprofil c wird auch als die **Masse der Spalte *j*** bezeichnet. Zu den erwarteten Häufigkeiten in Formel (1) besteht offenbar folgender Zusammenhang:

$$e_{ij} = \frac{n_{i.} n_{.j}}{n} = c_j n_{i.}$$

Als Gewicht bzw. **Masse der Zeile *i*** wird analog der Quotient aus ihrer Häufigkeit $n_{i.}$ und der Gesamthäufigkeit n verwendet:

$$r_i := \frac{n_{i.}}{n}$$

Nun kann man leicht erkennen, daß sich das oben definierte Zeilen-Randprofil (c_1 . . c_s) als gewichtete Summe der Zeilenprofile ergibt, wenn zur Gewichtung die Zeilenmassen r_i verwendet werden. Für die erste Spalte der Zeilenprofile gilt z.B.:

$$\sum_{i=1}^{z} r_i \; p_{i_1} = \sum_{i=1}^{z} r_i \frac{n_{i1}}{n_{i.}} = \sum_{i=1}^{z} \frac{n_{i.}}{n} \frac{n_{i1}}{n_{i.}} = \frac{1}{n} \sum_{i=1}^{z} n_{i1} = \frac{1}{n} n_{.1} = c_1$$

Mit obigen Vereinbarungen kann man die χ^2-Statistik nach Pearson auch folgendermaßen schreiben:

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{z} \sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \\
&= \sum_{i=1}^{z} n_{i.} \sum_{j=1}^{s} \frac{(\frac{n_{ij}}{n_{i.}} - \frac{e_{ij}}{n_{i.}})^2}{\frac{e_{ij}}{n_{i.}}} \\
&= n \sum_{i=1}^{z} \frac{n_{i.}}{n} \sum_{j=1}^{s} \frac{(p_{i_j} - c_j)^2}{c_j} \\
&= n \sum_{i=1}^{z} r_i \sum_{j=1}^{s} \frac{(p_{i_j} - c_j)^2}{c_j}
\end{aligned}$$

Der i-te Summand der äußeren Summe ist der quadrierte und mit der Zeilenmasse r_i gewichtete **χ^2-Abstand** des i-ten Zeilenprofils p_i vom Zeilen-Randprofil c:

$$\| p_i - c \|_c := \sqrt{\sum_{j=1}^{s} \frac{(p_{ij} - c_j)^2}{c_j}} =: d_i$$

Mit dieser Bezeichnungsvereinbarung können wir für die χ^2-Statistik schreiben:

$$\chi^2 = n \sum_{i=1}^{z} r_i \| p_i - c \|_c^2 = n \sum_{i=1}^{z} r_i d_i^2$$

Dieser Ausdruck wird um so größer, je stärker sich die Zeilenprofile unterscheiden. Er quantifiziert somit die gesamte Variation bzw. Dispersion einer Tabelle F. Der Quotient aus diesem Ausdruck und der Gesamthäufigkeit n wird in der englischsprachigen KA-Literatur als **Inertia In(*F*)** bezeichnet:

$$\mathrm{In}(F) := \frac{\chi^2}{n} = \sum_{i=1}^{z} r_i d_i^{\,2} \qquad (2)$$

In(F) ist also eine Summe gewichteter und quadrierter Abstände zwischen Profilvektoren und kann daher als Maß für die "räumliche Dispersion" der Zeilen in F aufgefaßt werden. In der KA wird diese verallgemeinerte „Varianz“ auf unterschiedliche Weise zerlegt. In gewisser Analogie zur Faktorenanalyse wird versucht, die Gesamtvariation durch ein möglichst niedrigdimensionales Model zu erklären.
Im Unterschied zur χ^2-Statistik ist In(F) normiert und damit leichter zu interpretieren. Es gilt:

$$0 \le \mathrm{In}(F) \le \min(z, s) - 1.$$

Der Maximalwert ist identisch mit der Dimensionalität des Problems (in unserem Fall also 4) und kommt dann zustande, wenn alle Profile verschieden sind, und dabei in jeder Zeile die gesamte Masse auf *eine* Spalte konzentriert ist (Greenacre 1993, S. 30, 87). Dann liegen alle Profilvektoren in den "Ecken" des potentiellen Aufenthaltsraumes. Ein typisches Profil sieht dann folgendermaßen aus:

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Der minimale Wert (In(F) = 0) resultiert, wenn alle Zeilenprofile identisch sind.

Bei einem Zeilenprofil p_i interessiert nicht nur sein Abstand vom Zeilen-Randprofil c. Es wird sich herausstellen, daß die Analyse bzw. grafische Darstellung von χ^2-Abständen zwischen Zeilenprofilen im Mittelpunkt der KA steht. Für die Zeilenprofile p_i und p_k ist der χ^2-Abstand definiert durch:

$$\left\| p_i - p_k \right\|_c := \sqrt{\sum_{j=1}^{s} \frac{(p_{i_j} - p_{k_j})^2}{c_j}}$$

Hier werden also für das betrachtete Paar von Profilen über alle Spalten die Differenzen ihrer relativen Häufigkeiten quadriert und gewichtet aufsummiert, wobei als Gewicht einer Spalte der Kehrwert $1/c_j$ ihrer Masse verwendet wird.

Ohne die Gewichtungsfaktoren $1/c_j$ würde sich die einfache, euklidische Distanzdefinition ergeben. Warum macht man sich in der KA das Leben so schwer? Die Begründung ist zumindest bei Häufigkeitstabellen schlüssig: Die Gewichtung verhindert, daß Spalten mit großer relativer Häufigkeit, die auch bei perfekter "Populations"-Homogenität zweier Zeilen durch Zufälligkeiten der Stichprobe große Unterschiede zwischen den Zeilen erwarten lassen, das Abstandsmaß dominieren. Die Normierung sorgt also dafür, daß alle Spalten eine annähernd gleiche Variation in die Korrespondenzanalyse einbringen. Somit erfüllt die Gewichtung der Spalten mit $1/c_j$ eine ähnliche Funktion wie die bei Faktorenanalysen übliche Standardisierung der Variablen (Greenacre 1993, S. 36).

3 Repräsentation der Zeilenprofile und χ^2-Abstände in einem niedrig-dimensionalen Raum

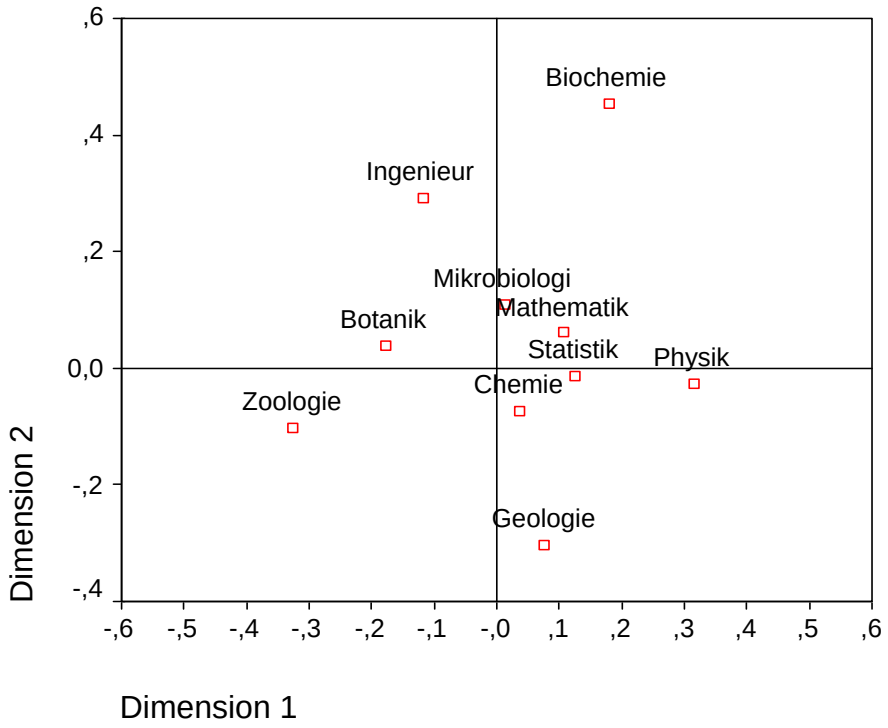
Nun können wir das zentrale Ziel der Korrespondenzanalyse schon sehr genau beschreiben:

> Die Zeilenprofile p_i sollen in einem möglichst niedrig-dimensionalen euklidischen Raum so durch Punkte repräsentiert werden, daß die einfache, euklidische Distanz zwischen je zwei Punkten in guter Näherung ihrem χ^2-Abstand entspricht.

