Universitäts-Rechenzentrum Trier

AWS.SPSS.13

Trier, den 30.1.1998

Hauptkomponentenanalyse für kategoriale Daten mit SPSS-HOMALS

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

Vorwort

Das Manuskript beschreibt die im SPSS-Zusatzmodul **Categories** verfügbare Hauptkomponentenanalyse für kategoriale Daten (Prozedur HOMALS). Man bestimmt dabei orthogonale Dimensionen, welche die Kategorien der manifesten kategorialen Variablen optimal separieren, also möglichst viel Unterschiedlichkeit der Kategorien enthalten.

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 > Hauptkomponentenanalyse für kategoriale Daten mit SPSS-HOMALS

Hinweise auf Unzulänglichkeiten im Manuskript werden mit Dank entgegen genommen

1 Einleitung

Der korrekte Name des vorzustellenden Verfahrens lautet eigentlich "Homogenitätsanalyse", es wird sich aber zeigen, daß man zu Recht auch die vertrautere Bezeichnung "Hauptkomponentenanalyse" verwenden darf. Historisch ist die Homogenitätsidee eng verwandt mit der Vorstellung, daß verschiedene Variablen "dasselbe messen". Bei homogenen *metrischen* Variablen sollten also nach geeigneter Normalisierung pro Fall die Variablenwerte zufällig um einen konstanten Wert streuen, so daß bei Ersetzung der Variablenwerte durch diese Konstante nur ein minimaler Informationsverlust eintritt. Eine Untersuchung der Homogenität von metrischen Variablen ist z.B. mit der altbekannten Hauptkomponentenanalyse möglich, die eine Dimensionsreduktion bei minimalem Informationsverlust anstrebt. Eine analoge multivariate Analyse von *kategorialen* Variablen erlaubt die von Gifi (1990) entwickelte und im SPSS-Modul CATEGORIES implementierte HOMALS-Analyse, die in diesem Manuskript vorgestellt werden soll.
Man bestimmt dabei orthogonale Dimensionen, welche die Kategorien der manifesten kategorialen Variablen optimal separieren, d.h. die möglichst viel Unterschiedlichkeit der Kategorien enthalten. Das Optimierungsprinzip kann für die erste Dimension auch so beschrieben werden: Man bestimmt deren Werte für alle Merkmalsträger und sucht simultan zu jeder kategorialen Variablen eine beliebige Transformation derart, daß die Summe der quadrierten Korrelationen zwischen der Dimension und den transformierten Variablen maximal wird. Für jede manifeste Variable wird ein Diskriminanzmaß bzgl. jeder Dimension berechnet, das analog zur quadrierten Ladung bei der Hauptkomponentenanalyse interpretiert werden kann. Entsprechend wird der gesamte Erklärungsbeitrag jeder Dimension durch die Summe ihrer Diskriminanzmaße charakterisiert.
Gelegentlich wird die HOMALS-Analyse auch als "Multiple Korrespondenzanalyse" bezeichnet. Bei Beschränkung auf zwei Variablen sind die HOMALS-Ergebnisse tatsächlich mit denen einer Korrespondenzanalyse vergleichbar. Während die klassische Korrespondenzanalyse (z.B. im SPSS-Modul CATEGORIES realisiert in der Prozedur ANACOR) jedoch auf *zwei* Variablen beschränkt ist, besteht ein wesentlicher HOMALS-Vorteil in der Möglichkeit zur multivariaten Datenanalyse.

Dieser Kurs basiert im Wesentlichen auf drei Informationsquellen:

- Gifi, A. (1990). *Nonlinear Multivariate Analysis*. New York: Wiley.
- Sonnemann, E. (1992). *Exploratorische Datenanalyse*. Mitschrift der Vorlesung im Sommersemester 1992.
- SPSS Inc. (1994). *SPSS Categories*. Chicago, IL.

2 Der Gifi-Ansatz in der multivariaten Datenanalyse

Eine Arbeitsgruppe von Statistikern an der niederländischen Universität Leiden hat eine Familie von statistischen Verfahren und zugehörigen Programmen zur multivariaten Analyse kategorialer und ordinaler Daten entwickelt. Weil diese Arbeitsgruppe die Gesamtdarstellung ihrer bisherigen Ergebnisse unter dem Pseudonym "Gifi" veröffentlicht hat, wird in diesem Manuskript vom "Gifi-Ansatz" gesprochen.
Das gesamte Gifi-Projekt umfaßt neben der in diesem Kurs behandelten HOMALS-Methode u.a. noch eine Homogenitätsanalyse für *ordinale* Daten, Varianten der kanonischen Korrelationsanalyse für kategoriale und ordinale Daten sowie eine Korrespondenzanalyse.

2.1 Gifis Einstellung zu einigen forschungsstrategischen Prinzipien

Vor dem Einstieg in die Homogenitätsanalyse für kategoriale Daten sollen einige zentrale forschungsstrategische Standpunkte der Gifi-Gruppe referiert werden.

i) Kritik am Normalverteilungsmodell

In der klassischen multivariaten Analyse werden üblicherweise Hypothesen über Parameter in einem Modell für kontinuierliche Variablen getestet. Bestandteil der dabei, meist ohne Test, vorausgesetzten generellen Modelleigenschaften ist in der Regel die Annahme der multivariaten Normalverteilung von beobachteten Variablen. Eine wichtige Ursache für die Bevorzugung parametrischer Modelle für kontinuierliche Variablen in den Sozialwissenschaften, insbesondere in der Psychologie, ist das Bestreben, eine exakte, modellorientierte Wissenschaft nach dem Vorbild der Physik aufzubauen (Urväter dieser Idee: Pearson, Spearman).
Gifi kritisiert, daß die Normalverteilungsannahme oft nicht plausibel ist und kaum getestet werden kann.

ii) Flexible exploratorische Datenanalyse statt mechanischer, konfirmatorischer Modelltests

Nach traditionellem Verständnis sollte empirische Forschung folgendermaßen ablaufen:

- Formulierung eines Modells aufgrund des vorhandenen Wissens
- Datenerhebung
- Mechanisch durchzuführender Modelltest, basierend auf der Likelihood der Daten unter dem Modell

Eine besonders ausgefeilte Technologie für eine modellorientierte empirische Forschung stellt z.B. der LISREL-Ansatz bzw. das LISREL-Programm bereit (Jöreskog & Sörbom 1989).
Gifi kritisiert, daß bei diesem Vorgehen **apriori-Wissen** für eine erfolgreiche Forschung vorausgesetzt wird, das häufig nicht oder nur fragmentarisch gegeben ist, z.B. bei breit angelegten sozialwissenschaftlichen Studien mit vielen Variablen. In solchen Fällen verläuft die Modellwahl teilweise willkürlich bzw. zufällig. Erfolgt die Ergebnisinterpretation dann ausschließlich in termini der Parameter des zufällig gewählten Modells, so ist sie ebenfalls teilweise willkürlich bzw. zufällig.

Gifi betont, daß der Einsatz hochgradig strukturierter Modelle unsinnig und gefährlich ist in hochgradig unstrukturierten Situationen und befürwortet statt dessen das folgende datenanalytische Vorgehen:

> Wähle eine Analysetechnik bzw. ein spezielles Computerprogramm **aufgrund des Datenformates** und analysiere den Output.

Statistische Methoden werden hier als Werkzeug aufgefaßt. Theoretisches Vorwissen und Kreativität sind dabei natürlich nicht irrelevant; sie werden jedenfalls nicht auf der Phase des Modellentwufs beschränkt.
Gifis Methoden zielen darauf ab, in sehr umfangreichen Variablensätzen dominante Zusammenhänge aufzudecken und sind nicht dazu geeignet, spezifische Fragen an die Daten zu beantworten. Hier zeigt sich Orientierung des Gifi-Projektes an den Sozialwissenschaften, wo breit angelegte Untersuchungen und nicht vollspezifizierte Fragestellungen die Regel sind. Gifi beschränkt sich auf niedrig-strukturierte Modelle und vermeidet damit das Problem der Modellselektion.

iii) Betonung der grafischen Darstellung von Beobachtungen und/oder Variablen in niedrig-dimensionalen euklidischen Räumen

Hier zeigt sich der starke Bezug zur Faktorenanalyse (vgl. Einleitung), der im Gifi-Ansatz mehr Beachtung zukommt als der regressions- bzw. varianzanalytischen Methodenfamilie.
Noch enger als zur Faktorenanalyse sind die Beziehungen zur Multidimensionalen Skalierung (MDS), da in Gifis Methoden die Variablen unter Verwendung monotoner Funktionen optimal transformiert werden, um eine niedrigdimensionale Repräsentation zu verbessern. Die Faktorenanalyse versucht, durch zusätzliche Faktoren und durch Rotation den Fit zu optimieren.

iv) Konzentration auf Methoden für ordinale und kategoriale Daten

Ein Schwerpunkt im Gifi-Projekt ist die Entwicklung von nicht-linearen Versionen gängiger multivariater Auswertungsmethoden, d.h. die Entwicklung von Techniken, deren Ergebnisse sich unter monotonen oder eineindeutigen Abbildungen nicht ändern. Gifi geht davon aus, daß alle Daten diskret sind, und daß Modelle mit kontinuierlichen Variablen gelegentlich sinnvoll sind als Approximationen, die zu einfacheren Rechenformeln führen.
Auch bei der Untersuchung der Stabilität multivariater Repräsentationen werden Methoden bevorzugt, die nur minimale Voraussetzungen benötigen (z.B. bootstrap, jackknife).

v) Geringe Bedeutung des Konzepts der Zufallsstichprobe

Das in traditionellen statistischen Analysen zentrale Begriffspaar "Zufallsstichprobe - Population" spielt in Gifis Methodologie nur eine untergeordnete Rolle.

vi) Multivariate Analyse durch simultane Betrachtung bivariater Assoziationen

Gifi analysiert multivariate Interdependenzen über *bivariate* Assoziationsmaße, wobei natürlich die Assoziationen zu allen Variablenpaaren *simultan* analysiert werden. Dieser, auch in der traditionellen multivariaten Statistik übliche, bivariate Ansatz setzt voraus, daß die bivariaten Randverteilungen alle relevanten Informationen über die Interdependenzen in der multivariaten Verteilung enthalten, was z.B. bei der multivariaten Normalverteilung stets der Fall ist.

2.2 Traditionelle Verfahren zur Analyse qualitativer Daten

Als nächstes sollen kurz zwei traditionelle Verfahren zur Analyse qualitativer Daten angesprochen werden, deren Mängel durch die Gifi-Methoden teilweise überwunden werden können.

i) χ^2-Assoziationstest

Dieses beliebte Verfahren produziert bei der in heutigen Projekten üblichen Variablenvielfalt (100 und mehr) zahllose Tabellen, wobei vor allem die Beziehungen zwischen den einzelnen Tabellen unklar bleiben. Bei der rein bivariaten Betrachtung besteht die Gefahr, daß aus "Scheinassoziationen" falsche Schlüsse gezogen werden (omitted variable error). Man ist in derselben Lage wie ein Forscher mit 100 kontinuierlichen Variablen, der nur die einzelnen Einträge in der riesigen Korrelationsmatrix diskutiert, ohne eine kompakte, datenreduzierende Beschreibung der gesamten Matrix zu entwickeln.

ii) Loglineare Methoden

Im Unterschied zu den χ^2-Assoziationstests werden hier mehrere Variablen simultan analysiert.
Problem: Für die normalerweise angestrebte inferenzstatistische Auswertung werden sehr viele Fälle benötigt. Bei 10 Variablen mit je vier Ausprägungen hat die 10-dimensionale Tabelle $4^{10} = 1048576$ Zellen. Weil für die asymptotische Statistik ca. 5 Fälle pro Zelle benötigt werden, ergibt sich für das Beispiel mit 10 Variablen die Forderung nach über 5 Millionen Beobachtungen. Folglich sind loglineare Methoden nur für kleinere Modelle mit drei bis vier Variablen geeignet.

Korrekterweise muß festgehalten werden, daß bei einer rein *deskriptiven* loglinearen Analyse keine Minimalanforderungen an die Stichprobengröße bestehen.

3 Homogenitätsanalyse bei metrischen Variablen

In diesem Abschnitt wird die Homogenitätsidee zunächst für den einfachen Fall metrischer Variablen entwickelt.

Die ($n \times m$)-Datenmatrix **H** enthalte die an n Objekten (Fällen) beobachteten Werte der Variablen $\mathbf{h_1}$, $\mathbf{h_2}$, .., $\mathbf{h_m}$:

$$\left[\mathbf{h_1} \quad \mathbf{h_2} \quad . \quad . \quad \mathbf{h_m}\right] = \begin{bmatrix} h1_1 & h2_1 & . & . & hm_1 \\ h1_2 & h2_2 & & & hm_2 \\ & & & & \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ & & & & \\ h1_n & h2_n & . & . & hm_n \end{bmatrix}$$

Die Variablen seien:

- zentriert
  Ein zentrierter Variablenvektor $\mathbf{h_j}$ hat den Mittelwert Null:

$$\frac{1}{n}\sum_{i=1}^{n} h_{j_i} \quad = 0 \tag{1}$$

- auf Länge 1 normiert
  Der Variablenvektor $\mathbf{h_j}$ hat die Länge 1, wenn gilt:

$$\mathbf{h_j'h_j} = \sum_{i=1}^{n} h_{j_i}^2 \quad = 1 \tag{2}$$

  Hier werden im wesentlichen die *Varianzen* der Variablen normiert, allerdings nicht auf den Wert Eins, sondern auf den Wert 1/n:

$$\mathrm{Var}(\mathbf{h_j}) = \frac{1}{n}\sum_{i=1}^{n}(h_{j_i} - \bar{h}_j)^2 = \frac{1}{n}\sum_{i=1}^{n} h_{j_i}^2 \quad = \frac{1}{n} \tag{3}$$

Es wird eine neue Variable **x** gesucht, die bei minimalem Verlust alle Variablen $\mathbf{h_j}$, j = 1,..,m, ersetzen kann. Bei der Auswahl der Variablen **x** spielt die gewählte Verlustfunktion eine zentrale Rolle.

3.1 Einfache Mittelwertsbildung

Bei dieser einfachen Idee wird als Ersatzvariable **x** einfach das Mittel über alle Variablen $\mathbf{h_j}$, j = 1,..,m, genommen:

$$\mathbf{x} := \frac{1}{m} \sum_{j=1}^{m} \mathbf{h_j} \tag{4}$$

Es läßt sich zeigen, daß durch dieses **x** die folgende Verlustfunktion minimiert wird:

$$\sigma(\mathbf{x}^+) \;=\; \frac{1}{m} \sum_{j=1}^{m} \mathrm{SSQ}(\mathbf{x}^+ - \mathbf{h_j}) = \frac{1}{m} \sum_{j=1}^{m}\sum_{i=1}^{n} (x_i^+ - h_{j_i})^2 \;, \quad \mathbf{x}^+ \in \mathrm{R}^n \tag{5}$$

Eine wesentliche Eigenschaft der einfachen Mittelwertsbildung besteht darin, daß in die Lösung **x** alle Ausgangsvariablen $\mathbf{h_j}$, j = 1,..,m, mit demselben Gewicht eingehen, d.h. es erfolgt keine differentielle Gewichtung der Variablen.

Als minimalen Wert der Verlustfunktion erhält man:

$$\sigma(\mathbf{x}) \;=\; \frac{1}{m}\;\sum_{j=1}^{m}\mathrm{SSQ}(\mathbf{x}\,\text{-}\,\mathbf{h_j}) = 1-\mathrm{SSQ}(\mathbf{x}) = 1-\bar{r}.. \tag{6}$$

Dabei ist $\bar{r}..$ der Mittelwert aller Korrelationen zwischen den Variablen $\mathbf{h_j}$ (unter Einschluß der trivialen Korrelationen aller Variablen mit sich selbst, $r_{jj} = 1$) (Gifi 1990, S. 85).
Für den Mittelwert $\bar{r}_{\mathbf{x},\mathbf{h}_\bullet}$ der Korrelationen $r_{\mathbf{x},\mathbf{h_j}}$ zwischen unserer Lösung **x** und den Variablen $\mathbf{h_j}$ gilt eine Beziehung, die für das weitere Vorgehen noch aufschlußreicher ist (vgl. Gifi 1990, S. 86):

$$\sigma(\mathbf{x}) \;=\; 1-\bar{r}^2_{\mathbf{x},\mathbf{h}_\bullet} \tag{7}$$

Offenbar optimiert **x** die mittlere Korrelation mit den Variablen $\mathbf{h_j}$ bzw. die Summe der Korrelationen $r_{\mathbf{x},\mathbf{h_j}}$.

Nun sind wir aber eher daran interessiert, die Summe bzw. den Mittelwert der *quadrierten* Korrelationen mit den Ausgangsvariablen $\mathbf{h_j}$ zu optimieren. Dieser letztgenannte Kriteriumswert kann größer werden als das Quadrat der mittleren Korrelation und wird daher im allgemeinen durch **x** *nicht* optimiert. Es gilt nämlich allgemein die Ungleichung (siehe z.B. Heuser, 1986, S. 98):

$$\left(\frac{1}{n}\sum_{j=1}^{m} r_{\mathbf{x}^+,\mathbf{h_j}}\right)^2 \le \frac{1}{n}\sum_{j=1}^{m} r^2_{\mathbf{x}^+,\mathbf{h_j}} \tag{8}$$

3.2 Lineare Hauptkomponentenanalyse

3.2.1 Kernidee

Damit sind wir auf die Kernidee der linearen Hauptkomponentenanalyse gestoßen. Bei diesem Ansatz zur Homogenitätsanalyse wird nach einer optimalen Variablen **x** gesucht, welche die Summe der quadrierten Korrelationen mit den Ausgangsvariablen $\mathbf{h_j}$, $j = 1,..,m$, maximiert. Diese optimale Variable **x** wird als "erste Hauptkomponente der Spalten in **H**" bezeichnet.

$$\sum_{j=1}^{m} r^2_{\mathbf{x},\mathbf{h_j}} \xrightarrow{\;!\;} \max \tag{9}$$

Weil eine Korrelation vom Mittelwert und von der Varianz der beteiligten Variablen unabhängig ist, können wir der Einfachheit halber annehmen, daß **x** zentriert und auf Einheitslänge normiert ist:

$$\sum_{i=1}^{n} x_i = 0 \tag{10}$$

$$\mathbf{x'x} \;=\; \sum_{i=1}^{n} x_i^2 \;= 1\,1 \tag{11}$$

Wenn wir eine geometrische Sichtweise anwenden und die *m* (zentrierten und auf Einheitslänge normierten) Spalten der Datenmatrix **H** als Vektoren im $\mathbb{R}^n$ betrachten, dann bestimmt die Hauptkomponentenanalyse den Lösungsvektor **x** so, daß die Projektionen der Ausgangsvariablen $\mathbf{h_j}$ auf **x** optimiert werden, genauer:

$$\sum_{j=1}^{m} \mathrm{SSQ}(\mathrm{pr}_{\mathbf{x}}(\mathbf{h_j})) \xrightarrow{\;!\;} \max \tag{12}$$

Da die quadrierte Länge eines Vektors durch die Summe seiner quadrierten Komponenten definiert ist, gibt $\mathrm{SSQ}(\mathrm{pr}_{\mathbf{x}}(\mathbf{h_j}))$ gerade die quadrierte Länge der Projektion von $\mathbf{h_j}$ auf **x** an. Der Zusammenhang zwischen der quadrierten Korrelation $r^2_{\mathbf{x},\mathbf{h_j}}$ und der quadrierten Projektionslänge $\mathrm{SSQ}(\mathrm{pr}_{\mathbf{x}}(\mathbf{h_j}))$ wird in folgenden Gleichungen beschrieben:

$$r_{\mathbf{x},\mathbf{h_j}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(h_{j_i} - \bar{h}_j)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(h_{j_i} - \bar{h}_j)^2}} = \frac{\sum_{i=1}^{n} x_i h_{j_i}}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} h_{j_i}^2}} = \sum_{i=1}^{n} x_i h_{j_i} \tag{13}$$

$$\mathrm{SSQ}(\mathrm{pr}_{\mathbf{x}}(\mathbf{h_j})) = (\sum_{i=1}^{n} x_i h_{j_i})^2 \; \mathrm{SSQ}(\mathbf{x}) = (\sum_{i=1}^{n} x_i h_{j_i})^2 = r^2_{\mathbf{x},\mathbf{h_j}} \tag{14}$$

Unter unseren Voraussetzungen ist also die Korrelation identisch mit der Projektionslänge, und beide sind gleich dem Skalarprodukt der beiden Vektoren **x** und $\mathbf{h_j}$:

$$r_{\mathbf{x},\mathbf{h_j}} = \sqrt{\mathrm{SSQ}(\mathrm{pr}_{\mathbf{x}}(\mathbf{h_j})} = \sum_{i=1}^{n} x_i h_{j_i} = \langle \mathbf{x}, \mathbf{h_j} \rangle \tag{15}$$

Zwar darf der gesuchte optimale Vektor **x** prinzipiell ein beliebiges (zentriertes und normiertes) Element aus dem $\mathbb{R}^n$ sein, doch als Hauptkomponente der Vektoren $\mathbf{h_j}$, $j = 1,..,m$, liegt **x** sicherlich im Spaltenraum von **H**, d.h. **x** ist eine lineare Funktion der Vektoren $\mathbf{h_j}$:

$$\mathbf{x} = \sum_{j=1}^{m} a_j \mathbf{h_j} \tag{16}$$

Im Vergleich zu Formel 4 ist erwähnenswert, daß in die optimale Variable im Sinne der Hauptkomponentenanalyse die Ausgangsvariablen mit *unterschiedlichen* Gewichten eingehen, wobei die Wahl der Gewichte im Rahmen eines Optimierungsverfahrens erfolgt (siehe unten).

3.2.2 Gifi-Formulierung des Optimierungsproblems

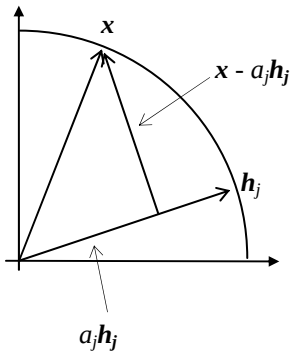
Die naheliegendste Methode zur Lösung des Homogenitätsproblemes im Sinne der Hauptkomponentenidee besteht nun darin, die Gewichte in Formel 16 so zu wählen, daß die resultierende Linearkombination **x** das Kriterium in Formel 9 optimiert, zentriert ist und Einheitslänge besitzt. Dieser Lösungsansatz kann mit der sogenannten Hauptachsentransformation realisiert werden, die z.B. in der SPSS-Faktorenanalyse verwendet wird.
Gifi bevorzugt allerdings eine alternative Vorgehensweise. Zunächst einmal formuliert er die Optimierungsaufgabe 9 in ein äquivalentes Quadratsummen-Minimierungsproblem um, das als natürliche Verallgemeinerung des Quadratsummen-Minimierungsproblems in Formel 5 (Mittelwertsidee) aufgefaßt werden kann:

$$\begin{aligned} &\sigma(\mathbf{x}^+, \mathbf{a}^+) \xrightarrow{\;!\;} \min \\ \text{mit:}\quad &\sigma(\mathbf{x}^+, \mathbf{a}^+) := \frac{1}{m} \sum_{j=1}^{m} \mathrm{SSQ}(\mathbf{x}^+ - a_j^+ \mathbf{h_j}), \quad \mathbf{x}^+ \in \mathrm{IR}^n, \; a^+ \in \mathrm{IR}^m \end{aligned} \tag{17}$$

Bei dieser Optimierungsaufgabe sind "Längen" a_j zu den gegebenen Vektoren $\mathbf{h_j}$ **und** ein Vektor **x** so zu wählen, daß die passend "abgelängten" Vektoren $a_j\mathbf{h_j}$ optimal durch **x** repräsentiert werden.
Es sind also zu bestimmen:

- Die Werte der Hauptkomponente **x**, bezeichnet als die **Scores der Objekte**.
- Die Längen a_j, bezeichnet als die **Gewichte der Variablen**, gesammelt im Vektor **a**.

Wir wollen uns zunächst überlegen, daß die Optimierungsaufgaben (9) und (17) zur selben Lösung **x** führen, also zur ersten Hauptkomponenten der Spalten in **H**. Statt, wie in den Formeln 9 bzw. 12, die Summe der quadrierten Korrelationen bzw. Projektionslängen zu maximieren, *minimiert* Formel 17 die Summe der quadrierten Abstände des Vektors **x** von den Geraden in Richtung der Vektoren $\mathbf{h_j}$. Für jeden Vektor $\mathbf{h_j}$ wird dabei eine optimale Länge a_j so bestimmt, daß der Abstand von **x** zur Geraden durch $\mathbf{h_j}$ gerade der Distanz zwischen **x** und $a_j\mathbf{h_j}$ entspricht. Dies wird in der folgenden Abbildung für eine beliebige Variable $\mathbf{h_j}$ demonstriert:



Für *jeden* festen Vektor $\mathbf{x}^+$ muß gemäß Optimierungsaufgabe (17) zu $\mathbf{h_j}$ die Länge a_j^+ quadratsummenminimierend so gewählt werden, daß $a_j^+ \mathbf{h_j}$ gerade die Projektion von $\mathbf{x}^+$ auf $\mathbf{h_j}$ ist. Daher gilt nach dem Satz des Pythagoras:

$$\mathrm{SSQ}(a_j^+ \mathbf{h_j}) = \mathrm{SSQ}(\mathbf{x}^+) - \mathrm{SSQ}(\mathbf{x}^+ - a_j^+ \mathbf{h_j}) = 1 - \mathrm{SSQ}(\mathbf{x}^+ - a_j^+ \mathbf{h_j}) \tag{18}$$

Außerdem haben wir uns schon in Formel 14 überlegt, daß unter den gegebenen Voraussetzungen die Projektionslänge mit der Korrelation identisch ist:

$$SSQ(a_j^+ \mathbf{h_j}) = \mathrm{r}^2_{\mathbf{x}^+,\mathbf{h_j}} \tag{19}$$

Dabei macht es offenbar keinen Unterschied, ob wir $\mathbf{h_j}$ auf $\mathbf{x}^+$ projizieren oder umgekehrt. Damit gilt für die Summanden in Formel 17:

$$SSQ(\mathbf{x}^+ - a_j^+ \mathbf{h_j}) = 1 - \mathrm{r}^2_{\mathbf{x}^+,\mathbf{h_j}} \tag{20}$$

Folglich ist auch nach Formel (17) der Scoresvektor **x** so zu bestimmen, daß er die Summe der quadrierten Korrelationen $\mathrm{r}^2_{\mathbf{x},\mathbf{h_j}}$ maximiert, d.h. der Lösungsvektor **x** ist die erste Hauptkomponente der Spalten von **H**.

3.2.3 Hauptkomponentenanalyse mit dem ALS-Algorithmus

Nachdem Gifi schon eine etwas ungewöhnliche Beschreibung der bei einer Hauptkomponentenanalyse zu lösenden Optimierungsaufgabe bevorzugt, wählt er zur konkreten Bestimmung der Lösung natürlich nicht die übliche Hauptachsentransformation, sondern den **ALS-Algorithmus** (Alternating Least Squares). Weil dieser Algorithmus in fast allen Gifi-Methoden und -Programmen verwendet wird, soll er in seiner Anwendung auf das vertraute Problem der Hauptkomponentenanalyse beispielhaft erläutert werden. Daß es sich bei dem aus diesem Algorithmus resultierenden Vektor **x** tatsächlich um die erste Hauptkomponente der Vektoren $\mathbf{h_j}$ handelt, wird später noch belegt.

3.2.3.1 Das Verfahren

Bei der hier vorgestellten Variante der ALS-Methode werden normierte Scores erzeugt:

$$\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2 = 12$$

Eine alternative Variante erzeugt normierte Gewichte.
Es wird vorausgesetzt, daß die Vektoren $\mathbf{h_j}$ zentriert und auf Länge Eins normiert sind. Damit ist auch der resultierende Scoresvektor $\mathbf{x} = \sum_{j=1}^{m} a_j \mathbf{h_j}$ zentriert.

Der Algorithmus startet mit einem beliebigen Gewichtungsvektor $\tilde{\mathbf{a}} \neq 0$ 3und durchläuft dann folgende Schritte:

(1) Zunächst werden zu dem aktuellen Gewichtungsvektor $\tilde{\mathbf{a}}$ 4neue Objektscores $\tilde{\mathbf{x}}$ 5 bestimmt, indem jedem Objekt der Mittelwert seiner gewichteten Variablenwerte zugewiesen wird:

$$\tilde{\mathbf{x}} := \frac{1}{m} \sum_{j=1}^{m} \tilde{a}_j \mathbf{h_j} = \frac{1}{m} \mathbf{H}\tilde{\mathbf{a}} \tag{21}$$

Dieses Mittel der gewichteten Variablen $\tilde{a}_j \mathbf{h_j}$ 6minimiert die Verlustfunktion *bei festem Gewichtungsvektor* $\tilde{\mathbf{a}}$ 7(siehe Abschnitt 3.1). $\tilde{\mathbf{x}}$ ist zentriert, weil alle Variablen $\mathbf{h_j}$ zentriert sind.

(2) Dann wird der Scoresvektor auf Länge Eins normiert:

$$\mathbf{x}^+ := \frac{\tilde{\mathbf{x}}}{\sqrt{\sum_{i=1}^{n} \tilde{x}_i^2}} \tag{22}$$

Mit $\tilde{\mathbf{x}}$ ist natürlich auch $\mathbf{x}^+$ zentriert.

(3) Nun werden neue Variablengewichte so ermittelt, daß mit dem gerade ermittelten $\mathbf{x}^+$ 8die Verlustfunktion in Formel (17) minimal wird (daher die Bezeichnung "alternierende kleinste Quadrate"). Es zeigt sich, daß dazu das neue Gewicht $\tilde{a}_j$ 9gerade identisch mit der Korrelation von $\mathbf{h_j}$ und $\mathbf{x}^+$ 10zu wählen ist. Da die Variablen $\mathbf{h_j}$ ebenso zentriert und normiert sind wie der Vektor $\mathbf{x}^+$ 11erhält man den neuen Gewichtsvektor $\mathbf{a}^+$ 12durch (vgl. Formeln 13 und 14):

$$\mathbf{a}^+ := \mathbf{H}'\mathbf{x}^+ = \begin{bmatrix} \mathbf{h'_1} \\ \mathbf{h'_2} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{h'_j} \end{bmatrix} \mathbf{x}^+ = \begin{bmatrix} h_{1_1} & h_{1_2} & \cdot & \cdot & \cdot & h_{1_n} \\ h_{2_1} & h_{2_2} & \cdot & \cdot & \cdot & h_{2_n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ h_{m_1} & h_{m_2} & \cdot & \cdot & \cdot & h_{m_n} \end{bmatrix} \begin{bmatrix} x_1^+ \\ x_2^+ \\ \cdot \\ \cdot \\ \cdot \\ x_n^+ \end{bmatrix} \tag{23}$$

(4) Falls sich die Werte von $\mathbf{a}^+$ 13und $\mathbf{x}^+$ 14noch nicht genügend stabilisiert haben, setze $\tilde{\mathbf{a}} = \mathbf{a}^+$ und15 gehe zurück zu Schritt (1).