In unserem Beispiel sollen folglich 10 Zeilenprofile, die als 5-Tupel eigentlich in einem 5-dimensionalen Raum "zu Hause" sind, mit all ihren paarweisen χ^2-Abständen bei möglichst geringem Informationsverlust in einem möglichst einfachen Raum, am besten in der Ebene des $\mathbb{R}^2$, dargestellt werden. Es sind nun zwei wichtige Probleme zu lösen:

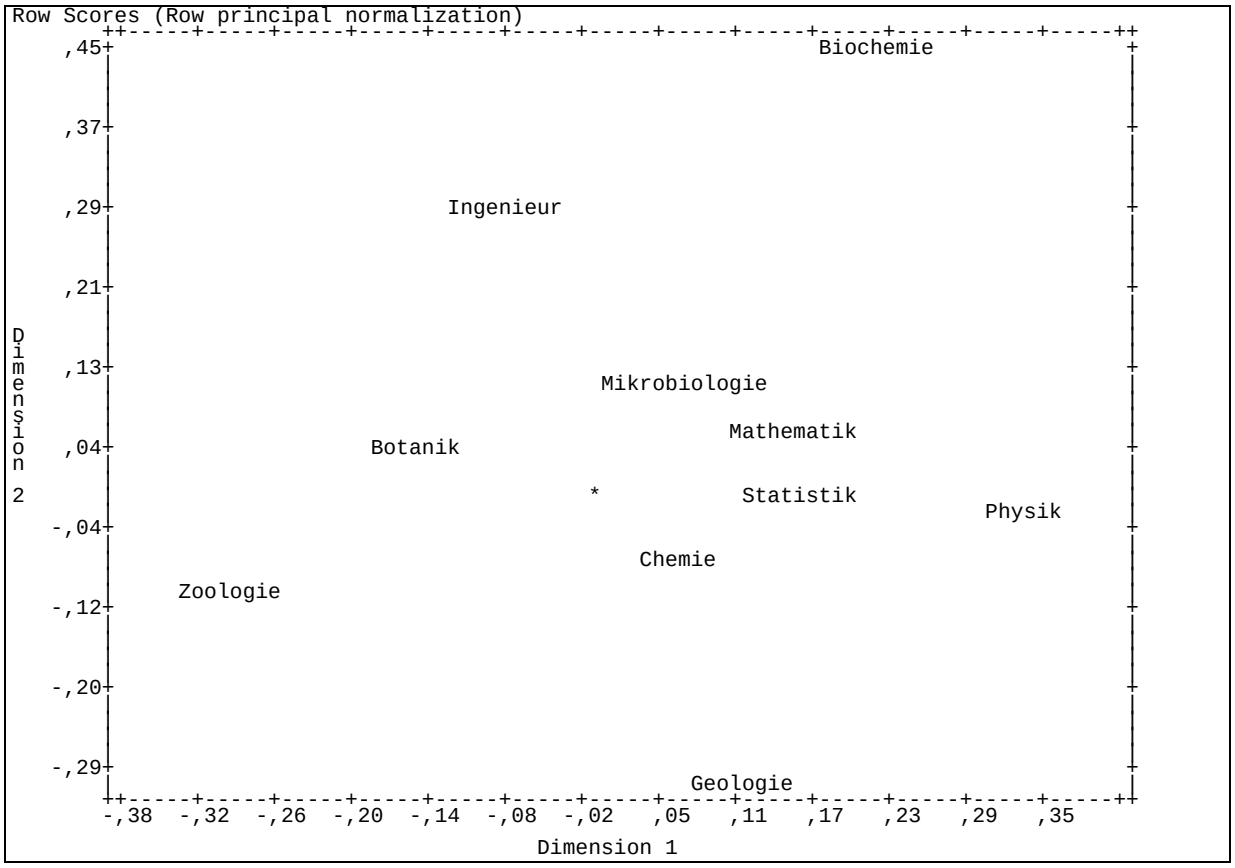
- Es ist eine euklidische Repräsentation der χ^2-Abstände zu bestimmen (vgl. Abschnitt 0).
- Es ist eine Dimensionsreduktion bei möglichst geringem Informationsverlust vorzunehmen (vgl. Abschnitt 3.2).

SPSS-ANACOR liefert folgende Lösung, deren Zustandekommen und deren Eigenschaften noch besprochen werden müssen:



Hier können wir z.B. unmittelbar ablesen, daß - jeweils in der χ^2-Metrik - Mathematik und Statistik relativ ähnliche Förderungsprofile haben (geringer euklidischer Abstand, hohe Korrespondenz) und daß Biochemie am weitesten vom Durchschnittsprofil (dargestellt durch einen Stern) entfernt ist.
Das Durchschnittsprofil der Zeilen wurde übrigens der Übersichtlichkeit halber in den Nullpunkt des Koordinatensystems verschoben, indem von allen Zeilenprofilen ihr gewichtetes Mittel c subtrahiert wurde.

Die obige hochaufgelöste Grafik ist nicht sehr übersichtlich. Zur Korrespondenzanalyse sind die Drucker-Semigrafiken generell besser gelungen, so daß sie in diesem Manuskript häufig bevorzugt werden:



3.1 Die euklidische Repräsentation der χ^2-Abstände

Nun zum ersten der oben angesprochenen Probleme: Wie können Punkte in einem euklidischen Koordinatensystem so gewählt werden, daß ihre Entfernungen voneinander gerade den χ^2-Abständen zwischen den Zeilenprofilen entsprechen?
Dazu definieren wir zu einer Korrespondenzmatrix *F* die Matrix *B*, deren Elemente folgendermaßen aussehen:

$$b_{ij} := \frac{p_{ij} - c_j}{\sqrt{c_j}}$$

Für zwei Zeilen(profile) b_i und b_k aus *B* erhalten wir die (euklidische) Entfernung:

$$\|b_i - b_k\|_2 := \sqrt{\sum_{j=1}^{s}(b_{ij} - b_{kj})^2} = \sqrt{\sum_{j=1}^{s}\left(\frac{p_{ij} - c_j - (p_{kj} - c_j)}{\sqrt{c_j}}\right)^2} = \sqrt{\sum_{j=1}^{s}\frac{\left(p_{ij} - p_{kj}\right)^2}{c_j}} = \|p_i - p_k\|_c$$

Die Zeilen bzw. Punkte b_i aus der Matrix *B* leisten also tatsächlich eine euklidische Repräsentation der χ^2-Abstände, aber natürlich handelt es sich wie bei den ursprünglichen Zeilenprofilen p_i aus immer noch um Punkte im $\mathbb{R}^5$, die wir nicht grafisch darstellen können. Wir kommen somit zum zweiten angekündigten Problem: zur Dimensionsreduktion.

3.2 Dimensionsreduktion

Zunächst machen wir uns in einer Nebenüberlegung kurz klar, warum unser Problem die folgende maximale Dimensionalität *K* besitzt:

$$K := \min(z, s) - 1$$

Die *z* Zeilenprofile p_i, liegen als *s*-Tupel trivialerweise in einem Raum der maximalen Dimensionalität min(*z*, *s*). Weil sie konstruktionsgemäß die folgende Gleichung:

$$\sum_{j=1}^{s} p_{ij} = 1$$

erfüllen, befinden sie sich in einer Hyperebene mit Dimension min(*z*, *s*).
Diese Hyperebene enthält im allgemeinen nicht den Ursprung des Koordinatensystems, so daß man nicht auf lineare Abhängigkeit der Zeilenprofile p_i schließen kann. Demgegenüber ergeben die Vektoren b_i bei Zeilenmassen-gewichteter Summation den Nullvektor, wie folgende Überlegung für die beliebig herausgegriffene Spalte *j* zeigt:

$$\sum_{i=1}^{z} r_i b_{ij} = \frac{1}{\sqrt{c_j}} \sum_{i=1}^{z} \frac{n_{i.}}{n} (p_{ij} - c_j) = \frac{1}{n\sqrt{c_j}} \left(\sum_{i=1}^{z} n_{i.} \frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n} \sum_{i=1}^{z} n_{i.} \right) = \frac{1}{n\sqrt{c_j}} \left(n_{.j} - n_{.j} \right) = 0$$

Damit sind die *z* Zeilen von *B* linear abhängig voneinander und befinden sich in einem *linearen* Unterraum mit der maximalen Dimension:

$$K := \min(z, s) - 1$$

Bei einer Korrespondenzanalyse ist nun folgende Optimierungsaufgabe zu lösen:

> Bestimme einen *k* (meist 2)-dimensionalen Unterraum *S* des $\mathbb{R}^K$ derart, daß die gewichtete Summe der quadrierten euklidischen Abstände $d_i(S)$ zwischen *S* und den Profilen b_i minimal wird, wobei zur Gewichtung wiederum die Zeilenmassen r_i herangezogen werden sollen:

$$\sum_{i=1}^{z} r_i \left[d_i(S) \right]^2 \xrightarrow{!} \min$$

Dieses Problem läßt sich mit der sogenannten Singulärwertzerlegung der linearen Algebra lösen, die hier nur grob skizziert werden soll (siehe Greenacre 1984, Gifi 1990):

Im dem durch die Zeilen von *B* aufgespannten Raum wird eine erste Hauptachse so bestimmt, daß sie minimalen Abstand von den Zeilen bzw. Profilen in *B* hat, d.h. daß die gewichtete Summe der quadrierten Projektionen aller Zeilen bzw. Profile in *B* auf diese Hauptachse maximal wird. Dann wird eine dazu orthonormale Hauptachse nach demselben Kriterium bestimmt.
So ergeben sich maximal *K* Hauptachsen, mit denen die Zeilen von *B perfekt* dargestellt werden können, und durch die folglich auch die gesamte Variation (Inertia) der Zeilenprofile erklärt wird. Damit ergibt sich eine **Zerlegung der Inertia** in die sukzessiv schrumpfenden Anteile der einzelnen Hauptachsen.
Wenn wir mit "ϕ_ν" die von der ν-ten Dimension erklärte Variation bezeichnen, erhalten wir folgende Zerlegung:

$$\text{In}(F) = \sum_{\nu=1}^{\min(z,s)-1} \phi_\nu$$

Mit den Erklärungsbeiträgen bzw. Variationsanteilen ϕ_ν wollen wir uns später noch etwas näher beschäftigen.

In unserem Beispiel mit 5 Zeilen, 10 Spalten erhalten wir min(5,10)-1 = 4 Hauptachsen mit folgenden Inertia-Anteilen:

Dimension	Singular Value	Inertia	Proportion Explained	Cumulative Proportion
1	,19778	,03912	,472	,472
2	,17430	,03038	,367	,839
3	,10426	,01087	,131	,970
4	,05012	,00251	,030	1,000
Total		,08288	1,000	1,000

Zwar scheint das Problem der Dimensionsreduktion nur verlagert worden zu sein, doch es steht kurz vor seiner Lösung. Denn die Singulärwertzerlegung liefert uns die beste Lösung im obigen Sinne für jede gewünschte Dimensionalität *k*: Dazu müssen wir uns lediglich auf den Raum der ersten *k* Hauptachsen beschränken. In unserem Beispiel erfahren wir etwa aus den obigen Ergebnissen, daß durch eine zweidimensionale Lösung 84 % der Gesamtvariation (= 0,08288 = 65,97/796, siehe Formel (2)) aufgeklärt werden. Ein Großteil der räumlichen Variation der Zeilenprofile spielt sich also in der Ebene ab, die von den ersten beiden Dimensionen aufgespannt wird.