Das Verfahren konvergiert mit den Lösungsvektoren $\mathbf{a}^+$ und $\mathbf{x}^+$, weil der Verlustwert immer kleiner wird und natürlich nach unten durch die Null beschränkt ist.

3.2.3.2 Zur Äquivalenz der ALS-Methode mit der Hauptachsentransformation

Für den per ALS-Methode bestimmten Gewichtungsvektor ***a*** kann man leicht zeigen, daß es ein (nicht-normierter) Eigenvektor zum Eigenwert $\mathbf{a}'\mathbf{a} = \sum_{j=1}^{m} a_j^2$ 16der Korrelationsmatrix $\mathbf{H}'\mathbf{H}$ zu den17 Variablen $\mathbf{h_j}$ ist. Weil sich das Verfahren als Vektoriteration mit der symmetrischen Matrix $\mathbf{H}'\mathbf{H}$ beschreiben läßt, muß dies der größte Eigenwert sein (Stoer & Bulirsch 1978, S. 44f). Der ALS-Algorithmus konvergiert also tatsächlich in die erste Hauptkomponente von **H**. Außerdem liefert der Algorithmus die Darstellung dieser Hauptkomponente als Linearkombination der Ausgangsvariablen $\mathbf{h_j}$, wobei die Gewichte bis auf gewisse Normierungen den Komponenten von **a** entsprechen (siehe Formeln 21 und 22).

3.3 Multiple Lösungen in der Linearen Hauptkomponentenanalyse

Bei der linearen Hauptkomponentenanalyse will man in der Regel an eine Matrix **H** mit zentrierten und normierten Variablen $\mathbf{h_1}$, $\mathbf{h_2}$,..,$\mathbf{h_m}$ nicht nur *eine* Hauptkomponente **x** anpassen, sondern *p* Hauptkomponenten $\mathbf{x^{(1)}}$, $\mathbf{x^{(2)}}$,.., $\mathbf{x^{(p)}}$, die paarweise orthonormal sind (alle Längen Eins, alle Korrelationen Null). Dabei wird die Hauptkomponente $\mathbf{x^{(k)}}$ im Sinne der Formel 17 an die linear transformierten Variablen

$$a_{1k}\mathbf{h_1}, a_{2k}\mathbf{h_2}, ...,.a_{mk}\mathbf{h_m}$$

angepaßt, d.h. für jede Hauptkomponente wird ein unabhängiger Satz von Gewichten zugelassen. Gesucht ist also eine Matrix **X**, welche die Hauptkomponenten bzw. Objektscores als Spalten $\mathbf{x^{(k)}}$ enthält, sowie eine Gewichtsmatrix **A**, deren Spalte $\mathbf{a^{(k)}}$ gerade die Gewichte zur *k*-ten Hauptkomponente enthält. Diese Matrizen müssen folgende Verlustfunktion minimieren:

$$\sigma(\mathbf{X},\mathbf{A}) := \sum_{k=1}^{p} \sigma(\mathbf{x^{(k)}},\mathbf{a^{(k)}}) \qquad (24)$$

unter der Nebenbedingung $\mathbf{X'X} = \mathbf{I}_p$ (≅ Orthonormalität der Scoresvektoren).

4 Maximierung der Homogenität durch nichtlineare Transformationen: Hauptkomponentenanalyse für kategoriale Daten (HOMALS)

In diesem Abschnitt gehen wir davon aus, daß die Variablen in der Datenmatrix **H** *nominal* sind, entweder "von Natur aus" oder durch künstliche Kategorisierung. Das Gifi- bzw. Categories-Programm HOMALS kann mit solchen Daten eine spezielle Homogenitäts- bzw. Hauptkomponentenanalyse durchführen. Dabei kommt eine Variante der schon in Abschnitt 3.2 beschriebenen ALS-Schätzmethode zum Einsatz, woraus sich auch der Name des Programms erklärt: "HOMALS" steht für "**HOM**ogeneity **A**nalysis by **A**lternating **L**east **S**quares".

4.1 Gifis Begriff von Homogenität bei kategorialen Variablen

4.1.1 Anwendungsbeispiel: Hartigan's Hardware

Die weiteren Überlegungen sollen anhand des Beispiels "Hartigan's Hardware" aus Gifi (1990, S. 128ff) veranschaulicht werden. An 24 Objekten aus einer Eisenwarenhandlung (Schrauben, Bolzen, Nägel und Stifte) wurden die folgenden sechs Variablen erhoben:

Variablen			Kategorien und Kodierung		
Nr.	Erläuterung	Name[1]	Kategorie	num. Code	engl. Abk.
1	Gewinde	THREADN	nein ja	1 2	N Y
2	Kopfform	HEADN	flach kegelförmig rund tassenförmig zylinderförmig	1 2 3 4 5	F O R U Y
3	Kopf-Einkerbung	INHEADN	Schlitz Keine Stern	1 2 3	L N T
4	Fußform	BOTTOMN	Flach Spitz	1 2	F S
5	Länge	LENGTH	½ Zoll 1 Zoll 1 ½ Zoll 2 Zoll 2 ½ Zoll	1 2 3 4 5	1 2 3 4 5
6	Messing	BRASS	nein ja	1 2	N Y

[1] Die Bezeichnungen der Variablen und Ausprägungen werden nur teilweise übersetzt, um konsistent mit dem Gifi-Originaltext zu bleiben.

Uns fallen spontan Begriffe aus dem Bereich Eisenwaren ein, mit deren Hilfe wir Ordnung in die Vielfalt der kleinen Teile bringen können (z.B. Schrauben, Nägel, Bolzen). Versetzen wir uns jedoch einmal in die Lage einer außerirdischen Intelligenz vom Gummiplaneten Plastillin, der solche metallischen Kleinteile noch nie gesehen hat. Er wird vielleicht versuchen, mit Hilfe einer Homogenitätsanalyse für kategoriale Variablen die wesentlichen Strukturen des Gegenstandsbereiches herauszuarbeiten. HOMALS liefert ihm zunächst eine erste **quantitative** Hauptkomponente **x** mit sogenannten Scores $x_1, x_2, ..., x_{24}$ zu den 24 Objekten. Diese Scores werden so bestimmt, daß für jede der sechs Variablen die zugehörigen Kategorien möglichst verschiedene **x**-Mittelwerte haben. Etwas einfacher ausgedrückt, wird die Hauptkomponente **x** so bestimmt, daß sie

möglichst gut zwischen allen Kategorien diskriminiert. Zusätzlich können noch weitere Hauptkomponenten zur Beschreibung des Gegenstandsbereichs bestimmt werden, die sukzessiv schwächere Diskriminanzleistungen erbringen und orthogonal zueinander sind.
Zunächst wird unser Forscher vom Mars die folgende (Objekte × Variablen) Datenmatrix **H** notieren:

$$\mathbf{H} = \begin{bmatrix} \\ \text{N F N S 1 N} \\ \text{N F N S 4 N} \\ \text{N F N S 2 N} \\ \text{N F N S 2 N} \\ \text{N F N S 2 N} \\ \text{N F N S 2 N} \\ \text{N U N S 5 N} \\ \text{N U N S 3 N} \\ \text{N U N S 3 N} \\ \text{Y O T S 5 N} \\ \text{Y R L S 4 N} \\ \text{Y Y L S 4 N} \\ \text{Y R L S 2 N} \\ \text{Y Y L S 2 N} \\ \text{Y R L F 4 N} \\ \text{Y O L F 1 N} \\ \text{Y Y L F 1 N} \\ \text{Y Y L F 1 N} \\ \text{Y Y L F 1 N} \\ \text{Y Y L F 1 N} \\ \text{N F N S 1 Y} \\ \text{N F N S 1 Y} \\ \text{N F N S 1 Y} \\ \text{Y O L S 1 Y} \end{bmatrix} \qquad \text{OBJECT} = \begin{bmatrix} \\ \text{TACK} \\ \text{NAIL1} \\ \text{NAIL2} \\ \text{NAIL3} \\ \text{NAIL4} \\ \text{NAIL5} \\ \text{NAIL6} \\ \text{NAIL7} \\ \text{NAIL8} \\ \text{SCREW1} \\ \text{SCREW2} \\ \text{SCREW3} \\ \text{SCREW4} \\ \text{SCREW5} \\ \text{BOLT1} \\ \text{BOLT2} \\ \text{BOLT3} \\ \text{BOLT4} \\ \text{BOLT5} \\ \text{BOLT6} \\ \text{TACK1} \\ \text{TACK2} \\ \text{NAILB} \\ \text{SCREWB} \end{bmatrix}$$

Neben der Matrix **H** steht noch die Variable OBJECT mit Namen zu den Objekten. Diese Variable spielt bei den Berechnungen keine Rolle, kann aber in SPSS-HOMALS-Programmausgaben zur Verbesserung der Interpretierbarkeit verwendet werden.

4.1.2 Indikatormatrizen

Während die Hauptkomponentenanalyse für metrischen Variablen direkt bei der Matrix **H** ansetzt, werden die kategorialen Variablen zunächst durch ihre **Indikatormatrizen** repräsentiert. Zu jeder Variablen $\mathbf{h_j}$ mit k_j verschiedenen Ausprägungen wird die Indikatormatrix $\mathbf{G^{(j)}}$ 18 definiert, die für jede Ausprägung eine Spalte und für jedes Objekt eine Zeile enthält. Die Zeile eines Objektes enthält eine Eins in der Spalte zur Kategorie, zu der es gehört, und sonst Nullen.

Nebeneinander gestellt ergeben die einzelnen Indikatormatrizen $\mathbf{G^{(j)}}$ 19die gesamte Indikatormatrix **G** mit $\sum_{j=1}^{m} k_j$ Spalten. In unserem Beispiel erhalten wir:

$$\mathbf{G} = \begin{bmatrix}
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 1\,0 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,0\,0\,1\,0 & 1\,0 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\
1\,0 & 0\,0\,0\,1\,0 & 0\,1\,0 & 0\,1 & 0\,0\,0\,0\,1 & 1\,0 \\
1\,0 & 0\,0\,0\,1\,0 & 0\,1\,0 & 0\,1 & 0\,0\,1\,0\,0 & 1\,0 \\
1\,0 & 0\,0\,0\,1\,0 & 0\,1\,0 & 0\,1 & 0\,0\,1\,0\,0 & 1\,0 \\
0\,1 & 0\,1\,0\,0\,0 & 0\,0\,1 & 0\,1 & 0\,0\,0\,0\,1 & 1\,0 \\
0\,1 & 0\,0\,1\,0\,0 & 1\,0\,0 & 0\,1 & 0\,0\,0\,1\,0 & 1\,0 \\
0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 0\,1 & 0\,0\,0\,1\,0 & 1\,0 \\
0\,1 & 0\,0\,1\,0\,0 & 1\,0\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\
0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\
0\,1 & 0\,0\,1\,0\,0 & 1\,0\,0 & 1\,0 & 0\,0\,0\,1\,0 & 1\,0 \\
0\,1 & 0\,1\,0\,0\,0 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\
0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\
0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\
0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\
0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1 \\
1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1 \\
0\,1 & 0\,1\,0\,0\,0 & 1\,0\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1
\end{bmatrix}$$

$$\mathbf{G^{(1)}} \quad \mathbf{G^{(2)}} \quad \mathbf{G^{(3)}} \quad \mathbf{G^{(4)}} \quad \mathbf{G^{(5)}} \quad \mathbf{G^{(6)}}$$

4.1.3 Die Minimierungsaufgabe bei der Analyse kategorialer Variablen

Diese Matrix **G** spielt bei der Homogenitätsanalyse für kategoriale Daten dieselbe Rolle wie die Datenmatrix **H** bei der Hauptkomponentenanalyse für metrische Daten. Es wird also aufgrund der Gifi-Vorliebe für Quadratsummen-Minimierungs-Formulierungen (vgl. Abschnitt 3.2) nach einer optimalen Variablen **x** gesucht, welche die Summe der quadrierten Distanzen zu den optimal "abgelängten" Spalten in **G** minimiert:

$$\sigma(\mathbf{x}^+, \mathbf{y}^+) \xrightarrow{!} \min$$

$$\text{mit:} \quad \sigma(\mathbf{x}^+, \mathbf{y}^+) := \frac{1}{m} \sum_{j=1}^{m} \sum_{\nu=1}^{k_j} \mathrm{SSQ}(\mathbf{x}^+ - y_\nu^{(j)+} \mathbf{G}_\nu^{(j)}), \quad \mathbf{x}^+ \in \mathrm{IR}^n, \; y_\nu^{(j)+} \in \mathrm{IR} \tag{25}$$

Mit $y_\nu^{(j)}$ 20 wird das Gewicht der Spalte $\mathbf{G}_\nu^{(j)}$ zur ν-ten Kategorie von Variable *j* bezeichnet und alle $y_\nu^{(j)}$ 21-Gewichte ergeben hintereinander plaziert den gesamten Gewichtungsvektor **y**, der dieselbe Rolle spielt wie der Vektor **a** in der Hauptkomponentenanalyse für metrische Daten (vgl. Abschnitt 3.2).
Die triviale Lösung **x** = **y** = 0 wird durch die Nebenbedingung Var(**x**) = 1 ausgeschlossen. Damit haben wir eine Normierung analog zu Formel (11) in Abschnitt 3.2 vorgenommen, wobei Gifi sich bei kategorialen Variablen allerdings dafür entschieden hat, die *Varianz* (statt der Vektorlänge) auf Eins zu normieren. Damit sind manche geometrische Veranschaulichungen nicht mehr so leicht möglich, andererseits sind Scores mit Varianz Eins in vielen Situationen besonders angenehm zu interpretieren. Mit **G** in der Rolle von **H** ist ein entscheidendes Argument für die Normierung von **x** auf Einheitslänge entfallen: Die Spalten von **G** haben offenbar im Unterschied zu den Spalten der Matrix **H** in Abschnitt 3.2 keine Einheitslänge. Sie sind auch nicht zentriert. Jedoch wird für **x** auch bei der Homogenitätsanalyse für kategoriale Variablen sehr wohl Zentriertheit verlangt:

$$\sum_{i=1}^{n} x_i = 0$$

Da wir uns gerade mit der Homogenitätsanalyse für **qualitative** Variablen beschäftigen, muß nachdrücklich betont werden, daß **x** für eine **quantitative** Beschreibungsdimension steht, auf der jedes Objekt durch einen metrischen Score lokalisiert ist.