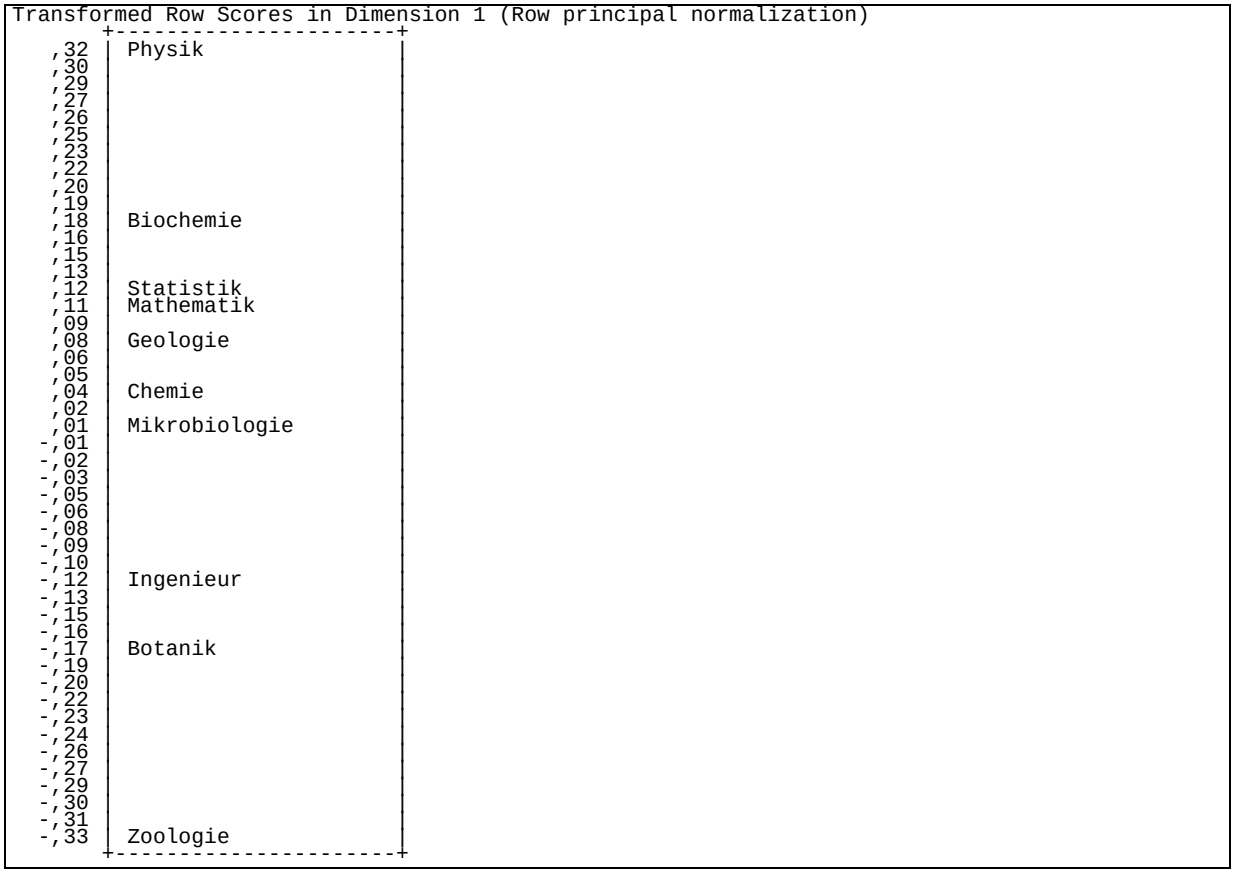
Als Koordinaten der Zeilenprofile in den Hauptachsen erhalten wir:

Row Scores:

FACH	Marginal Profile	Dim 1	2
1 Geologie	,107	,076	-,303
2 Biochemi	,036	,180	,455
3 Chemie	,163	,038	-,073
4 Zoologie	,151	-,327	-,102
5 Physik	,143	,316	-,027
6 Ingenieu	,111	-,117	,292
7 Mikrobio	,046	,013	,110
8 Botanik	,108	-,179	,039
9 Statisti	,036	,125	-,014
10 Mathemat	,098	,107	,061

Genau diese Koordinaten werden in der oben aus Motivationsgründen vorgezogenen Abbildung grafisch dargestellt. Es versteht sich von selbst, daß die Entfernungen zwischen zwei Punkten im reduzierten Raum nur noch approximativ den χ^2-Abständen zwischen den ursprünglichen Profilen entsprechen. In der Spalte "Marginal Profile" sind die Zeilenmassen r_i nochmals angegeben.
Die Scores zur ν-ten Hauptachse stellen Quantifizierungen der Zeilenkategorien dar, mit deren Eigenschaften wir uns später noch näher beschäftigen werden.

ANACOR kann die Scores der Zeilenprofile auf den Hauptachsen auch einzeln grafisch darstellen. Für die erste Hauptachse erhalten wird etwa:



Es wird sich zeigen, daß diese Darstellung in grober Näherung als Förderungshitliste der Wisenschaftsdisziplinen interpretiert werden kann (siehe Seite 17 und Abschnitt 6.3). Daß die Tabelle *F* mehr Information über die Zeilenprofile enthält als eine Quantifizierung nach Förderungsumfang, wird durch die Tatsache belegt, daß die erste Dimension nur 47,2 % der Gesamtvariation repräsentiert.

4 Repräsentation der Spalten

In der bisherigen Darstellung sind die Spaltenkategorien zu kurz gekommen. Es wäre durchaus sinnvoll, die Förderkategorien in die grafischen Darstellungen aufzunehmen, um etwa für eine Wissenschaftdisziplin die relativen Anteile der einzelnen Förderkategorien durch Distanzen zwischen dem Zeilenprofil der Disziplin und entsprechenden Repräsentationen der Spalten- bzw. Förderkategorien darstellen zu können.
Im Raum der ursprünglichen Zeilenprofile (bei uns: IR^5) kann die Förderkategorie *j* durch ein künstliches (Eck-)Profil dargestellt werden, das eine Eins enthält für die relative Häufigkeit der *j*-ten Kategorie und ansonsten Nullen, z.B. für Kategorie C (vgl. Abschnitt 2):

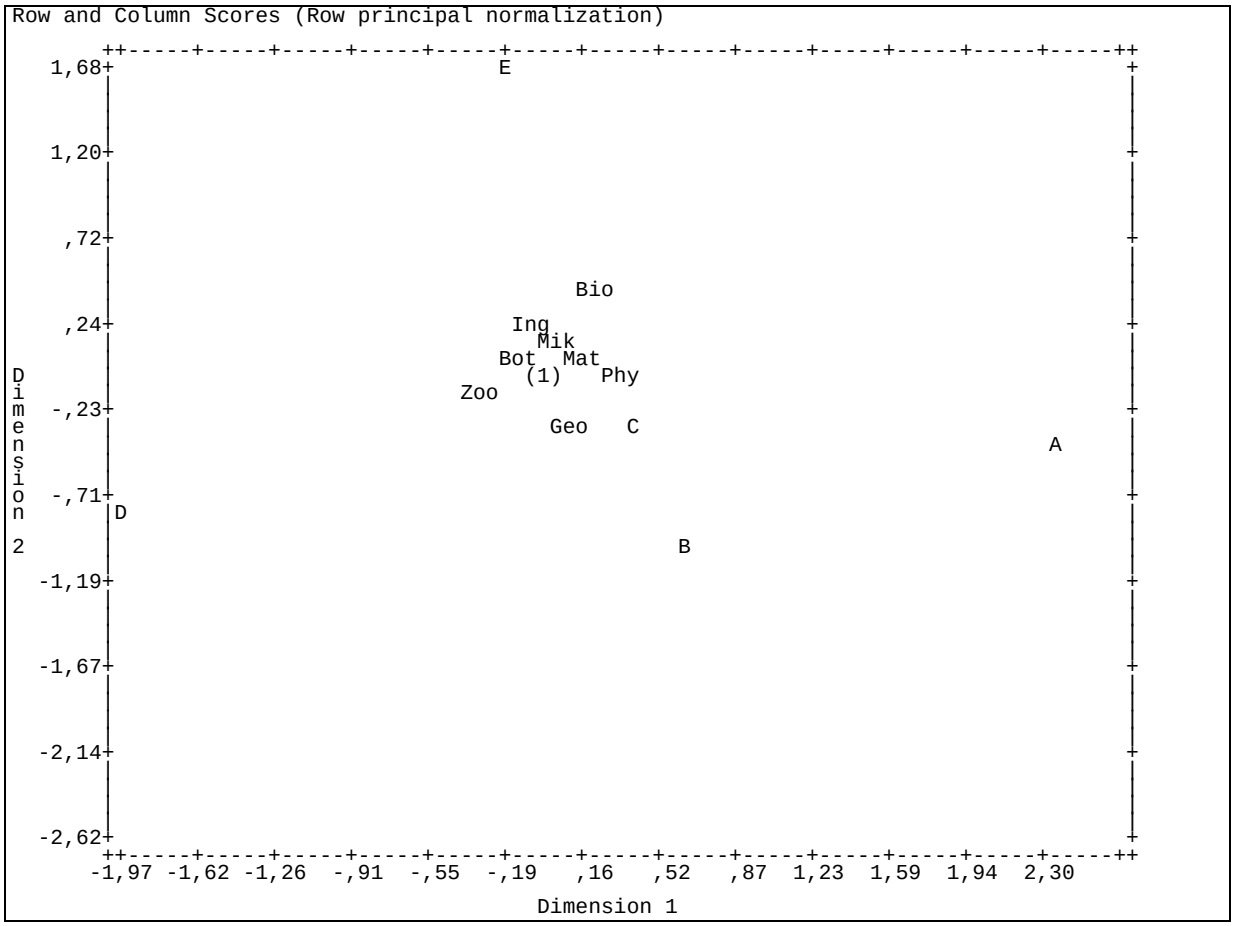
$$(0 \quad 0 \quad 1 \quad 0 \quad 0)$$

Auch diese künstlichen Profile, die jeweils eine Spaltenkategorie repräsentieren, können auf die oben hergeleitete optimale Ebene projiziert werden. Als Koordinaten und damit als Quantifizierungen der Spaltenkategorien erhalten wir in unserem Beispiel:

Column Scores:

FUND	Marginal Profile	Dim 1	2
1 A	,039	2,417	-,415
2 B	,161	,643	-,995
3 C	,389	,417	-,286
4 D	,162	-1,974	-,799
5 E	,249	-,161	1,676

Natürlich sind die zur Repräsentation der Spaltenkategorien herangezogenen Förderungsprofile sehr extrem, was sich durch entsprechende räumliche Verhältnisse auch in der folgenden gemeinsamen Darstellung von Zeilenprofilen und Spaltenkategorien zeigt. Durch die erforderlichen Skalenwechsel sind die Zeilenprofile ins Zentrum gerückt.



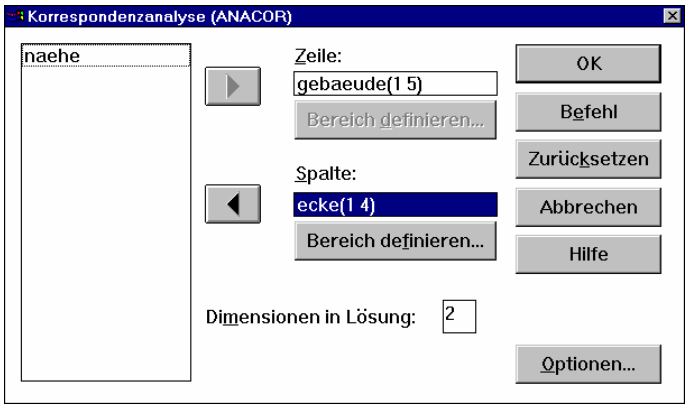
Die Extremprofile werden als Ecken aus einem hochdimensionalen Simplex in dem gewählten Unterraum schlechter lokalisiert als die Zeilenprofile. Der Unterraum wurde schließlich so bestimmt, daß *die Zeilenprofile* so präzise wie möglich dargestellt werden können. In der SPSS-Ausgabe wird dies durch den Zusatz „Row principal normalization“ zum Ausdruck gebracht. Trotzdem können die projizierten Spaltenkategorien-Punkte als Bezugspunkte bei der Interpretation der Zeilenprofil-Positionen verwendet werden (Greenacre 1993, S. 44).
Außerdem leisten sie wertvolle Beiträge zur Interpretation der Dimensionen. Wir finden etwa auf der ersten Dimension die Kategorien A, B, C und D in ihrer natürlichen Ordnung vor. Die Kategorie E (Antrag abgelehnt) erhält überraschenderweise auf der ersten Dimension einen mittleren Wert, setzt sich aber auf der zweiten Dimension von allen anderen Kategorien ab. Wir können also z.B. unsere erste Dimension als „Förderungshitliste der Wissenschaftsdisziplinen“ interpretieren.

5 Testfall: Was macht ANACOR mit den Campus-Daten?

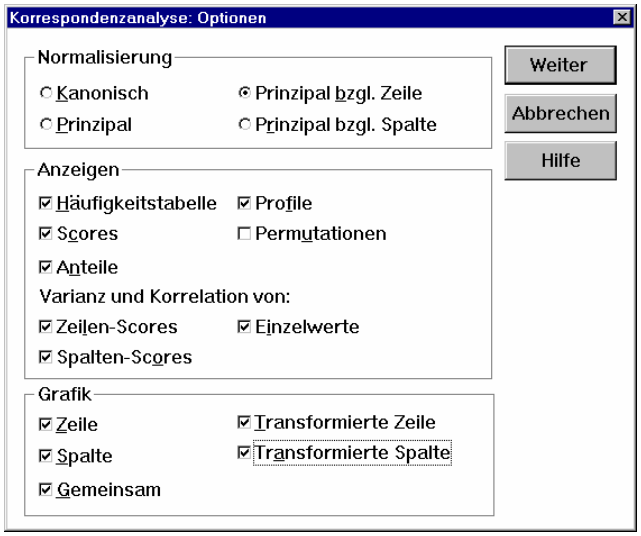
Nach so viel anstrengender Mathematik wollen wir uns nun anhand einiger Ergebnisse zum Beispiel 1 vergewissern, ob die Qualität der Ergebnisse den Aufwand einer intensiven Beschäftigung mit der Korrespondenzanalyse rechtfertigt.
Außerdem soll die EDV-Praxis nicht ganz vernachlässigt werden, so daß an dieser Stelle die Anforderung einer Korrespondenzanalyse mit SPSS anhand der Campus-Daten demonstriert werden soll.
Nach den obigen Ausführungen wundert es nicht, daß die Korrespondenzanalyse unter **Datenreduktion** einsortiert ist. Der Menübefehl lautet also:

Statistik > Datenreduktion > Korrespondenzanalyse...

In der Dialogbox legen wir GEBAEUDE mit dem Bereich von 1 bis 5 als Zeilenvariable, sowie ECKE im Bereich von 1 bis 4 als Spaltenvariable fest:



In der **Optionen**-Subdialogbox wählen wir die zeilenorientierte Normalisierung, die der obigen Beschreibung entspricht. Zur spaltenorientierten Normalisierung folgen im Verlauf des Manuskripts noch einige Hinweise. Die übrigen ANACOR-Normalisierungsmethoden sind im SPSS-Categories-Handbuch erläutert. Außerdem fordern wir noch einige zusätzliche Ausgaben an:



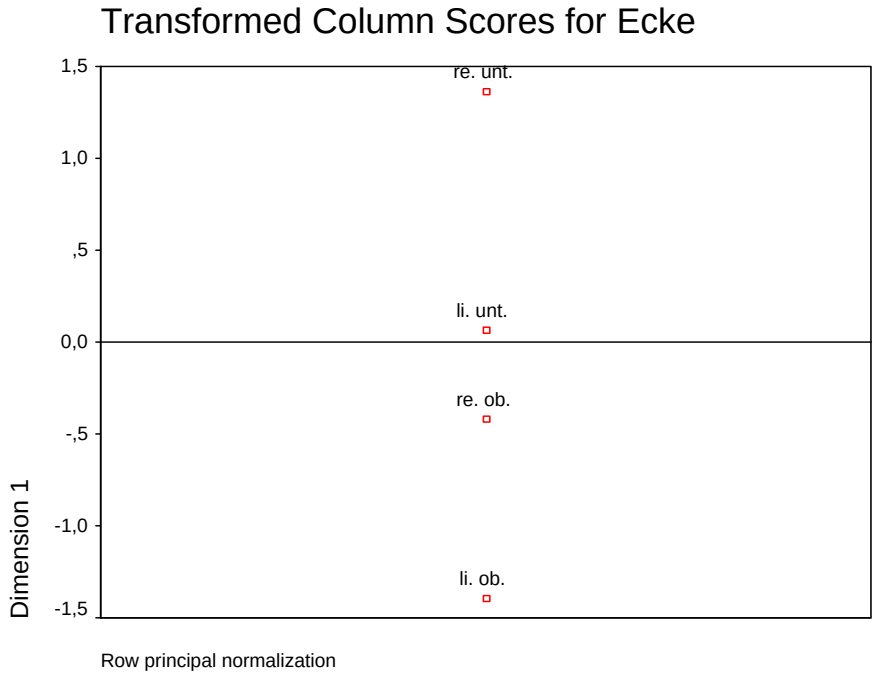
Das Syntaxäquivalent zu unserer Analyseanforderung sieht so aus:

```
ANACOR
  TABLE=gebaeude(1 5)  BY ecke(1 4)
  /DIMENSION=2
  /NORMALIZATION RPRINCIPAL
  /PRINT TABLE SCORES CONTRIBUTIONS PROFILES
  /VARIANCES ROWS COLUMNS SINGULAR
  /PLOT ROWS COLUMNS JOINT TRROWS TRCOLUMNS NDIM(ALL,MAX).
```

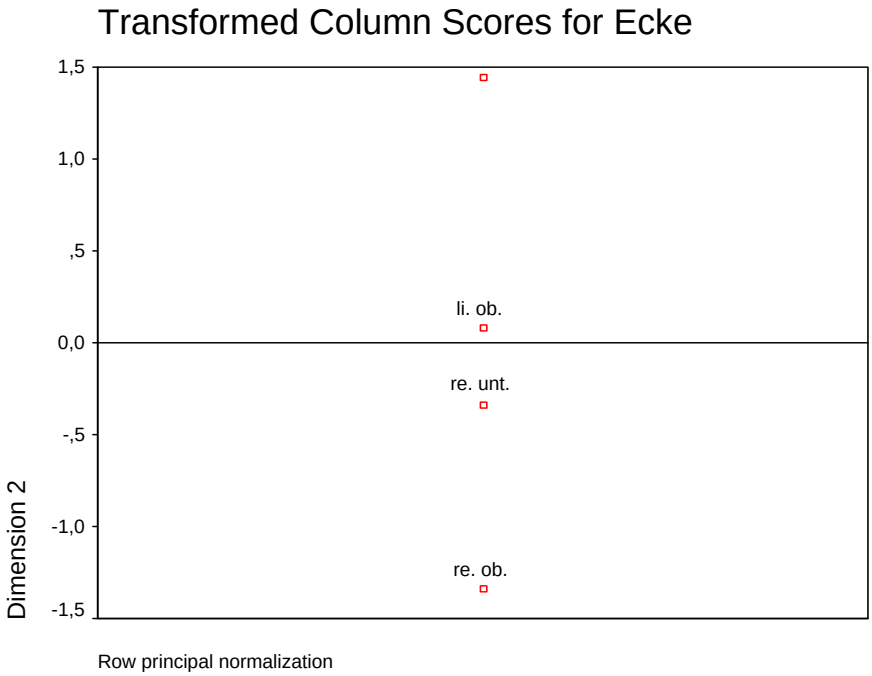
In der resultierenden Ausgabe stellen wir zunächst fest, daß erwartungsgemäß die zweidimensionale Lösung fast die gesamte "räumliche Dispersion" unserer Zeilenprofile erklärt:

Dimension	Singular Value	Inertia	Proportion Explained	Cumulative Proportion
1	,12815	,01642	,689	,689
2	,08559	,00733	,307	,996
3	,01008	,00010	,004	1,000
Total		,02385	1,000	1,000

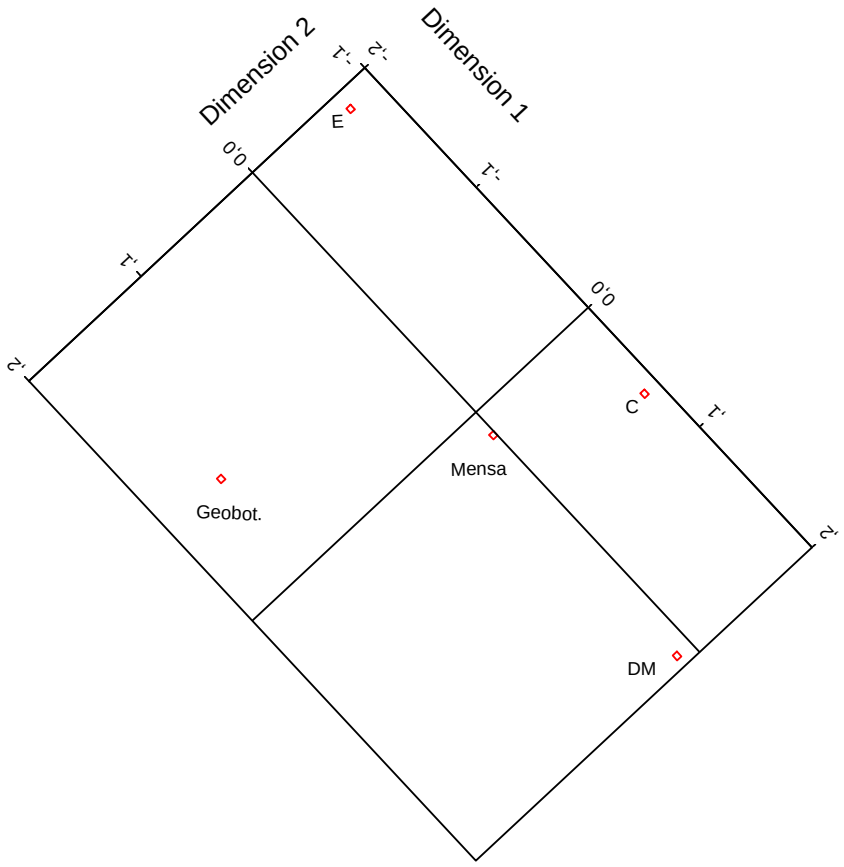
Als nächstes betrachten wir die Projektionen der Ecken auf die beiden Dimensionen, um deren Bedeutung zu verstehen:



Hier handelt es sich offenbar um eine Raumachse, die von links oben (kleinster Wert) nach rechts unten (größter Wert) verläuft. Wir werden die X-Achse der ANACOR-Plots also entsprechend drehen müssen, um eine bequeme Ansicht zu erhalten.
Wie wir wissen, muß die zweite Achse orthogonal zur ersten verlaufen. Allerdings ist damit die Richtung nur bis auf eine Drehung um 180° festgelegt. Leider verläuft in unserem Beispiel die zweite Achse von rechts oben nach links unten, so daß wir eine Spiegelung an der ersten Raumachse vornehmen müssen, um unser übliches Weltbild einzustellen.



Nach den beiden mathematisch unwesentlichen, für uns aber äußerst hilfreichen Korrekturen, liefert der zweidimensionale ANACOR-Plot der Zeilenprofile tatsächlich eine brauchbare Campus-Karte:



6 Inertia-Zerlegungen

In diesem Abschnitt werden Methoden der "Variationsanalyse" besprochen, die durch unterschiedliche Zerlegungen der Gesamtvariation einer Korrespondenztabelle eine detaillierte Interpretation der KA-Lösung ermöglichen.
Wir haben bereits zwei Zerlegungen kennengelernt:

- Zerlegung in Beiträge der Zeilenprofile (vgl. Abschnitt 2)

$$\mathrm{In}(F) = \sum_{i=1}^{z} r_i \left\| p_i - c \right\|_c^2 = \sum_{i=1}^{z} r_i d_i^2$$

  Leider werden die Zeilenbeiträge $r_i d_i^2$ von SPSS-ANACOR nicht ausgegeben.
- Zerlegung in die Erklärungsbeiträge der Dimensionen (vgl. Abschnitt 3.2)
  Wenn wir mit "ϕ_ν" die von der ν-ten Dimension erklärte Variation bezeichnen, erhalten wir folgende Zerlegung:

$$\mathrm{In}(F) = \sum_{\nu=1}^{\min(z,s)-1} \phi_\nu$$

  Mit den Erklärungsbeiträgen bzw. Variationsanteilen ϕ_ν wollen wir uns anschließend noch etwas näher beschäftigen.

6.1 Die durch den Unterraum S erklärte Inertia bzw. Variation In_S

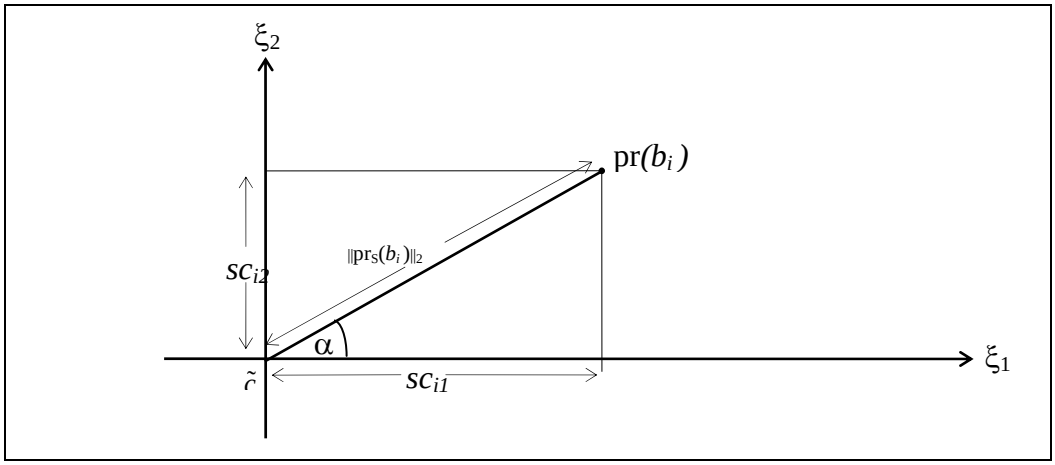
Wie in Abschnitt 3.2 erläutert, wird der optimale Unterraum *S* so gewählt, daß bei festgelegter Dimensionalität, z.B. $k = 2$, die gewichtete Summe der quadrierten euklidischen Abstände $d_i(S)$ zwischen *S* und den Profilen b_i minimal wird, wobei zur Gewichtung die Zeilenmassen r_i herangezogen werden. Dem minimalen euklidischen Abstand $d_i(S)$ eines Profils b_i vom Unterraum *S* entspricht die maximale euklidische Länge seiner Projektion $\mathrm{pr}_S(b_i)$ auf *S*. Daher kann das in Abschnitt 3.2 angegeben Optimierungsziel äquivalent auch folgendermaßen formuliert werden:

$$\sum_{i=1}^{z} r_i \left\| \mathrm{pr}_S(b_i) \right\|_2^2 \xrightarrow{!} \max$$

Weil das Zeilenzentroidprofil der Einfachheit halber in den Nullpunkt verschoben worden ist, sind die euklidischen Längen $\left\| \mathrm{pr}_S(b_i) \right\|_2$ gerade die χ^2-Abstände zwischen den projizierten Zeilenprofilen und dem zentrierten Zeilenzentroidprofil $\tilde{c}$. Die r_i-gewichtete Summe der quadrierten χ^2-Abstände vom Zeilenzentroidprofil ergibt nach der Definitionsgleichung (2) aber gerade die Inertia der Projektionen auf *S*. Damit ist das bei obiger Optimierungsprozedur erreichte Maximum gerade die durch *S* erklärte Inertia bzw. Variation In_S:

$$\mathrm{In}_S(F) = \sum_{i=1}^{z} r_i \left\| \mathrm{pr}_S(b_i) \right\|_2^2$$

Die geometrischen Verhältnisse bei einem zweidimensionalen Unterraum *S* sind in folgender Abbildung wiedergegeben:



Den Inertia-Anteil der *i*-ten Zeile *im reduzierten Raum S* erhält man mit:

$$r_i \left\| \mathrm{pr}_S(b_i) \right\|_2^{\,2}$$

6.2 Erklärungsleistung der ν-ten Dimension

Die Quadrate der euklidischen Längen $\left\| \mathrm{pr}_S(b_i) \right\|_2$ lassen sich nach dem Satz des Pythagoras additiv zerlegen in Beiträge der einzelnen Dimensionen. Dabei sind die Koordinaten der projizierten Profile in den Dimensionen heranzuziehen, die wir schon als "Scores" der Zeilenprofile in den Dimensionen kennengelernt haben. Bei einer Reduktion auf einen 2-dimensionalen Unterraum *S* gilt also:

$$\left\| \mathrm{pr}_S(b_i) \right\|_2^{\,2} = sc_{i1}^{\,2} + sc_{i2}^{\,2}$$

Dabei bezeichnet sc_{i1} den Score des *i*-ten Zeilenprofil auf der ersten Dimension. Sein Quadrat gibt also den *Beitrag der ersten Dimension zum i-ten Zeilenprofil* an.
Der oben ohne exakte Definition eingeführte Variationsanteil $\phi_\nu$der ν-ten Dimension ist nichts anderes als die Summe ihrer gewichteten Beiträge zu allen Zeilenprofilen, wobei wie üblich mit den Zeilenmassen r_i gewichtet wird:

$$\phi_\nu = \sum_{i=1}^{z} r_i \, sc_{i\nu}^{\,2}$$

Aus Gründen, die in Abschnitt 8 näher erläutert werden, variiert ϕ_ν zwischen Null und Eins.
Für die durch den Unterraum *S* mit Dimensionalität *k* erklärte Inertia $\mathrm{In}_S(F)$ gilt:

$$\mathrm{In}_S(F) = \sum_{\nu=1}^{k} \phi_\nu$$

6.3 Einfluß des i-ten Zeilenprofils auf die ν-te Dimension

Man möchte bei einer Korrespondenzanalyse natürlich auch verstehen, was die ermittelten Dimensionen bedeuten. Die Zeilenscores können dazu wertvolle Hinweise liefern. Wieviel Information das *i*-te Zeilenprofil über die ν-te Dimension enthält, spiegelt sich darin wieder, welcher Anteil ihrer Erklärungsleistung ϕ_ν auf die *i*-te Zeile entfällt. Erinnern wir uns zur Verdeutlichung dieses Sachverhaltes daran, aufgrund welches Optimierungsprinzips die Dimensionen festgelegt werden. Die erste Dimension wird so gewählt, d.h. die Scores der Zeilenprofile auf der ersten Dimension ξ_1 werden so festgelegt, daß die Summe der Zeilen-

massen-gewichteten quadrierten Scores sc_{i1} maximal wird. Geometrisch betrachtet, wird also die erste Dimension ξ_1 so ausgerichtet, daß die gewichteten Projektionen $\mathrm{pr}_{\xi_1}(b_i)$ der Zeilenprofile b_i auf ξ_1 maximal lang werden:

$$\phi_1 = \sum_{i=1}^{z} r_i \left\| \mathrm{pr}_{\xi_1}(b_i) \right\|_2^2 = \sum_{i=1}^{z} r_i \, sc_{i1}^2 \xrightarrow{!} \max$$

Der Einfluß des *i*-ten Zeilenprofils auf die Lokalisation der ersten Dimension hängt, analog zur physikalischen Gravitationstheorie, von zwei Faktoren ab:

- Lage des *i*-ten Zeilenprofils b_i im Raum
- Masse des *i*-ten Zeilenprofils

Man betrachtet daher den folgenden Einfluß $consol_{ik}$ der *i*-ten Zeile auf die Lösung zur *k*-ten Dimension:

$$consol_{ik} := \frac{r_i \, sc_{ik}^2}{\phi_k} = \frac{r_i \, sc_{ik}^2}{\sum_{i=1}^{z} r_i \, sc_{ik}^2}$$

Die Einflußgröße $consol_{ik}$ schwankt zwischen folgenden Extremwerten:

- Null
  Das *i*-te Zeilenprofil hatte überhaupt keine Anziehungskraft auf die *k*-te Dimension. Beide liegen orthogonal zueinander.
- Eins
  Das *i*-te Zeilenprofil hat die *k*-te Dimension ganz in seine Richtung gezogen. Alle übrigen Zeilenprofile liegen orthogonal dazu.

In unserm Beispiel erhalten wir:

Contribution of row points to the inertia of each dimension:

FACH		Marginal Profile	Dim 1	Dim 2
1	Geologie	,107	,016	,322
2	Biochemi	,036	,030	,248
3	Chemie	,163	,006	,029
4	Zoologie	,151	,413	,052
5	Physik	,143	,365	,003
6	Ingenieu	,111	,039	,310
7	Mikrobio	,046	,000	,018
8	Botanik	,108	,088	,005
9	Statisti	,036	,014	,000
10	Mathemat	,098	,029	,012
			1,000	1,000

Es zeigt sich, daß die erste Dimension vor allem von Zoologie und Physik markiert wird, während bei der zweiten Dimension Geologie, Biochemie und Ingenieurwesen herausragen.
In der Spalte "Marginal Profile" enthält die Tabelle zu Vergleichszwecken wieder die Zeilengewichte.

Beitrag der *k*-ten Dimension zur Erklärung des *i*-ten Zeilenprofils

Die Scores sc_{ik} lassen sich noch auf andere Weise zu aufschlußreichen Verhältnisindikatoren verrechnen. Wir möchten zu jedem Zeilenprofil wissen, wie gut es insgesamt durch die zwei-

dimensionale Lösung erklärt werden kann, und welche Erklärungbeiträge die einzelnen Dimensionen leisten. Als Erklärungsbeitrag $conex_{ik}$ der k-ten Dimension zum i-ten Zeilenprofil definiert man:

$$conex_{ik} := \frac{sc_{ik}^{\,2}}{\|p_i - c\|_c^{\,2}} = \frac{sc_{ik}^{\,2}}{d_i^{\,2}} = \frac{sc_{ik}^{\,2}}{\|b_i\|_2^{\,2}}$$

Im Nenner steht der quadrierte χ^2-Abstand des i-ten Zeilenprofils vom Zeilenzentroidprofil, im Zähler steht der Anteil des quadrierten χ^2-Abstandes "in Richtung der k-ten Dimension".
Die Erklärungsanteile bewegen sich wie die obigen Einflußgrößen interpretationsfreundlich zwischen Null und Eins, was an das Verhalten von Determinationskoeffizienten erinnert. Tatsächlich läßt sich $conex_{ik}$ als quadrierte "Korrelation zwischen dem Zeilenprofil b_i und der k-ten Dimension" interpretieren. Wie obige Abbildung zeigt, ist $conex_{ik}$ nämlich gerade das Quadrat aus dem Kosinus des Winkels α zwischen dem Vektor b_i und der k-ten Dimension (Greenacre, 1993, S. 90).
Für unser Beispiel erhalten wir folgende Ergebnisse:

Contribution of dimensions to the inertia of each row point:

FACH		Marginal Profile	Dim 1	2	Total
1	Geologie	,107	,055	,861	,916
2	Biochemi	,036	,119	,762	,881
3	Chemie	,163	,134	,510	,644
4	Zoologie	,151	,846	,083	,929
5	Physik	,143	,880	,006	,886
6	Ingenieu	,111	,121	,749	,870
7	Mikrobio	,046	,009	,671	,680
8	Botanik	,108	,625	,029	,654
9	Statisti	,036	,554	,007	,561
10	Mathemat	,098	,240	,079	,319

Hier ist z.B. zu ersehen, daß die Profile von Zoologie und Physik überwiegend durch die erste Dimension erklärt werden, während die Profile von Geologie, Biochemie und Ingenieurwesen überwiegend durch die zweite Dimension beschrieben werden. Grundsätzlich korreliert ein Zeilenprofil natürlich relativ hoch mit *der* Hauptachse, deren Lage es stark mitbestimmt hat. Allerdings ist es möglich, daß die Position eines Profils erheblich durch eine Dimension erklärt wird, obwohl das Profil praktisch keinen Beitrag zu dieser Dimension geleistet hat. Dies wird durch das Zeilenprofil von Mikrobiologie und die Dimension Zwei demonstriert.

Wie gut können die Zeilenprofile im Unterraum *S* erklärt werden?

In der letzten Spalte der obigen Tabelle ("Total") sind die dimensionsspezifischen Anteile pro Zeilenprofil addiert. Sie gibt damit für jedes Zeilenprofil an, wie gut es insgesamt durch die zweidimensionale Lösung erklärt wird. Diese Information ist unbedingt bei der Interpretation der Distanzverhältnisse in den grafischen Darstellungen zu berücksichtigen. Z.B. wird das Mathematikprofil sehr schlecht aufgeklärt, so daß Aussagen über seinen Abstand von anderen Profilen, d.h. über seine Ähnlichkeit zu anderen Profilen, aufgrund der zweidimensionalen Lösung sehr unsicher sind.

7 Symmetrie von Zeilen- und Spaltenanalyse

Bisher haben wir uns fast ausschließlich mit der Analyse der Zeilenprofile befaßt. Die Spalten traten nur an zwei Stellen in Erscheinung:

- Die Massen bzw. Gewichte c_j der Spalten gingen wesentlich bei der Definition des χ^2-Abstandes in Abschnitt 2 ein.
- Im Abschnitt 4 haben wir die *j*-te Spaltenkategorie durch ein extremes Zeilenprofil mit vollständiger Massenkonzentration auf der Spalte *j* repräsentiert, z. B. bei *j* = 3 durch:

  $$(0 \quad 0 \quad 1 \quad 0 \quad 0)$$

  Diese Extremprofile bilden gerade die Ecken des vierdimensionalen Teilraums, in dem sich die 10 Zeilenprofile befinden. Sie spielen bei der Festlegung der Dimensionen ξ_k keine Rolle, können aber wie die normalen Zeilenprofile auf den reduzierten Raum *S* projiziert und dort gemeinsam mit den Zeilenprofilen dargestellt werden.

Völlig analog können wir aber auch die Spaltenprofile q_j untersuchen, d.h. wir können ...

- die Spaltenprofile in einem mehrdimensionalen Raum darstellen,
- einen optimalen, niedrig-dimensionalen Unterraum *S* bestimmen, um die χ^2-Abstände zwischen den Spaltenprofilen bei minimalem Informationsverlust einfach darzustellen,
- zusätzlich zu den Spaltenprofilen auch die extremen (Zeilen-)Eckpunkte auf *S* projizieren, um ihre Distanzen zu den projizierten Spaltenprofilen zu beurteilen.