4.1.4 Nichtlineare Transformationen der kategorialen Variablen

Analog zum Vorgehen in Abschnitt 3.2 wollen wir uns vor einer Beschäftigung mit dem ALS-Verfahren für kategoriale Variablen zunächst bemühen, die Gifi-Zielfunktion in Formel (25) plausibel zu machen. In diesem Zusammenhang werden wir auch endlich auf die in der Überschrift zum aktuellen Abschnitt 4.1 angesprochenen **nichtlinearen** Transformationen zur Homogenitätsmaximierung stoßen, die in den bisherigen Ausführungen noch nicht explizit erwähnt worden sind.
Wir fassen für jede Variable $\mathbf{h_j}$ die Gewichte $y_\nu^{(j)}$, ν=1,..,k_j, zu ihren Kategorien im Vektor $\mathbf{y^{(j)}}$ zusammen:

$$\mathbf{y^{(j)}} := (y_1^{(j)} \quad . \quad . \quad . \quad y_{k_j}^{(j)})'$$

Nun definieren wir zur Ausgangsvariablen $\mathbf{h_j}$ die Variable $\mathbf{q_j}$ durch die Linearkombination der Spalten von $\mathbf{G^{(j)}}$, die durch den Koeffizientenvektor $\mathbf{y^{(j)}}$ 22beschrieben wird 23:

$$\mathbf{q_j} := \mathbf{G^{(j)}}\,\mathbf{y^{(j)}} = \sum_{\nu=1}^{k_j} y_\nu^{(j)}\ \mathbf{G}_\nu^{(j)} =: \mathbf{q_j}(\mathbf{h_j}) \qquad (26)$$

Bei der Variablen $\mathbf{q_j}$ erhalten alle Objekte in Kategorie Eins der Variablen $\mathbf{h_j}$ den Wert $y_1^{(j)}$ 24, die Objekte in Kategorie Zwei den Wert $y_2^{(j)}$ 25 usw. Weil die $y_\nu^{(j)}$ beliebige Koeffizienten sind, ist $\mathbf{q_j}$ tatsächlich eine ***nichtlineare Funktion*** *von* $\mathbf{h_j}$. Wen die Bezeichnung "nichtlineare Funktion von $\mathbf{h_j}$" bei einer echt-kategorialen Variablen $\mathbf{h_j}$ stört, der kann auch von einer "beliebigen Funktion" sprechen. Nichtlinearität im traditionellen Sinne können wir in unserem Hardware-Beispiel bei $\mathbf{q_5}(\mathbf{h_5})$ beobachten, da die künstlich kategorisierte Variable Länge numerisch kodiert worden ist. Sie wird in der später vorzustellenden Lösung folgendermaßen transformiert:

$$\mathbf{h}_5 = \begin{bmatrix} 1 \\ 4 \\ 2 \\ 2 \\ 2 \\ . \\ . \\ . \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \mathbf{G}^{(5)}\mathbf{y}^{(5)} = \begin{bmatrix} 1\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0 \\ 0\,1\,0\,0\,0 \\ 0\,1\,0\,0\,0 \\ 0\,1\,0\,0\,0 \\ \vdots \\ \vdots \\ \vdots \\ 1\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0 \end{bmatrix} \begin{bmatrix} -0.34 \\ 0.44 \\ 1.25 \\ -0.60 \\ 0.31 \end{bmatrix} = \begin{bmatrix} -0.34 \\ -0.60 \\ 0.44 \\ 0.44 \\ 0.44 \\ . \\ . \\ . \\ -0.34 \\ -0.34 \\ -0.34 \\ -0.34 \\ -0.34 \end{bmatrix} = \mathbf{q}_5$$

Nachdem wir endlich die nichtlinearen Transformationen ausfindig gemacht haben, können wir die Optimierungsaufgabe in Formel (25) äquivalent in Termini der transformierten Variablen $\mathbf{q_j}$ formulieren:
Der HOMALS-Algorithmus bestimmt $\mathbf{x}$ und $\mathbf{y}$ gerade so, daß die nichtlinearen Funktionen $\mathbf{q_j}$ der Ausgangsvariablen $\mathbf{h_j}$ im euklidischen Raum minimale Abstände zu $\mathbf{x}$ haben, d.h. daß die folgende Verlustfunktion minimal wird:

$$\frac{1}{m}\sum_{j=1}^{m}\mathrm{SSQ}(\mathbf{x}^+ - \mathbf{G}^{(j)}\mathbf{y}^{(j)+}) = \frac{1}{m}\sum_{j=1}^{m}\mathrm{SSQ}(\mathbf{x}^+ - \mathbf{q}_\mathbf{j}^+) = \frac{1}{m}\sum_{j=1}^{m}\mathrm{SSQ}(\mathbf{x}^+ - \mathbf{q}_\mathbf{j}^+(\mathbf{h}_\mathbf{j})) \xrightarrow{!} \min \quad (27)$$

Der Nachweis, daß die Minimierungskriterien in den Formeln (25) und (27) tatsächlich dieselben Lösungsvektoren $\mathbf{x}$ und $\mathbf{y}$ liefern, bleibt dem Leser überlassen.

4.1.5 Kontrastmaximierung

Von der Formel (27) ist es zum Glück nur noch ein kurzer Weg zu einer sehr vertrauten und plausiblen Formulierung der Homogenitätsidee für kategoriale Variablen. Für jedes feste $\mathbf{x}^+$ wird $\sum_{j=1}^{m}\mathrm{SSQ}(\mathbf{x}^+ - \mathbf{q}_\mathbf{j}^+)$ genau dann minimal, wenn $\mathbf{y}^+$ so gewählt wird, daß für alle j=1,..,m und alle ν=1,..,k_j gilt:

> Alle Objekte in der Kategorie ν von Variable j erhalten als Wert auf der Variablen $\mathbf{q}_\mathbf{j}^+$ ihren $\mathbf{x}^+$-Mittelwert.

Ist $\mathbf{q_j}$ in diesem Sinn $\mathbf{x}^+$-bedingt quadratsummen-minimierend gebildet, dann ist $\mathrm{SSQ}(\mathbf{x}^+ - \mathbf{q}_\mathbf{j}^+)$ gerade die Fehlerquadratsumme aus der Regression von $\mathbf{x}^+$ auf $\mathbf{q}_\mathbf{j}^+$ und steht in folgendem Zusammenhang zur quadrierten Korrelation $r^2_{\mathbf{x}^+,\mathbf{q}_\mathbf{j}^+}$:

$$r^2_{\mathbf{x}^+,\mathbf{q}_\mathbf{j}^+} = 1 - \frac{\mathrm{SSQ}(\mathbf{x}^+ - \mathbf{q}_\mathbf{j}^+)}{\mathrm{SSQ}(\mathbf{x}^+)} = 1 - \mathrm{SSQ}(\mathbf{x}^+ - \mathbf{q}_\mathbf{j}^+) \quad (28)$$

Folglich ist die Minimierung der Fehlerquadratsummen $\mathrm{SSQ}(\mathbf{x}^+ - \mathbf{q}_\mathbf{j}^+)$ äquivalent mit der Suche nach demjenigen Scoresvektor, welcher die Summe der quadrierten Korrelationen mit passend nonlinear

transformierten Ausgangsvariablen maximiert. Das Kriterium in Formel (27) ist also äquivalent zum Optimierungsziel:

$$\sum_{j=1}^{m} r^2_{\mathbf{x}^+,\mathbf{q}_j^+} \xrightarrow{!} \max 27 \qquad (29)$$

Die quadrierte Korrelation $r^2_{\mathbf{x},\mathbf{q}_j}$ 28 der ersten Hauptkomponenten **x** mit der Variablen $\mathbf{q_j}$ ist gerade der Determinationskoeffizient aus der einfaktoriellen Varianzanalyse mit der unabhängigen Variablen $\mathbf{h_j}$ und der abhängigen Variablen **x**. Sie ist also ein Maß dafür, wie stark sich die Kategorien der Variablen $\mathbf{h_j}$ hinsichtlich der abhängigen Variablen **x** unterscheiden (siehe Gleichung (28)).

Die erste Hauptkomponente x wird also so bestimmt, daß über alle Variablen hinweg die Unterschiede zwischen den Kategorien maximal werden. Etwas einfacher ausgedrückt, wird die Hauptkomponente x so bestimmt, daß sie möglichst gut zwischen allen Kategorien diskriminiert.

Die Hauptkomponente **x** enthält **Quantifizierungen für die Objekte**, die von Gifi als **Objektscores** bezeichnet werden. Ihr Mittelwert und ihre Varianz sind durch die Forderung nach maximaler Summe quadrierter Korrelationen mit den Variablen $\mathbf{q_j}$ nicht determiniert. Um zu besonders leicht interpretierbaren Scores zu kommen, werden diese vom HOMALS-Programm standardisiert, d.h. auf Mittelwert Null und Varianz Eins gebracht.

Die im Verlauf der Fit-Optimierung berechneten $y_\nu^{(j)}$ 29-Gewichte für die Spalten von *G* stellen **Quantifizierungen für die Kategorien** dar. Oben wurde schon erwähnt, daß $y_\nu^{(j)}$ 30 gerade der mittlere **x**-Wert aller Objekte in Kategorie ν von Variable *j* sein muß.

4.1.6 Optimale Skalierung

Die Bestimmung von optimalen Transformationen durch Minimierung einer Verlustfunktion bezeichnet man als **optimale Skalierung** (engl.: **optimal scaling**). Dementsprechend heißen die Variablen **x**, und $\mathbf{q_j}$, j = 1,..,m, auch **optimal skaliert**. Weil **x** simultan an mehrere Variablen angepaßt wird, kann man hier sogar von einer **generalisierten optimalen Skalierung** sprechen.
Eine Eigenschaft der optimal skalierten Variablen $\mathbf{q_j}$, j = 1,..,m, soll noch festgehalten werden: Es ist die beste "Kleinst-Quadrat-Schätzung" der zentrierten Variablen **x**. Daher muß auch $\mathbf{q_j}$ den Mittelwert Null haben.

4.2 Multiple HOMALS-Lösungen

Es bietet sich an, analog zu Abschnitt 3.3 eine **multiple Lösung** mit *p* paarweise orthogonalen Scoresvektoren und jeweils optimal passenden nichtlinearen Transformationen der Spalten von **G** zu bestimmen. Wenn die Scoresvektoren als Spalten in der Matrix **X** gesammelt werden und die zugehörigen Transformationsvektoren als Spalten der Matrix **Y**, analog zu **G** in $\mathbf{Y^{(j)}}$, j=1,..,m, partitioniert, dann läßt sich die multiple Lösung bestimmen durch Minimierung der folgenden Verlustfunktion:

$$\sigma(\mathbf{X},\mathbf{Y}) := \frac{1}{m}\sum_{j=1}^{m} \mathrm{SSQ}(\mathbf{X} - \mathbf{G^{(j)}}\,\mathbf{Y^{(j)}})\,31 \qquad (30)$$

unter der Nebenbedingung $\mathbf{X'X} = n\mathbf{I}_p$ (≅ unabhängige Scoresvektoren mit Varianz Eins). Hier wird die SSQ-Funktion auf eine Matrix angewendet, wobei alle Elemente quadriert und aufaddiert werden. Dies dient nur zur Vereinfachung der Schreibweise: Wir sparen den Index für die einzelnen Scoresvektoren bzw. Spalten.

Dies ist die grundlegende Verlustfunktion für den Gifi-Ansatz. Sie wird als **HOMALS-Verlustfunktion** bezeichnet, weil sie vom HOMALS-Programm minimiert wird.

Bei vollständigen Datensätzen (keine fehlenden Werte) ist die maximal mögliche Zahl der Dimensionen das Minimum der beiden folgenden Zahlen:

- Gesamte Anzahl der Kategorien minus Anzahl der Variablen: $\sum_{j=1}^{m} k_j - m$ 32

  Dies ist Spaltenrang der Indikatormatrix **G**, falls $\sum_{j=1}^{m} k_j - m > n$ 33.

- $n - 1$

Meist beschränkt man sich auf sehr wenige Dimensionen, bei der Analyse von Hartigans Hardware z.B. auf zwei (siehe unten). Von den höheren Dimensionen sind kaum bedeutsame, interpretierbare und stabile Unterschiede zwischen den Kategorien zu erwarten. Für eine zweidimensionale Lösung (Voreinstellung bei der SPSS-Prozedur HOMALS) spricht u.a. auch die gute grafische Darstellbarkeit.
In diesem Zusammenhang soll noch vermerkt werden, daß HOMALS-Lösungen **erweiterungsinvariant** sind, d.h. beim Übergang von einer *p*-dimensionalen zu einer (*p* + 1)-dimensionalen Lösung bleiben alle Teilergebnisse für die ersten *p* Dimensionen unverändert (Gifi 1990, S. 118).

4.3 Der HOMALS-Algorithmus

Wie wir schon wissen, ist "HOMALS" ein Akronym ist für "**HOM**ogeneity analysis by **A**lternating **L**east **S**quares". Mit den ersten drei Buchstaben haben wir uns in der Spezialisierung auf kategoriale Variablen in Abschnitt 4.1 bislang ausführlich beschäftigt. Nun sollen auch die letzten drei Buchstaben behandelt werden, d.h. wir wollen den ALS-Algorithmus, den wir in Abschnitt 3.2 schon in der Spezialisierung für die lineare Hauptkomponentenanalyse kennengelernt haben, nun in der Variante zur Minimierung der HOMALS-Verlustfunktion (siehe Gleichung 30) untersuchen. Dabei beschränken wir uns wie in Abschnitt 3.2 der Einfachheit halber auf die Bestimmung der ersten Dimension.

Im Vergleich zur linearen Hauptkomponentenanalyse spielt die Matrix $\mathbf{G} = (\mathbf{G}^{(1)}, \mathbf{G}^{(2)},..., \mathbf{G}^{(m)})$ die Rolle der Datenmatrix **H,** und an Stelle des Gewichtungsvektors **a** steht nun der Vektor $\mathbf{y} = (\mathbf{y}^{(1)\prime}, \mathbf{y}^{(2)\prime},..., \mathbf{y}^{(m)\prime})'$ mit den Kategorien-Quantifizierungen. Die Matrix **G** und den Vektor **y** zu unserem Beispiel haben wir uns schon angesehen.

Der gesuchte Scoresvektor **x** soll zentriert sein und außerdem zur Vermeidung der trivialen Lösung $\mathbf{y} = \mathbf{x} = 0$ auf die Länge *n* und damit auf die *Varianz* Eins normiert werden, d.h.:

$$\sum_{i=1}^{n} x_i = 0$$

$$\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2 = n \quad \Rightarrow \quad \mathrm{Var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 1 \qquad 34$$

Startvorbereitungen für den HOMALS-Algorithmus:

Ein kurzer Blick auf die Struktur der Indikatormatrix **G** und die Verlustfunktion in Formel (27) zeigt, daß wir bei Verwendung von Einservektoren mit passender Komponentenzahl für **x** und **y** eine ebenso perfekte wie unnütze Lösung erhalten:

$$\sum_{j=1}^{m} SSQ\left(\begin{bmatrix}1\\1\\.\\.\\.\\1\\1\end{bmatrix} - G^{(j)}\begin{bmatrix}1\\1\\.\\.\\1\\1\end{bmatrix}\right) = \sum_{j=1}^{m} SSQ\left(\begin{bmatrix}1\\1\\.\\.\\.\\1\\1\end{bmatrix} - \begin{bmatrix}1\\1\\.\\.\\.\\1\\1\end{bmatrix}\right) = 0 \qquad (31)$$

Unnütz ist diese Lösung vor allem deshalb, weil die Objektscores identisch sind, so daß sie keine Kategorien-Diskrimination leisten können. Wegen des optimalen Verlustwertes Null muß der HOMALS-Algorithmus durch Wahl eines geeigneten Startvektors daran gehindert werden, die unerwünschten Lösungsvektoren anzustreben. Aus der Eigenwerttheorie weiß man, daß alle vom Einservektor linear unabhängigen Lösungen orthogonal zu ihm sein, d.h. den Mittelwert Null haben müssen (siehe Gifi, 1990, S.109ff). Außerdem kann man zeigen, daß der HOMALS-Algorithmus den Mittelwert des Objektscores-Startvektors während der iterativen Optimierung beibehält. Folglich kann man die Einserlösung erfolgreich vermeiden durch Wahl eines Startvektors mit Mittelwert Null. Hier zwingt die besondere Struktur der Indikatormatrix ***G*** also zu einer gezielten Wahl des Startvektors, während der in Abschnitt 3.2 beschriebene Algorithmus für metrische Variablen mit einem beliebigen Vektor gestartet werden kann.