Zeilen- und Spaltenanalyse erweisen sich als weitgehend äquivalent:

- Beide Probleme haben dieselbe Dimensionalität:

  $$\min(z, s) - 1$$

  In unserem Beispiel werden bei einer Spaltenanalyse zwar die fünf Spalten im 10-dimensionalen Raum der Zeilen dargestellt, doch die fünf Spaltenprofil-Vektoren q_j , j = 1,..,5, halten sich in der vierdimensionalen Hyperebene auf, die durch folgende Gleichung definiert ist:

  $$\sum_{i=1}^{10} q_{j_i} = 1$$

- Es zeigt sich, daß in beiden Analysen dieselbe Inertia (Gesamtvariation) zugrundeliegt. Ferner ergibt sich exakt dieselbe Inertia-Verteilung auf die Dimensionen.
- Bei der Spaltenanalyse erhält man Scores der Spaltenprofile auf der *k*-ten Dimension, die bis auf einen Schrumpfungsfaktor identisch sind mit den Quantifizierungen der Spalten-Ecken in der Zeilenanalyse (vgl. Abschnitt 4). Der Schrumpfungsfaktor ist gerade die Wurzel aus dem Inertia-Anteil ϕ_k der *k*-ten Dimension:

  $$\sqrt{\phi_k}$$

  In unserem Beispiel lagen die Inertia-Anteile zwischen ϕ_1=0,03912 und ϕ_4=0,00251, so daß wir also Schrumpfungseffekte von 0,19779 bis 0,05010 erhalten.
  Andererseits sind die (Zeilen-)Eck-Quantifizierungen zur der *k*-ten Dimension aus der Spaltenanalyse bis auf den Streckungsfaktor

$$\frac{1}{\sqrt{\phi_k}}$$

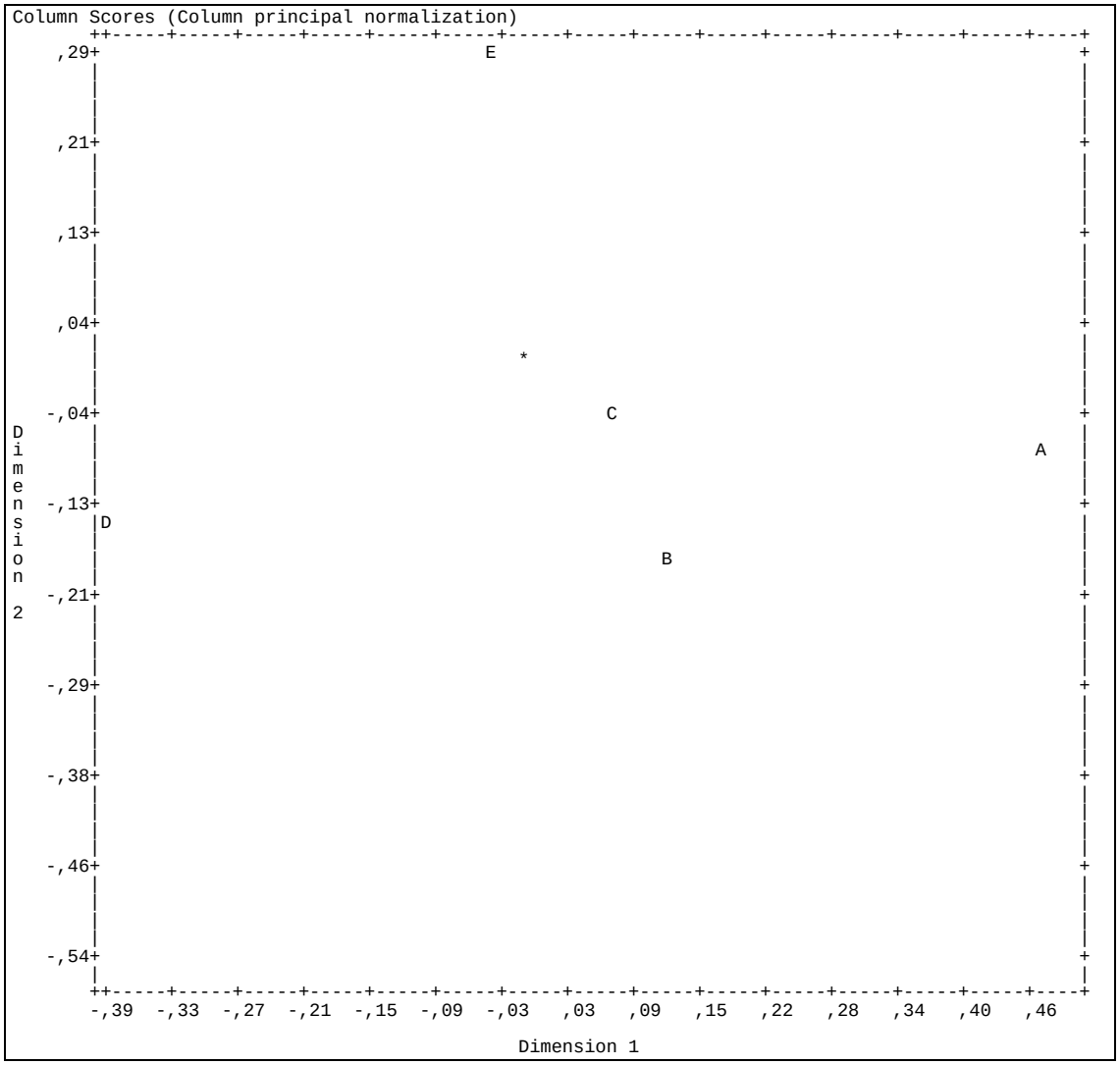
identisch mit den Scores der Zeilenprofile aus der Zeilenanalyse.
Für unser Beispiel liefert die Spaltenanalyse mit SPSS-ANACOR folgende Scores zu den Spaltenprofilen bzw. (Zeilen-)Eckpunkten:

```
Row Scores:

FACH           Marginal    Dim
                Profile       1        2

   1 Geologie     ,107     ,386   -1,736
   2 Biochemi     ,036     ,910    2,610
   3 Chemie       ,163     ,190    -,421
   4 Zoologie     ,151   -1,655    -,587
   5 Physik       ,143    1,595    -,155
   6 Ingenieu     ,111    -,594    1,674
   7 Mikrobio     ,046     ,065     ,629
   8 Botanik      ,108    -,904     ,221
   9 Statisti     ,036     ,630    -,081
  10 Mathemat     ,098     ,540     ,352

          .   .   .   .   .   .
          .   .   .   .   .   .

Column Scores:

FUND           Marginal    Dim
                Profile       1        2

   1 A            ,039     ,478    -,072
   2 B            ,161     ,127    -,173
   3 C            ,389     ,083    -,050
   4 D            ,162    -,390    -,139
   5 E            ,249    -,032     ,292
```

Für die Förderungsstufe A erhalten wir z.B. zur ersten Dimension den Score 0,478. Dies ist gerade das 0,19779-fache des korrespondierenden (Spalten-)Eck-Wertes 2,417 aus der Zeilenanalyse (vgl. Seite 16).
Bei der Spaltenanalyse erhalten wir eine räumliche Darstellung der χ^2-Abstände zwischen den Spaltenprofilen. Bei Berücksichtigung der eben erläuterten, dimensionsspezifischen Schrumpfungsfaktoren folgt daraus unmittelbar, daß die Distanzen zwischen den Spalten-*Eck*profilen, die wir bei einer Zeilenanalyse erhalten, *keine* (approximativen) χ^2-Abstände sind und daher nicht wie solche interpretiert werden dürfen. Analoges gilt natürlich für die Zeilen-*Eck*profilen, die wir bei einer Spaltenanalyse erhalten.
Zur Vermeidung von Mißverständnissen und zur Vereinfachung der Ausdrucksweise hat man für die verschiedenen Scores-Vektoren folgende Bezeichnungen eingeführt:

- Bei einer Zeilenanalyse bezeichnet man ...
  die Scores der Zeilenprofile als "Zeilen-Hauptkoordinaten",
  die Scores der Spalten-Eckpunkte als "Spalten-Standardkoordinaten".
- Bei einer Spaltenanalyse bezeichnet man ...
  die Scores der Spaltenprofile als "Spalten-Hauptkoordinaten",
  die Scores der Zeilen-Eckpunkte als "Zeilen-Standardkoordinaten".

Die Bezeichnung "Hauptkoordinaten" ist plausibel, weil es sich hierbei um Koordinaten auf den Hauptachsen der Korrespondenztabelle handelt. Die Begründung für die Bezeichnung "Standardkoordinaten" folgt in Abschnitt 8.
Die folgende Abbildung zeigt für unser Forschungsbeispiel die Spalten-Hauptkoordinaten, deren Distanzen als (approximative) χ^2-Abstände interpretiert werden können:



- Alle Aussagen in Abschnitt 0 über Inertia-Zerlegungen, Beiträge der Profile bzw. Dimensionen gelten völlig analog auch bei der Spaltenanalyse, d.h. wir können betrachten:
    - Zerlegung der Gesamtvariation (Inertia) in Beiträge der Spaltenprofile (vgl. Abschnitt 2)
    - Einfluß des *j*-ten Spaltenprofils auf die *k*-te Dimension
      Bei der Spaltenanalyse schreiben wir den Spaltenprofilen Anziehungskräfte auf die optimal auszurichtenden Hauptachsen zu, wobei der Einfluß des *j*-ten Spaltenprofils auf die Lokalisation der *k*-ten Dimension abhängt von ...
        - der Lage des *j*-ten Spaltenprofils im Raum,
        - der Masse des *j*-ten Spaltenprofils.