Es wird also zunächst ein beliebiger zentrierter Scores-Startvektor $\hat{\mathbf{x}}$ gesucht mit:

$$\sum_{i=1}^{n} \hat{\mathbf{x}}_i = 0 \qquad \sum_{i=1}^{n} \hat{\mathbf{x}}_i^2 = n$$

Für unser Beispiel (n = 24) erfüllt z.B. folgender Startvektor $\hat{\mathbf{x}}$ diese Voraussetzungen:

$$\hat{\mathbf{x}} = (-1\ -1\ -1\ -1\ -1\ -1\ -1\ -1\ -1\ -1\ -1\ -1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1)'$$

Dann werden die zugehörigen Kategorien-Quantifizierungen $\tilde{\mathbf{y}}$ 36so bestimmt, daß die HOMALS-Verlustfunktion für das gegebene $\hat{\mathbf{x}}$ bedingt minimiert wird. Wir haben uns schon in Abschnitt 4.1 überlegt, daß man dazu einfach jeder Kategorie den mittleren Score aus allen in ihr enthaltenen Objekten zuordnen muß. Um diese einfache Lösung bequem in Matrixschreibweise notieren zu können, wird noch die Diagonalmatrix **D** benötigt, deren Diagonalelemente gerade die Fallzahlen zu den einzelnen Kategorien enthalten:

$$\mathbf{D} := \text{diag}(\mathbf{G}'\mathbf{G})$$

Für Hartigans Hardware ergibt sich:

$$\mathbf{D} = \mathrm{diag}\left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ & & & & & & & & & & . & . & . & . & & & & & & & & & & \\ & & & & & & & & & & . & . & . & . & & & & & & & & & & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 0 & 1 & . & . & . & . & 1 & 0 \\ 1 & 0 & . & . & . & . & 0 & 1 \\ 1 & 0 & . & . & . & . & 0 & 1 \\ 1 & 0 & . & . & . & . & 0 & 1 \\ 0 & 1 & . & . & . & . & 0 & 1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & . & . & . & & & & & & & & \\ & & & & & & & & . & . & . & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

Den $\tilde{\mathbf{y}}$ -Startvektor, dessen Komponenten gerade die $\hat{\mathbf{x}}$ -Mittelwerte in den Kategorien sind, erhalten wir also durch:

$$\tilde{\mathbf{y}} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{x}$$

In unserem Beispiel ergibt sich:

$$\tilde{\mathbf{y}} = \begin{bmatrix} \frac{1}{12} & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{12} & \dots & 0 & 0 \\ & & \dots & & \\ 0 & 0 & \dots & \frac{1}{20} & 0 \\ 0 & 0 & \dots & & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ & & & & & & & & & . & . & . & . & . & & & & & & & & & \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Nach diesen Startvorbereitungen durchläuft der HOMALS-Algorithmus folgende Schritte:

(1) Berechnung neuer Objekt-Scores aufgrund der aktuellen Kategorien-Quantifizierungen

Der neue Scores-Vektor $\tilde{\mathbf{x}}$ 37wird so bestimmt, daß die HOMALS-Verlustfunktion bei den gegebenen aktuellen Kategorien-Quantifizierungen $\tilde{\mathbf{y}}$ 38minimal wird. Wie wir schon seit Abschnitt 3.1 über die Mittelwertsbildung zur Homogenisierung *metrischer* Variablen wissen, ist dazu für jedes Objekt der Mittelwert seiner Werte in den Vektoren $\mathbf{G^{(j)}}\tilde{\mathbf{y}}^{(j)}$, j = 1,..,m, zu bilden. Dies ist gerade die mittlere Quantifizierung aller Kategorien, denen das Objekt angehört:

$$\tilde{\mathbf{x}} \;=\; \frac{1}{m}\mathbf{G}\tilde{\mathbf{y}} = \frac{1}{m}\sum_{j=1}^{m}\tilde{\mathbf{q}}_{\mathbf{j}}, \quad \text{mit: } \tilde{\mathbf{q}}_{\mathbf{j}} := \mathbf{G}^{(j)}\tilde{\mathbf{y}}^{(j)}$$

In unserm Beispiel:

$$\tilde{\mathbf{x}} = \frac{1}{6}\begin{bmatrix} 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 1\,0 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,0\,0\,1\,0 & 1\,0 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\ 1\,0 & 0\,0\,0\,1\,0 & 0\,1\,0 & 0\,1 & 0\,0\,0\,0\,1 & 1\,0 \\ 1\,0 & 0\,0\,0\,1\,0 & 0\,1\,0 & 0\,1 & 0\,0\,1\,0\,0 & 1\,0 \\ 1\,0 & 0\,0\,0\,1\,0 & 0\,1\,0 & 0\,1 & 0\,0\,1\,0\,0 & 1\,0 \\ 0\,1 & 0\,1\,0\,0\,0 & 0\,0\,1 & 0\,1 & 0\,0\,0\,0\,1 & 1\,0 \\ 0\,1 & 0\,0\,1\,0\,0 & 1\,0\,0 & 0\,1 & 0\,0\,0\,1\,0 & 1\,0 \\ 0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 0\,1 & 0\,0\,0\,1\,0 & 1\,0 \\ 0\,1 & 0\,0\,1\,0\,0 & 1\,0\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\ 0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 0\,1 & 0\,1\,0\,0\,0 & 1\,0 \\ 0\,1 & 0\,0\,1\,0\,0 & 1\,0\,0 & 1\,0 & 0\,0\,0\,1\,0 & 1\,0 \\ 0\,1 & 0\,1\,0\,0\,0 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\ 0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\ 0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\ 0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\ 0\,1 & 0\,0\,0\,0\,1 & 1\,0\,0 & 1\,0 & 1\,0\,0\,0\,0 & 1\,0 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1 \\ 1\,0 & 1\,0\,0\,0\,0 & 0\,1\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1 \\ 0\,1 & 0\,1\,0\,0\,0 & 1\,0\,0 & 0\,1 & 1\,0\,0\,0\,0 & 0\,1 \end{bmatrix}\begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \\ y_1^{(2)} \\ y_2^{(2)} \\ y_3^{(2)} \\ y_4^{(2)} \\ y_5^{(2)} \\ y_1^{(3)} \\ y_2^{(3)} \\ y_3^{(3)} \\ y_1^{(4)} \\ y_2^{(4)} \\ y_1^{(5)} \\ y_2^{(5)} \\ y_3^{(5)} \\ y_4^{(5)} \\ y_5^{(5)} \\ y_1^{(6)} \\ y_2^{(6)} \end{bmatrix}$$

(2) Normierung des neuen Scores-Vektors auf Länge *n* bzw. Varianz 1

$$\mathbf{x}^{+} = \sqrt{n}\,\tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-\frac{1}{2}} = \sqrt{\frac{n}{\sum_{i=1}^{n}\tilde{x}_i^2}}\,\tilde{\mathbf{x}} \qquad (32)$$

(3) Berechnung neuer Kategorien-Quantifizierungen

Wie schon unter "Startvorbereitungen" diskutiert, ist als neue Quantifizierung einer Kategorie gerade der mittlere Score aller darin enthaltenen Objekte zu verwenden:

$$\mathbf{y}^{+} = \mathbf{D}^{-1}\mathbf{G}'\,\mathbf{x}^{+} \tag{33}$$

(4) Konvergenztest

Falls sich die Werte von $\mathbf{x}^{+}$ und $\mathbf{y}^{+}$ noch nicht genügend stabilisiert haben, setze $\tilde{\mathbf{y}} := \mathbf{y}^{+}$ und gehe zurück zu Schritt (1). Ansonsten sind $\mathbf{x}^{+}$ und $\mathbf{y}^{+}$ die gesuchten Lösungen.

4.4 Anmerkungen zum HOMALS-Algorithmus

Es folgen einige Anmerkungen zum ALS-Algorithmus für kategoriale Daten, die sich vor allem mit Unterschieden zum ALS-Algorithmus für metrische Daten beschäftigen. Die Hinweise ii) und folgende sind sehr technischer Art und können vom anwendungsorientierten Leser übersprungen werden.

i) Reziprokes Mitteln und Proportionalitätseigenschaften der HOMALS-Lösung

Wie wir eben gesehen haben, kann das ALS-Verfahren bei Anwendung auf die Indikatormatrix **G** zurecht als **reziproke Mittelwertsbildung** bezeichnet werden. Dieser Name wird gelegentlich in der Literatur verwendet (engl.: "reciprocal averaging").
Wenn Konvergenz erreicht ist, gelten aufgrund dieser reziproken Mittelwertsbildung für die beiden Lösungsvektoren **x** und **y** folgende Proportionalitätseigenschaften:

$$\mathbf{x} \propto \frac{1}{m}\sum_{j=1}^{m}\mathbf{G}^{(j)}\mathbf{y}^{(j)} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{q}_j \tag{34}$$

$$\mathbf{y} = \mathbf{D}^{-1}\mathbf{G}'\,\mathbf{x} \tag{35}$$

Die Formel (34) besagt, daß der Score zu jedem Objekt proportional ist zur mittleren Quantifizierung aller Kategorien, zu denen das Objekt gehört. Nach Gleichung (35) ist die Quantifizierung einer Kategorie identisch mit dem mittleren Score aller Objekte in dieser Kategorie.

ii) Zur Optimalität der ALS-Lösung

In Abschnitt 3.2 haben wir den Nachweis, daß der ALS-Algorithmus tatsächlich das optimale Paar aus einem Scores- und einem Gewichtsvektor findet, durch matrixalgebraische Argumente erbracht. In der aktuellen, etwas komplizierteren Situation, soll der Hinweis auf ein analytisches Argument genügen: Setzt man die partiellen Ableitungen der Verlustfunktion in Formel (27) nach den Komponenten von **x** und **y** gleich Null, so resultierten gerade die Proportionalitätsbedingungen in den Formeln (34) und (35) als Gleichungen. Damit erfüllen also die ALS-Lösungen, abgesehen von gewissen Normalisierungs-Faktoren, die notwendigen Bedingungen für Extremwerte (Gifi, 1990, S. 106). Aufgrund der einfachen Gestalt von Formel (27) als Summe quadrierter Terme sind diese Bedingungen auch hinreichend.

*iii) Die Spalten von **G** sind nicht varianzhomogen*

Im ALS-Algorithmus für metrische Variablen (vgl. Abschnitt 3.2) wurde vorausgesetzt, daß die Spalten der Datenmatrix **H** auf Länge Eins normiert sind, was bei der Matrix **G**, die bei kategorialen Variablen die Rolle von **H** übernimmt, nicht der Fall ist. Diese Varianzheterogenität wird im HOMALS-Algorithmus korrigiert, indem in Schritt (3) bei der Bestimmung neuer Gewichte mit der Matrix $\mathbf{D}^{-1}$ vormultipliziert wird (vgl. Formel (23)). Diese enthält auf der Hauptdiagonalen gerade die quadrierten und invertierten Längen der **G**-Spalten (siehe Gifi, S. 95 und S. 197). Damit wird statt der Matrix **G** implizit die Matrix $\mathbf{GD}^{-1/2}$ analysiert, deren Spalten auf Einheitslänge normiert sind. Für eine beliebige Spalte $\mathbf{d}_v^{(j)-½}\mathbf{G}_v^{(j)}$ dieser Matrix kann das optimale Gewicht wie in Formel (23) über das Skalarprodukt mit **x** ermittelt werden:

$$\left\langle \mathbf{d}_\nu^{(j)-½}\mathbf{G}_\nu^{(j)}, \mathbf{x} \right\rangle = \mathbf{d}_\nu^{(j)-½}\left\langle \mathbf{G}_\nu^{(j)}, \mathbf{x} \right\rangle$$

Dies ist der passende Längenfaktor für $\mathbf{d}_\nu^{(j)-½}\mathbf{G}_\nu^{(j)}$. Daraus ergibt sich für $\mathbf{G}_\nu^{(j)}$ der Faktor $\mathbf{d}_\nu^{(j)-1}\left\langle \mathbf{G}_\nu^{(j)}, \mathbf{x} \right\rangle$. Dementsprechend tritt im HOMALS-Schritt (3) der Vorfaktor $\mathbf{D^{-1}}$ auf.

*iv) Die Spalten von **G** sind nicht zentriert*

Bei der Homogenitätsanalyse für metrische Daten wurde mit gutem Grund vorausgesetzt, daß die Spalten der Datenmatrix ***H*** zentriert sind: Lageunterschiede zwischen den Variablen könnten durch die einzig erlaubten Längenanpassungen über die Gewichte a_j nicht ausgeglichen werden. Bei der Homogenitätsanalyse für kategoriale Variablen ist hingegen für jede Variable eine beliebige Transformation erlaubt. Wenn ohnehin für jede Kategorie ein eigener Parameter verfügbar ist, dann ist offenbar eine vorherige Zentrierung ohne Belang für die erreichbare Homogenität und damit überflüssig.

4.5 Das HOMALS-Datenchema für kategoriale Variablen im Überblick

Nach den langen, technisch geprägten Ausführungen sollen in diesem Abschnitt die aus Anwendersicht wesentlichen HOMALS-Konzepte an unserem Hardware-Beispiel wiederholt werden.
Die HOMALS-Methode ermittelt Quantifizierungen $\mathbf{q_j}$ der sechs kategorialen Variablen und einen Vektor **x** mit Objektscores derart, daß die Summe der quadrierten Korrelationen zwischen **x** und den $\mathbf{q_j}$ maximal wird. Der Vektor $\mathbf{q_1}$ enthält z.B. für jedes Objekt in Kategorie Eins von Variable Eins den Wert $y_1^{(1)}$, für jedes Objekt aus der zweiten Kategorie den Wert $y_2^{(1)}$ u.s.w. Folglich können die Vektoren $\mathbf{q_j}$ als nichtlineare Funktionen der Ausgangsvariablen $\mathbf{h_j}$ aufgefaßt werden. In der folgenden Tabelle sind die Ergebnisse für die erste Dimension aus unseren Beispieldaten zusammengefaßt.

Objekt	Objekt-Scores	THREADN Kat.	THREADN Quant.	HEADN Kategorien	HEADN Qu.	INHEADN Kat.	INHEADN Qu.	BOTTOMN Kat.	BOTTOMN Qu.	LENGTH Kategorien	LENGTH Quant.	BRASS Kat.	BRASS Quant.
	x	N Y	$\mathbf{q_1}$	F U O R Y	$\mathbf{q_2}$	N T L	$\mathbf{q_3}$	S F	$\mathbf{q_4}$	1 2 3 4 5	$\mathbf{q_5}$	N Y	$\mathbf{q_6}$
1	0,75	1 0	0,96	1 0 0 0 0	0,90	0 1 0	0,96	0 1	0,43	1 0 0 0 0	-0,34	1 0	-0,11
2	0,68	1 0	0,96	1 0 0 0 0	0,90	0 1 0	0,96	0 1	0,43	0 0 0 1 0	-0,60	1 0	-0,11
3	0,96	1 0	0,96	1 0 0 0 0	0,90	0 1 0	0,96	0 1	0,43	0 1 0 0 0	0,44	1 0	-0,11
4	0,96	1 0	0,96	1 0 0 0 0	0,90	0 1 0	0,96	0 1	0,43	0 1 0 0 0	0,44	1 0	-0,11
.	.		.	. . . . .	.	. . .	.	. .	.	. . . . .	.	. .	.
.	.		.	. . . . .	.	. . .	.	. .	.	. . . . .	.	. .	.
.	.		.	. . . . .	.	. . .	.	. .	.	. . . . .	.	. .	.
21	0,93	1 0	0,96	1 0 0 0 0	0,90	0 1 0	0,96	0 1	0,43	1 0 0 0 0	-,34	0 1	0,56
22	0,93	1 0	0,96	1 0 0 0 0	0,90	0 1 0	0,96	0 1	0,43	1 0 0 0 0	-,34	0 1	0,56
23	0,93	1 0	0,96	1 0 0 0 0	0,90	0 1 0	0,96	0 1	0,43	1 0 0 0 0	-,34	0 1	0,56
24	-0,54	0 1	-0,96	0 1 0 0 0	0,44	1 0 0	-1,02	0 1	0,43	1 0 0 0 0	-,34	0 1	0,56

4.6 Die HOMALS-Ergebnisse

4.6.1 Anforderung einer Homogenitätsanalyse in SPSS

In diesem Abschnitt wird gezeigt, wie die HOMALS-Ergebnisse zu den oben beschriebenen Beispieldaten mit SPSS 6.1 (für Windows, Macintosh oder UNIX) per Menüsystem erzeugt werden können.[1] Die Daten befinden sich in der SPSS-Datei **HARTIG.SAV** an der im Vorwort vereinbarten Stelle.
Alle Variablen sind trotz Nominalskalenqualität vom SPSS-Variablentyp numerisch, weil die HOMALS-Prozedur Variablen in anderer Kodierung nicht verarbeiten kann.

Gehen Sie folgendermaßen vor:

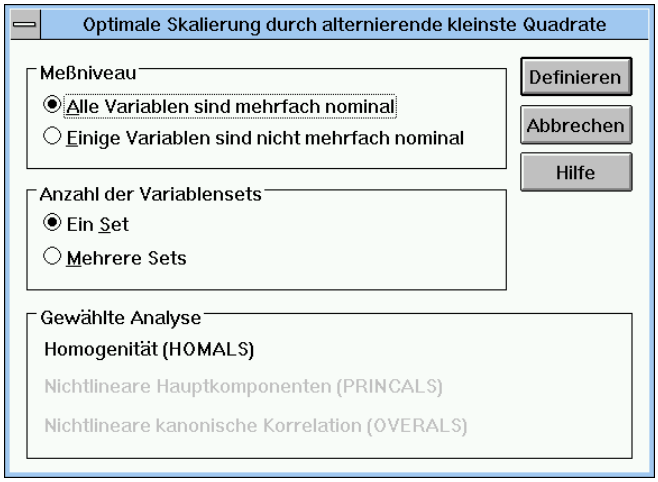
- Öffnen Sie die angegebene Datei mit

  Datei > Öffnen > Daten...

- Rufen Sie dann mit

  Statistik > Datenreduktion > Optimale Skalierung...

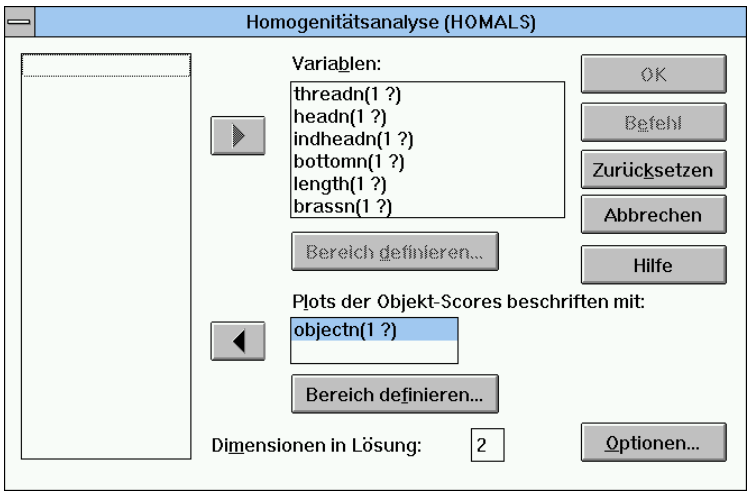
die Dialogbox **Optimale Skalierung durch alternierende kleinste Quadrate** auf:



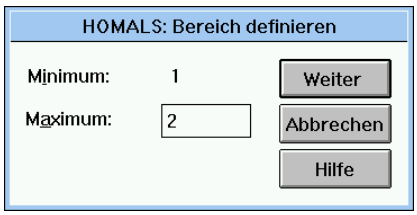
- Behalten Sie hier alle Voreinstellungen bei (**Alle Variablen sind mehrfach nominal**, **Ein Set**) und fordern Sie mit **Definieren** die HOMALS-Dialogbox an.

- Wählen Sie zur Analyse die Variablen THREADN, HEADN, INDHEADN, BOTTOMN, LENGTH und BRASSN sowie zur Objektbeschriftung die Variable OBJECTN aus:

[1] Die in Abschnitt 4.6 dargestellten Ergebnisse können alternativ auch mit einem SPSS-Programm erzeugt werden, das in der Datei **HARTIG.SPS** an der im Vorwort vereinbarten Stelle zu finden ist.



- Legen Sie nun für alle ausgewählten Variablen den maximalen Wert fest, z.B. für THREADN:



- Belassen Sie es bei der voreingestellten zweidimensionalen Lösung.
- Wählen Sie in der Optionen-Subdialogbox über die voreingestellten Ausgaben hinaus noch die Anzeige von Objektscores und die grafische Darstellung der Diskriminanzmaße.
- Lassen Sie die Analyse durchführen. Die Ergebnisse werden in den nächsten Abschnitten besprochen.

4.6.2 Objekt-Scores und Kategorien-Quantifizierungen

Wie oben erläutert gilt für die durch den HOMALS-Algorithmus bzw. durch das HOMALS-Programm ermittelten **Objekt-Scores**:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \;=\; \bar{\mathbf{x}} = 0$$

$$\mathbf{x}'\mathbf{x} = n \;\Leftrightarrow\; \frac{1}{n}\sum_{i=1}^{n} x_i^2 \;=\; \mathrm{Var}(\mathbf{x}) = 1$$

Aufgrund dieser Standardisierung können die Objektscores analog zu z-Werten interpretiert werden.

Für die **Kategorien-Quantifizierungen** zur Variablen $\mathbf{h}_j$ gilt:

$$\mathbf{y}^{(j)} = \mathbf{D}^{(j)-1}\mathbf{G}^{(j)}\mathbf{x} \qquad (36)$$

Also ist der Wert zu Kategorie ν von Variable $\mathbf{h}_j$ gerade der mittlere Score aller Objekte in dieser Kategorie. Damit sind die $\mathbf{y}$-Werte übrigens *nicht* z-standardisiert. Ihr Variationsbereich hängt von der Verteilung der Variablen $\mathbf{h}_j$ ab: Er ist bei stark besetzten Kategorien kleiner als bei schwach besetzten.

In der zweidimensionalen Lösung zu Hartigans Hardware erhalten wir folgende Objektscores bzw. Kategorien-Quantifizierungen, deren Interpretation erst im Zusammenhang mit den HOMALS-Plots erfolgen soll:

The object scores are:

Object	dimension 1	dimension 2
1	,75	,46
2	,68	,47
3	,96	,52
4	,96	,52
5	,96	,52
6	,96	,52
7	1,00	-1,69
8	1,25	-,74
9	1,25	-,74
10	-,38	-3,96
11	-,85	,23
12	-,91	,26
13	-,57	,28
14	-,63	,31
15	-1,31	,38
16	-1,18	-,51
17	-1,30	,40
18	-1,30	,40
19	-1,30	,40
20	-1,30	,40
21	,93	,67
22	,93	,67
23	,93	,67
24	-,54	-,45

In der folgenden Tabelle mit Kategorien-Quantifizierungen stehen jeweils links die Werte für die erste Dimension und rechts die Werte zur zweiten Dimension:

The labels in the plot correspond to the category values in the following way:

Variable THREADN			
No_Thread	=	,96	,15 = No_Thread
Yes Thread	=	-,96	-,15 = Yes Thread
Variable HEADN			
FLAT	=	,90	,56 = FLAT
CONE	=	-,70	-1,64 = CONE
ROUND	=	-,91	,30 = ROUND
CUP	=	1,16	-1,05 = CUP
CYLINDER	=	-1,12	,36 = CYLINDER
Variable INDHEADN			
SLIT	=	-1,02	,19 = SLIT
NONE	=	,96	,15 = NONE
STAR	=	-,38	-3,96 = STAR
Variable BOTTOMN			
flat	=	-1,28	,25 = flat
sharp	=	,43	-,08 = sharp
Variable LENGTH			
1/2 in	=	-,34	,31 = 1/2 in
1 in	=	,44	,44 = 1 in
1 1/2 in	=	1,25	-,74 = 1 1/2 in
2_in	=	-,60	,34 = 2_in
2_1/2 in	=	,31	-2,82 = 2_1/2 in
Variable BRASSN			
NotBr	=	-,11	-,08 = NotBr
YesBr	=	,56	,39 = YesBr

4.6.3 Diskriminanzmaße und Eigenwerte

4.6.3.1 Diskriminanz einer Variablen in Bezug auf eine Dimension

Gifi (1990, S. 113f) definiert für jede Dimension *s* mit Scoresvektor $\mathbf{x}^{(s)}$ und jede Variable $\mathbf{h}_j$ mit Kategorien-Quantifizierungen $\mathbf{y}^{(js)}$ 39bei Dimension *s* das folgende **Diskriminanzmaß** η^2_{js} :40

$$\eta^2_{js} := \frac{1}{n}\mathbf{y}^{(js)}\mathbf{D}^{(j)}\mathbf{y}^{(js)} \tag{37}$$

Man kann leicht zeigen, daß η^2_{js} identisch ist mit der quadrierten Korrelation zwischen dem Scoresvektor $\mathbf{x}^{(s)}$ und der optimal skalierten Variablen $\mathbf{q}^{(js)} = \mathbf{G}^{(j)}\mathbf{y}^{(js)}$ 41. Der allgemeine Ausdruck für die quadrierte Korrelation vereinfacht sich rasch, weil $\mathbf{x}^{(s)}$ und $\mathbf{q}^{(js)}$ zentriert sind, und $\mathbf{x}^{(s)}$ die Varianz Eins hat:

$$r^2_{\mathbf{x}^{(s)},\mathbf{q}^{(js)}} = \frac{\left(\sum_{i=1}^{n}(x_i^{(s)} - \bar{x}^{(s)})(q_i^{(js)} - \bar{q}^{(js)})\right)^2}{\sum_{i=1}^{n}(x_i^{(s)} - \bar{x}^{(s)})^2 \sum_{i=1}^{n}(q_i^{(js)} - \bar{q}^{(js)})^2} = \frac{\left(\sum_{i=1}^{n} x_i^{(s)} q_i^{(js)}\right)^2}{n\sum_{i=1}^{n}(q_i^{(js)})^2} = \frac{\left(\mathbf{x}^{(s)\prime}\mathbf{G}^{(j)}\mathbf{y}^{(js)}\right)^2}{n\left(\mathbf{y}^{(js)\prime}\mathbf{G}^{(j)\prime}\mathbf{G}^{(j)}\mathbf{y}^{(js)}\right)}$$

Wegen $\mathbf{G}^{(j)\prime}\mathbf{G}^{(j)} = \mathbf{D}^{(j)}$ und $\mathbf{D}^{(j)}\mathbf{y}^{(js)} = \mathbf{G}^{(j)\prime}\mathbf{x}^{(s)}$ (siehe z.B. Gleichung (36)) gilt:

$$r^2_{\mathbf{x}^{(s)},\mathbf{q}^{(js)}} = \frac{\left(\mathbf{x}^{(s)\prime}\mathbf{G}^{(j)}\mathbf{y}^{(js)}\right)^2}{n\left(\mathbf{y}^{(js)\prime}\mathbf{G}^{(j)\prime}\mathbf{G}^{(j)}\mathbf{y}^{(js)}\right)} = \frac{\left(\mathbf{y}^{(js)\prime}\mathbf{D}^{(j)}\mathbf{y}^{(js)}\right)^2}{n\left(\mathbf{y}^{(js)\prime}\mathbf{D}^{(j)}\mathbf{y}^{(js)}\right)} = \frac{\left(\mathbf{y}^{(js)\prime}\mathbf{D}^{(j)}\mathbf{y}^{(js)}\right)}{n} = \eta^2_{js}$$

Mit diesen quadrierten Korrelationen haben wir uns schon in Abschnitt 4.1 im Zusammenhang mit Gifis Homogenitätsbegriff bei kategorialen Variablen ausführlich beschäftigt. Das Diskriminanzmaß der Variablen $\mathbf{h_j}$ für die Dimension *s* wird Null, wenn der mittlere $\mathbf{x}^{(s)}$-Score in allen Kategorien von $\mathbf{h_j}$ identisch ist und damit zwangsläufig mit dem globalen $\mathbf{x}^{(s)}$-Mittel (= Null) übereinstimmt. In diesem Fall können die $\mathbf{h_j}$-Kategorien mit den $\mathbf{x}^{(s)}$-Scores nicht separiert werden.

η^2_{js} 42erreicht den optimalen Wert Eins, wenn in jeder Kategorie von $\mathbf{h_j}$ alle enthaltenen Fälle denselben $\mathbf{x}^{(s)}$-Wert haben, und die Kategorien-Mittelwerte nicht alle gleich sind.

Falls fehlende Werte auftreten, kann η^2_{js} 43allerdings auch Werte größer als Eins erreichen (Gifi 1990).

Bei der Interpretation der Diskriminanzmaße sollte noch folgender Inflationierungseffekt berücksichtigt werden:

η^2_{js} ist die quadrierte Korrelation zwischen $\mathbf{q_j}$ und $\alpha\sum_{j=1}^{m}\mathbf{q}^{(js)}$, $\alpha \in \mathrm{R}$. Daher ist auch bei völliger Unabhängigkeit der $\mathbf{q}^{(js)}$, *j* = 1,..,*m*, also bei vollständiger Heterogenität (gegebenenfalls nach Berücksichtigung von Dimensionen mit niedrigerer Ordnung), der Erwartungswert für η^2_{js} nicht Null, sondern 1/*m* (Greenacre 1993, S. 154).

4.6.3.2 Eigenwerte der Dimensionen

Offenbar kann das Diskriminanzmaß η^2_{js} auch 44analog zur quadrierten Faktorladung in der linearen Hauptkomponentenanalyse interpretiert werden (vgl. Gifi 1990, S. 120ff).

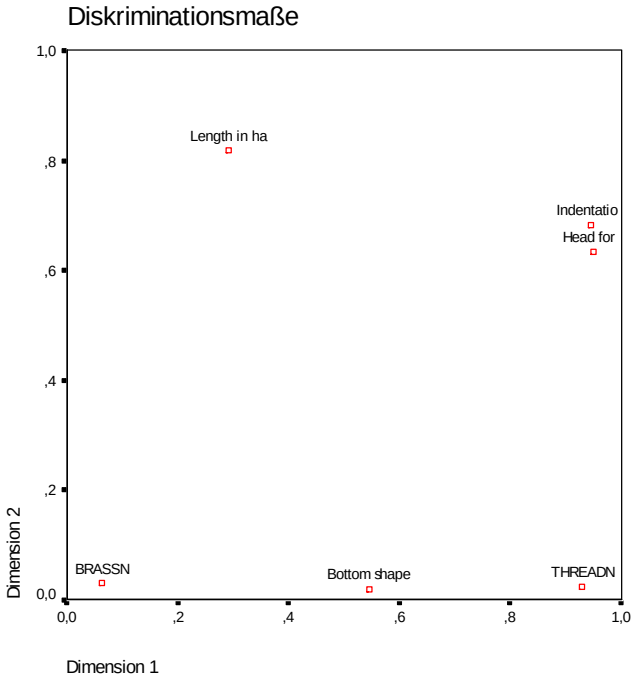
Wie wir in Abschnitt 4.1 festgestellt haben, sind $\mathbf{x}^{(s)}$ und $\mathbf{y}^{(js)}$ so zu bestimmen, daß die Summe bzw. der Mittelwert der Diskriminanzmaße η^2_{js} maximal wird. Das erreichte Maximum wird in der HOMALS-Programmausgabe als der **Eigenwert** der Dimension *s* bezeichnet:

$$EW_s := \frac{1}{m} \sum_{j=1}^{m} \eta^2_{js} \qquad (38)$$

Auch hier ist die Analogie zur linearen Hauptkomponentenanalyse unverkennbar.
Der HOMALS-Eigenwert variiert aufgrund seiner Definition wie die Diskriminanzmaße zwischen Null und Eins (vgl. obigen Hinweis zur Inflationierung von Diskriminanzmaßen).