      Für unsere Daten zur Forschungsförderung erhalten wir folgende Ergebnisse:

```
Contribution of column points to the inertia of each dimension:

FOUND          Marginal    Dim
                Profile       1        2

   1 A             ,039     ,228     ,007
   2 B             ,161     ,067     ,159
   3 C             ,389     ,068     ,032
   4 D             ,162     ,632     ,103
   5 E             ,249     ,006     ,699
                         -------- --------
                           1,000    1,000
```

Es zeigt sich daß 63,2 % der Inertia von Dimension Eins auf die Kategorie D zurückgeht. Ein weiterer bedeutsamer Anteil von 22,8 % stammt von Kategorie A. Diese beiden Spaltenkategorien legen also wesentlich die Orientierung der ersten Dimension fest. Vermutlich handelt es sich bei der ersten Hauptachse um eine Bewertunsdimension mit den Extrempunkten A und D.
Die zweite Dimension scheint überwiegend von der Kategorie E bestimmt zu sein.

- Beitrag der *k*-ten Dimension zur Erklärung des *j*-ten Spaltenprofils
  Die Erklärungsanteile bewegen sich wie die obigen Einflußgrößen zwischen Null und Eins und lassen als quadrierte "Korrelationen zwischen den Spaltenprofilen und den Hauptachsen" interpretieren.
  Die Ergebnisse für unser Beispiel:

```
Contribution of dimensions to the inertia of each column point:

FUND           Marginal    Dim                Total
                Profile       1        2

   1 A             ,039     ,574     ,013      ,587
   2 B             ,161     ,286     ,531      ,816
   3 C             ,389     ,341     ,124      ,465
   4 D             ,162     ,859     ,109      ,968
   5 E             ,249     ,012     ,978      ,990
```

In der mit "Total" betitelten letzten Tabellenspalte sind die dimensionsspezifischen Anteile für jedes Spaltenprofil addiert. Sie gibt damit für jedes Spaltenprofil an, wie gut es insgesamt durch die zweidimensionale Lösung erklärt wird.
Wie erwartet, korreliert ein Spaltenprofil deutlich mit *der* Hauptachse, deren Lage es stark mitbestimmt hat.

8 Optimal Scaling

Bislang wurden die KA-Ergebnisse vor allem in geometrischen bzw. gravitationstheoretischen Begriffen diskutiert (z.B. Zeilenprofile als Massenpunkte mit χ^2-Abständen in einem durch die Spalten-Eckpunkte begrenzten Raum), doch sind auch interessante alternative Deutungen möglich. Dies hat dazu geführt, daß die KA-Methodologie mehrfach unter verschiedenen Bezeichnungen erfunden worden ist (Greenacre 1993, S. 48). "Optimal scaling" ist eine solche alternative Bezeichnung und Betrachtungsweise, deren Behandlung zusätzliche Einsichten in die KA vermittelt.

8.1 Skalenwerte mit optimaler Kriteriumsvarianz

Wir können uns im Forschungsbeispiel die Aufgabe stellen, die kategoriale Spaltenvariable ("Förderungsstufe") so zu quantifizieren, d.h. für ihre Kategorien A bis E Zahlen *a* bis *e* derart festzulegen, daß sich die Wissenschaftsdisziplinen bzgl. der neuen Variablen optimal unterscheiden. Dabei wird unter dem Wert w_i der Wissenschaftsdisziplin *i* auf der neuen Variablen das folgende gewichtete Mittel verstanden:

$$w_i := p_{i_1}\, a + p_{i_2}\, b + p_{i_3}\, c + p_{i_4}\, d + p_{i_5}\, e$$

Dieser Wert w_i wird *jeder* Beobachtungseinheit (= Förderungsantrag) in der Disziplin *i* zugeordnet. Gesucht sind die Quantifizierungen *a*, .., *e*, welche die Varianz der w_i-Werte über alle Beobachtungseinheiten in der Stichprobe maximieren:

$$\sum_{i=1}^{z} r_i\, (w_i - \overline{w})^2 \xrightarrow{!} \max, \quad \text{mit } \overline{w} := \sum_{i=1}^{z} r_i\, w_i$$

Um eine eindeutige Lösung identifizieren zu können, wird für die Quantifizierungen *a*, .., *e* gefordert, daß sie, angewendet auf alle Beobachtungseinheiten, einen Mittelwert von Null und eine Varianz von Eins haben.
Es zeigt sich, daß die Scores der Spalten-Eckpunkte auf der ersten Hauptachse das Problem lösen, und daß als Wissenschaftsbewertungen w_i gerade die Scores der Zeilenprofile (die Zeilen-Hauptkoordinaten) auf der ersten Hauptachse resultieren.

Damit können wir für diese Koordinaten folgende Eigenschaften festhalten:

- Sie sorgen für eine optimale Trennung zwischen den Zeilenkategorien.
- Die Scores der Spalten-Eckpunkte sind standardisiert, so daß die in Abschnitt 7 eingeführte Bezeichnung "Standardkoordinaten" sinnvoll ist.

8.2 Die Inertia-Anteile der Hauptachsen als kanonische Korrelationen

Die Optimalität der Zeilen- bzw. Spaltenscores auf der ersten Hauptachse kann noch anders charakterisiert werden. Wenn wir in unserem Forschungsbeispiel die 10 Disziplinen und die fünf Förderungsklassen beliebig quantifizieren, dann können wir die beiden resultierenden numerischen Variablen über alle Beobachtungseinheiten hinweg korrelieren. Der dabei maximal erreichbare Wert wird als "erste kanonische Korrelation" zwischen der Zeilen- und der Spaltenvariablen bezeichnet. Dies ist ein sinnvolles Maß für die Stärke des Zusammenhangs zwischen der Zeilen- und der Spaltenvariablen. Es bestehen folgende Beziehungen zu den KA-Ergebnissen:

- Die Quantifizierungen, welche die maximal mögliche Korrelation liefern, sind gerade Zeilen- und Spaltenscores auf der ersten Hauptachse, die wir bei der KA ermitteln.
- Als Wert der ersten kanonischen Korrelation erhalten wir die Wurzel aus dem Inertia-Anteil der ersten Hauptachse, in unserem Beispiel also:

$$\sqrt{\phi_1} = 0{,}19779$$

Aus Abschnitt ist uns $\sqrt{\phi_k}$ als Schrumpfungsfaktor bekannt. Offenbar zeigt der Schrumpfungseffekt beim Übergang von Standard- auf Hauptkoordinaten also direkt die Stärke des Zusammenhangs zwischen der Zeilen- und der Spaltenvariablen an.

Mit unserem jetzigen Kenntnisstand können wir auch die Behauptung aus Abschnitt 0 nachvollziehen, daß der Inertia-Anteil der *k*-ten Dimension ϕ_k zwischen Null und Eins variiert: Es ist eine quadrierte Korrelation.

Die zweite kanonische Korrelation ist definiert als maximale Korrelation zwischen einer Zeilen- und einer Spalten-Quantifizierung, die jeweils unabhängig sind von den korrespondierenden Scores auf der ersten Hauptachse. Analog sind weitere kanonische Korrelationen bis zur Ordnung min(*z*,*s*)-1 definiert. In der Ausgabe von SPSS-ANACOR erscheinen die kanonischen Korrelationen gemeinsam mit den ihren Quadraten, also den Inertia-Anteilen der Dimensionen (siehe Seite 14).

9 Literatur

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Greenacre, M.J. (1984). *Theorie and applications of orrespondence analysis*. London: Academic Press.

Greenacre, M.J. (1993). *Correspondence analysis in practice*. London: Academic Press.

Greenacre, M.J. & Blasius, J. (Eds.). (1994). *Correspondence analysis in the social sciences*. London: Academic Press.

SPSS, Inc. (1994). *Categories 6.1*. Chicago.

10 Anhang: SPSS-Syntax zu Beispiel 2

Die in diesem Manuskript enthaltenen Ausgaben zum Forschungsbeispiel wurden mit folgendem SPSS-Programm erzeugt:

```
title 'Forschungsgelder nach Wissenschatfsdisziplinen'.

data list free /fach found freq.
begin data.
 1 1  3   1 2  19  1 3  39  1 4 14  1 5 10
 2 1  1   2 2   2  2 3  13  2 4  1  2 5 12
 3 1  6   3 2  25  3 3  49  3 4 21  3 5 29
 4 1  3   4 2  15  4 3  41  4 4 35  4 5 26
 5 1 10   5 2  22  5 3  47  5 4  9  5 5 26
 6 1  3   6 2  11  6 3  25  6 4 15  6 5 34
 7 1  1   7 2   6  7 3  14  7 4  5  7 5 11
          8 2  12  8 3  34  8 4 17  8 5 23
 9 1  2   9 2   5  9 3  11  9 4  4  9 5  7
10 1  2  10 2  11 10 3  37 10 4  8 10 5 20
end data.

weight by freq.

value labels
  fach  1 'Geologie'  2 'Biochemie' 3 'Chemie'   4 'Zoologie'  5 'Physik'
        6 'Ingenieur' 7 'Mikrobiologie'  8 'Botanik' 9 'Statistik' 10 'Mathematik'
 /found  1 'A' 2 'B' 3 'C' 4 'D' 5 'E'.

crosstabs table = fach by found
 /cell = count row column /stat = chi.

set lowres = on.

anacor
 table = fach(1,10) by found(1,5)
 /normalization = rprincipal
 /print = table profiles scores contributions
 /plot = rows(15) columns(15) joint.

anacor
 table = fach(1,10) by found(1,5)
 /normalization = cprincipal
 /print = table profiles scores contributions
 /plot = rows(15) columns(15) joint.
```

Anmerkungen:

- Die KA ist in SPSS über die Prozedur ANACOR realisiert. Diese ist Teil des Zusatzmoduls **Categories.**
- Ob eine Zeilen- oder Spaltenanalyse durchgeführt werden soll, wird mit dem Subkommando "normalization" festgelegt:

  "rprincipal" ergibt Zeilen-Hauptkoordinaten und Spalten-Standardkoordinaten,

  "cprincipal" ergibt Spalten-Hauptkoordinaten und Zeilen-Standardkoordinaten.

Sie finden dieses SPSS-Programm in der Datei **FOUND.SPS** an dem im Vorwort vereinbarten Ort.