Die SPSS-Prozedur HOMALS liefert für Hartigans Hardware bei Anforderung einer zweidimensionalen Lösung folgende Diskriminanz-Ergebnisse (im Ausgabefenster bzw. Grafik-Karussell):

```
Dimension   Eigenvalue
---------   ----------
    1          ,6212
    2          ,3681

Discrimination measures per variable per dimension
==================================================

Variable  Dimension

              1       2

THREADN     ,930    ,024
HEADN       ,951    ,634
INDHEADN    ,945    ,682
BOTTOMN     ,546    ,020
LENGTH      ,292    ,818
BRASSN      ,064    ,030
```



Einige Anmerkungen zur Interpretation:

- Die erste Dimension hat einen deutlich größeren Eigenwert, d.h. sie trennt erheblich besser zwischen den Kategorien. Vermutlich wird bei den meisten HOMALS-Analysen die erste Dimension die prägnantesten und stabilsten Ergebnisse bringen.
- Eine erste inhaltliche Interpretation der HOMALS-Dimensionen kann anhand der Diskriminanzmaße erfolgen.
  Als "**Markiervariablen**" zur ersten Dimension kann man THREADN (Gewinde: ja oder nein) und BOTTOMN (Fußform: spitz oder flach) ansehen. Beide Variablen "laden" relativ hoch auf der ersten Dimension und praktisch überhaupt nicht auf der zweiten.
  Die zweite Dimension wird nur durch LENGTH markiert, scheint also hauptsächlich zwischen unterschiedlich langen Objekten zu trennen. Bei Anwendung faktorenananalytischer Beurteilungs-Kriterien verdient diese Dimension als "**single**" wenig Beachtung, da sie nur *ein* Ausgangsmerkmal reflektiert und somit keinen Vereinfachungsgewinn bietet.
  HEADN (Kopfform) und INDHEADN (Kopfeinkerbung) weisen auf beiden Dimensionen nennenswerte Diskriminanzmaße, d.h. Unterschiede zwischen den Kategorien, auf.
- Die Variable BRASSN spielt (zumindest bei den beiden ersten Dimensionen) keine Rolle. Hier waren auch keine Diskriminanzleistungen der Hauptkomponenten zu erwarten, da alle Objektarten (Schrauben, Nägel etc.) sowohl in Messing als auch in anderen Materialien vertreten waren. Dementsprechend "korreliert" BRASSN schlecht mit den übrigen Variablen und hat daher wenig Einfluß auf die ersten Hauptkomponenten.
  Von einem analogen Argument ist eigentlich auch die Variable LENGTHN betroffen. Daß sie sich trotzdem in Bezug auf Dimension 2 als relativ trennscharf hervortut, liegt vermutlich an einem einzigen Ausreißer: Objekt Nr. 10 hat eine extreme Länge und einen extremen Wert auf der zweiten Dimension.

4.7 Die HOMALS-Plots

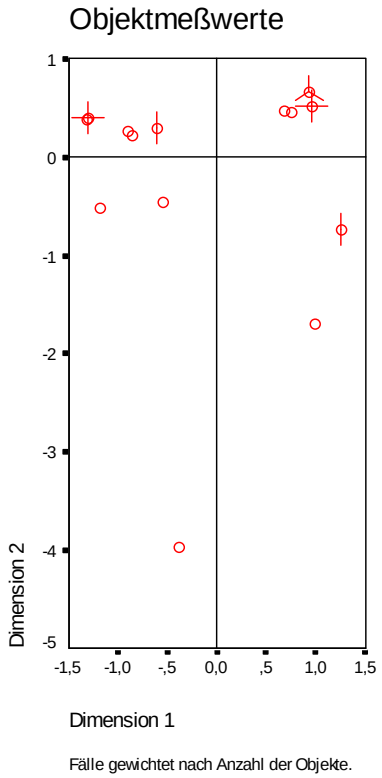
Im vollständigen HOMALS-Plot werden Kategorien und Objekte gemeinsam in der Ebene dargestellt, die von den beiden ersten Dimensionen aufgespannt wird. Seine wesentlichen Eigenschaften:

- Kategorienpunkte sind die Schwerpunkte der zugehörigen Objektpunkte.
- Eine Variable diskriminiert um so besser, je weiter ihre Kategorienpunkte auseinander liegen und je enger die Objektpunkte bei ihrem zugehörigen Kategorienpunkt liegen.
- Die Distanz zwischen zwei Objektpunkten erlaubt Schlüsse auf ihre Ähnlichkeit: Zwei Fälle mit identischem Profil haben dieselbe Plot-Position. Allerdings kann man aus einer ähnlichen Plotposition noch nicht auf sicher auf große Profilähnlichkeit schließen, weil der Plot nur die beiden ersten Dimensionen zeigt.
- Objekte mit durchschnittlichem Antwortverhalten liegen nahe beim Mittelpunkt (Koordinaten (0,0)), während Objekte mit untypischem Muster am Rand erscheinen. Eventuell wird diese Randständigkeit allerdings erst auf höheren Dimensionen sichtbar.
- Da eine Kategorien-Quantifizierung gerade der mittlere Score der zugehörigen Objekte ist, liegen zwei Kategorienpunkte um so näher, je mehr Objekte sie gemeinsam haben.

Eine gemeinsame Darstellung aller Objekt- *und* Kategorienpunkte kann sehr unübersichtlich werden. Deshalb bietet die SPSS-Prozedur HOMALS mehrere Plot-Typen mit unterschiedlichen Teilinformationen an.

4.7.1 Plot der Objektscores

Beim Plotten der Objektscores kann man die Objekte auf unterschiedliche Weise im Plot kennzeichnen lassen. Im einfachsten Fall erfolgt die Darstellung wie beim bivariaten Scattergramm, d.h. im Prinzip wird jedes Objekt durch einen Punkt markiert. Bei Mehrfachbelegung einer Plot-Position wendet SPSS die Sonnenblumendarstellung an. Diese Darstellung ist vor allem bei großen Stichproben sinnvoll. In unserem Beispiel erhalten wir:



Anmerkung zur Interpretation:

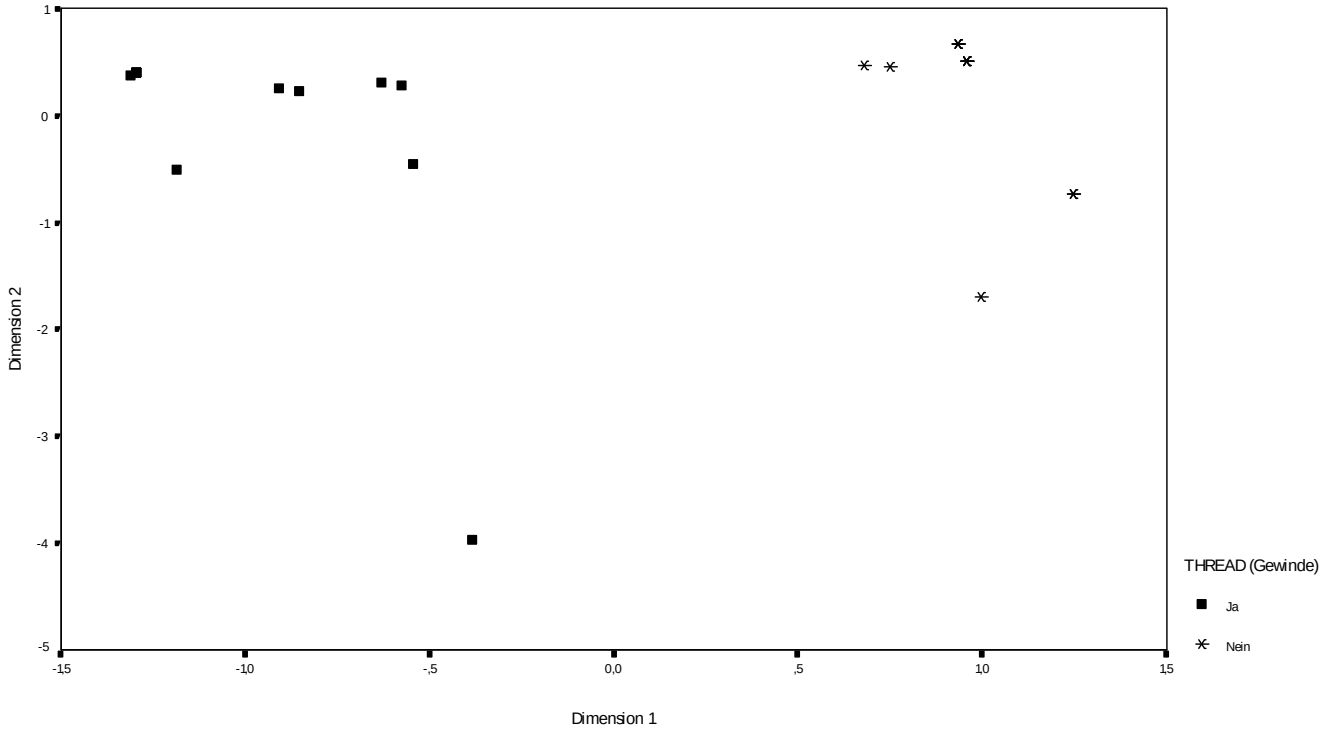
> Ein Objekt (Nr. 10, vgl. obige Tabelle mit den Objektscores) hat einen extremen Score bei Dimension 2 (- 3,96). Hier handelt es sich um einen Ausreißer mit starkem Einfluß auf die Ergebnisse, vor allem in Bezug auf Dimension 2. Nach Elimination dieses Objektes zeigt Dimension 2 deutlich veränderte Diskriminanzmaße.

SPSS-HOMALS bietet weiterhin Scoresplots an, in denen die Objekte durch Angabe ihrer Kategorie bei einer bestimmten Variablen gekennzeichnet sind. Wir erhalten in unserem Beispiel also 6 Plots mit identischen Objektpositionen, aber unterschiedlichen Objektmarkierungen, die zur Beurteilung der Diskriminanzleistungen der Variablen sowie bei der inhaltlichen Interpretation der Dimensionen nützlich sein können. Leider sind die hochaufgelösten Varianten dieser Abbildungen mißglückt, weil sich die Beschriftungen i.a. erheblich überlagern und dadurch unleserlich werden. Die Semigrafiken, die man in einem Syntaxfenster mit dem Kommando

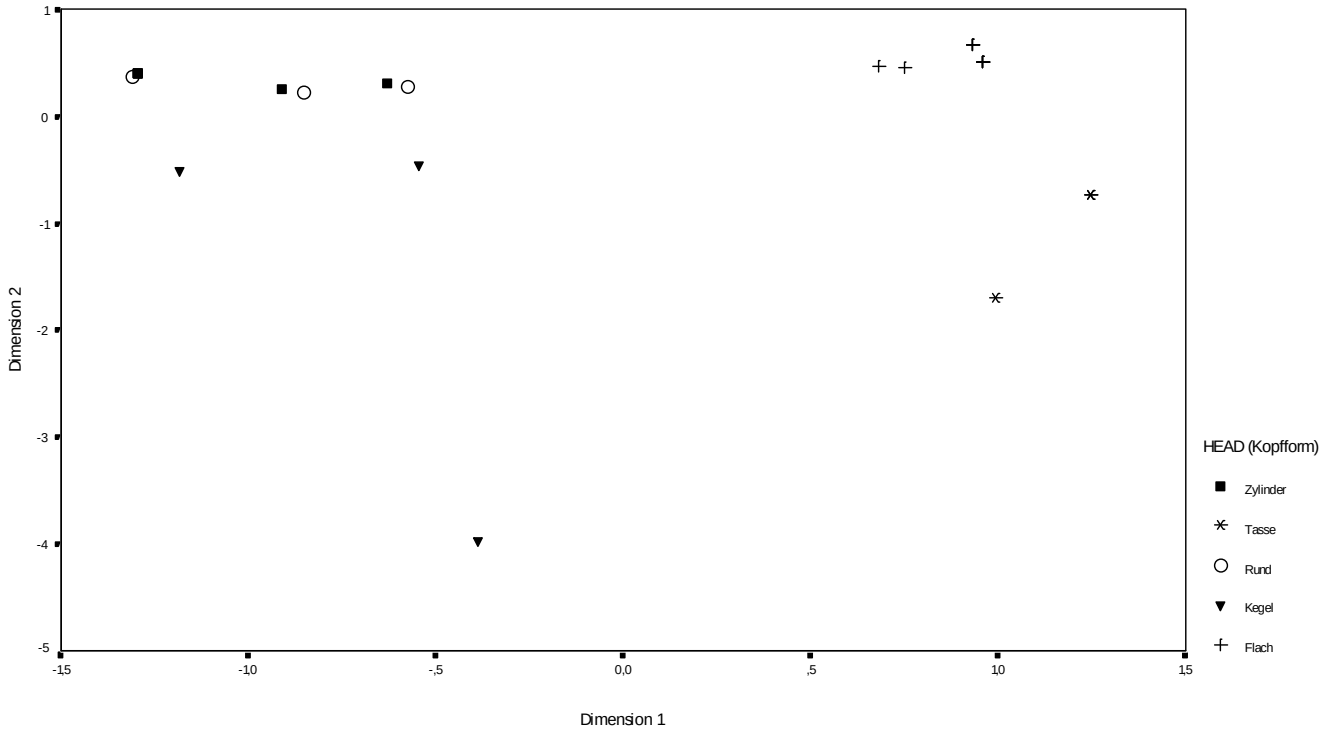
```
set lowres = on.
```

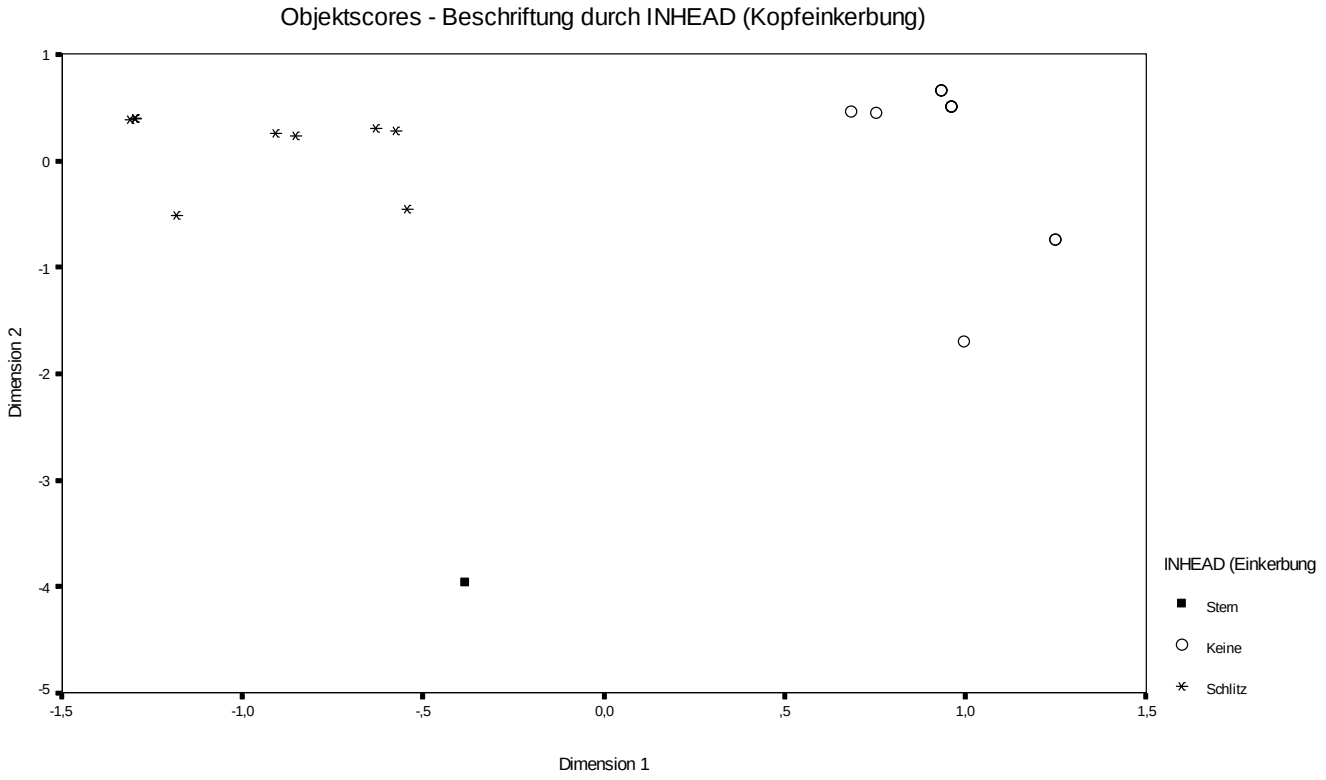
einschalten kann, sind ebenfalls nicht sehr anschaulich. Bei Mehrfachbesetzungen wird ein Fußnotenzeichen eingetragen und unterhalb des Plots erklärt. Eine sinnvolle Lösung besteht darin, die Scores-Variablen zu speichern und dann Scatterplots mit den gewünschten Gruppierungsvariablen zu erstellen. Das Speichern der Objektscores kann in der Optionen-Subdialogbox der HOMALS-Prozedur per Kontrollkästchen angefordert werden. Auf diese Weise entstanden die folgenden sechs Abbildungen:

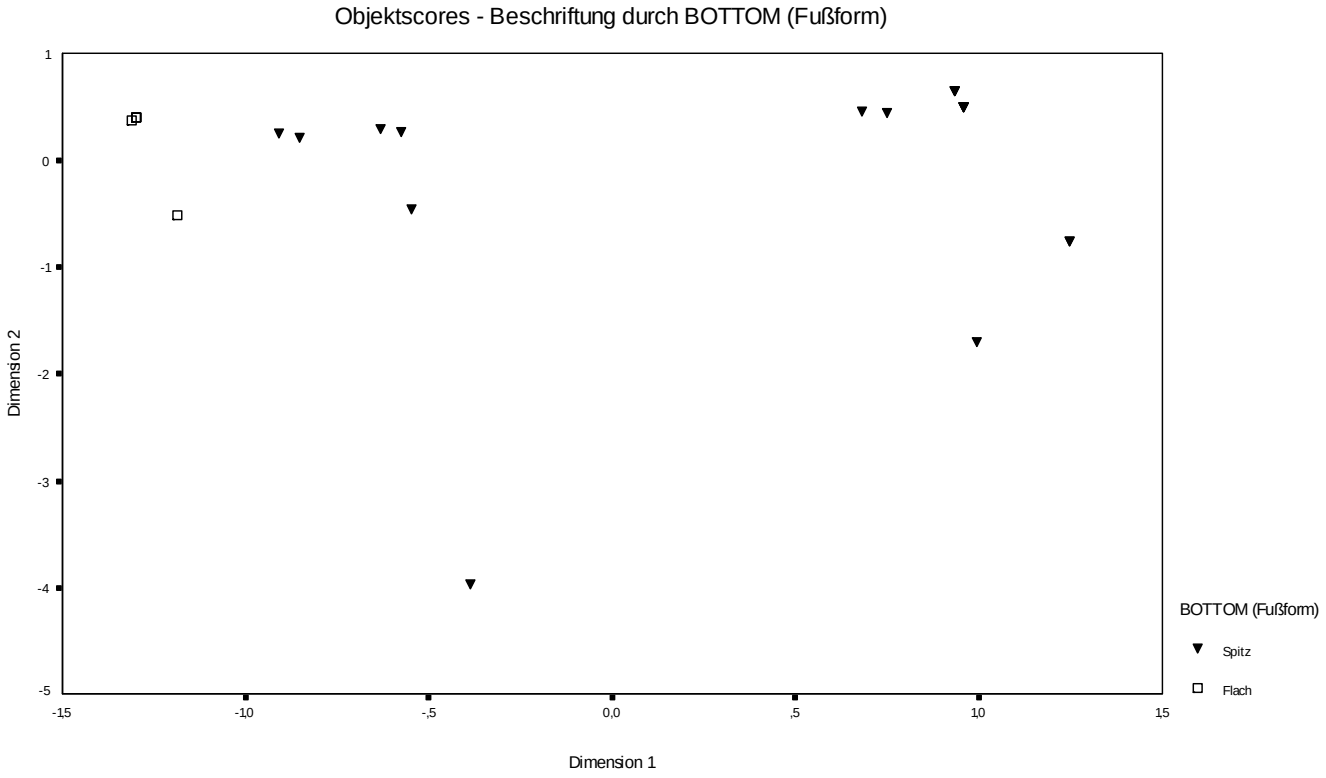
Objektscores - Beschriftung durch THREAD (Gewinde)

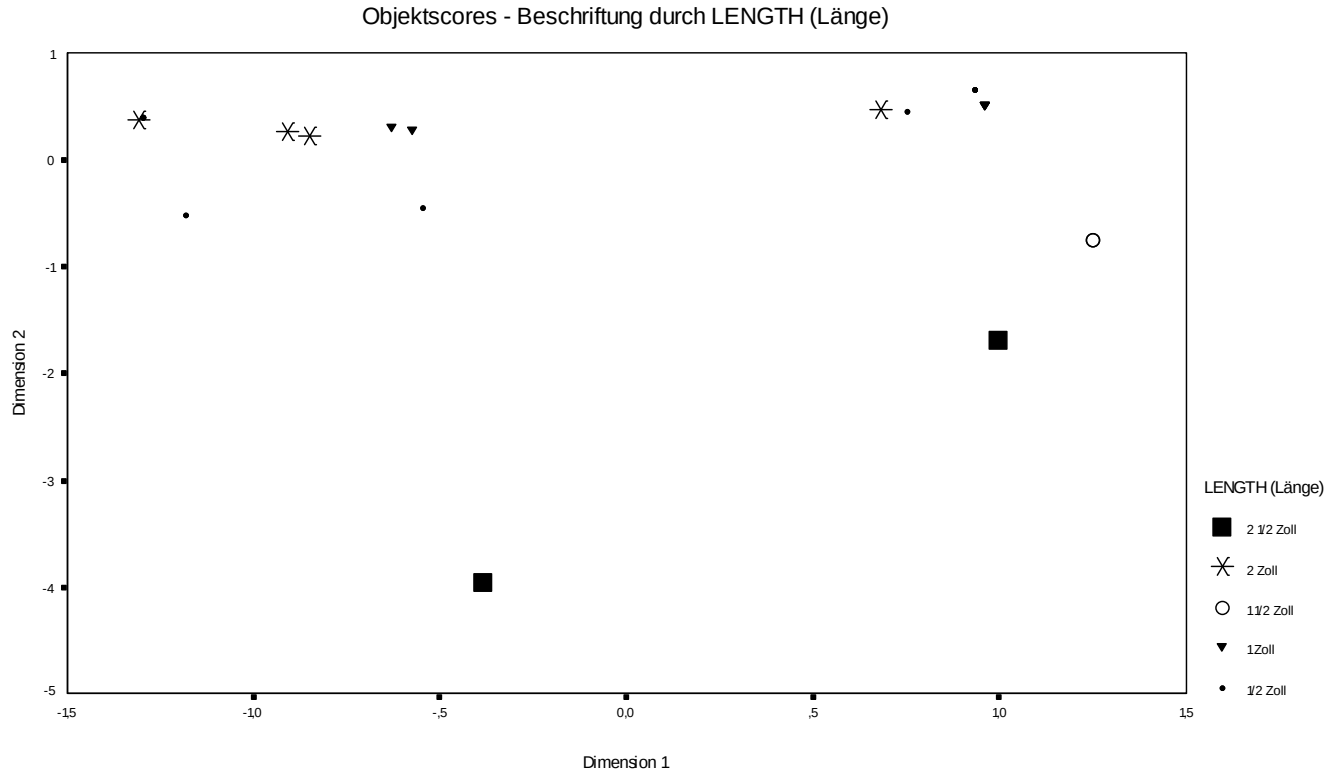


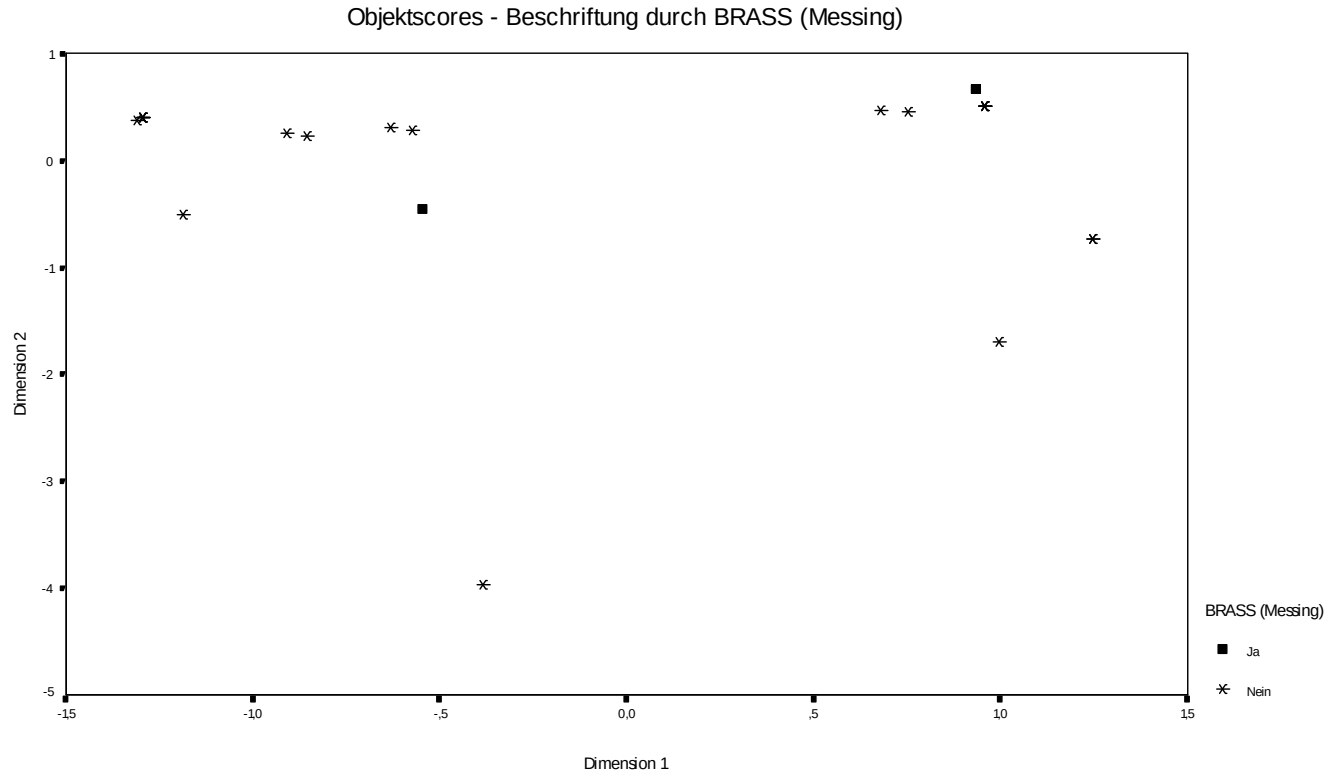
Objektscores - Beschriftung durch HEAD (Kopfform)




Objektscores - Beschriftung durch INHEAD (Kopfeinkerbung)
Dimension 2
Dimension 1
1
0
-1
-2
-3
-4
-5
-1,5
-1,0
-,5
0,0
,5
1,0
1,5
INHEAD (Einkerbung
Stern
Keine
Schlitz


Objektscores - Beschriftung durch BOTTOM (Fußform)
Dimension 2
Dimension 1
1
0
-1
-2
-3
-4
-5
-15
-10
-,5
0,0
,5
1,0
15
BOTTOM (Fußform)
Spitz
Flach





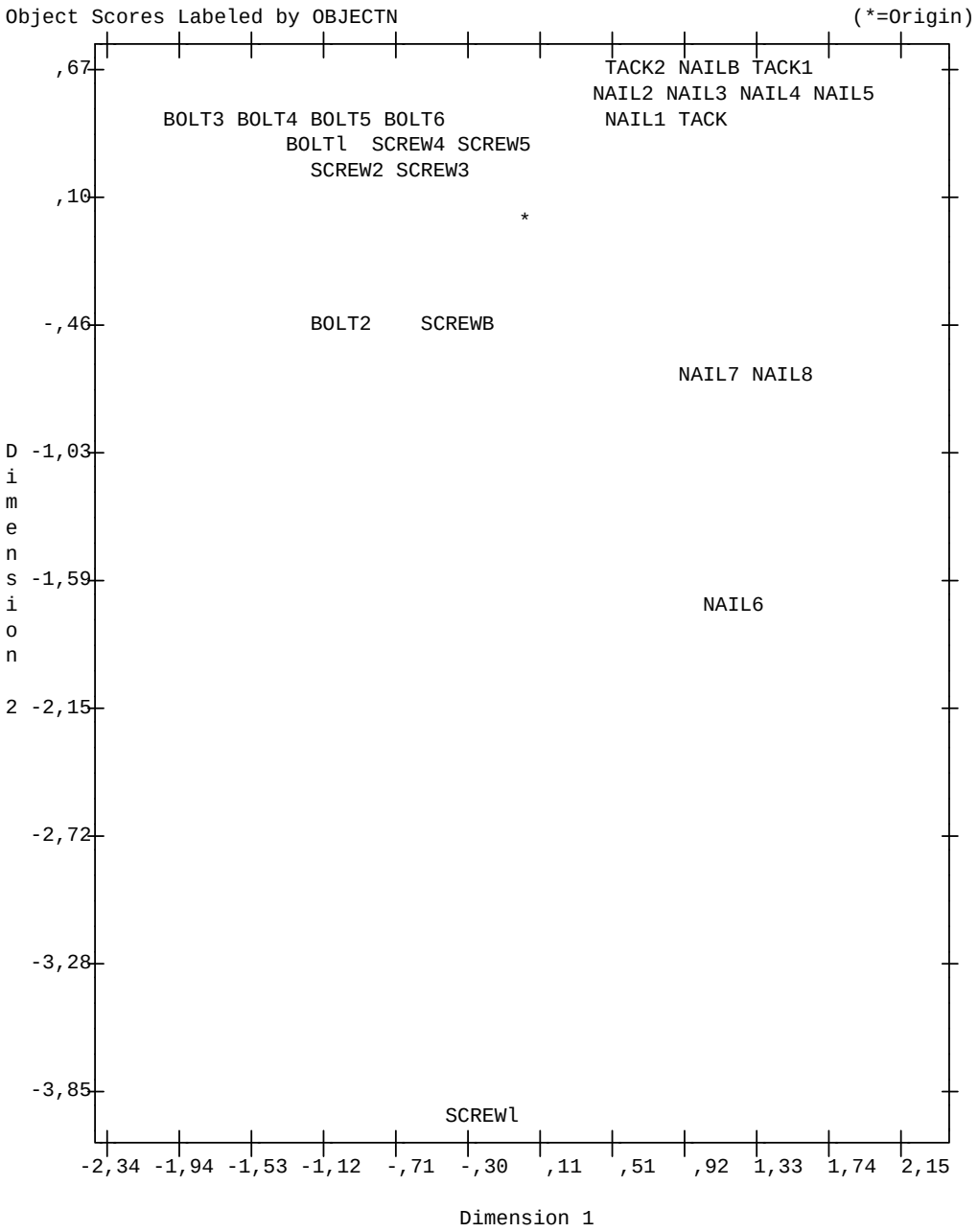
Anmerkungen zur Interpretation:

- Die wesentliche Diskrimination findet entlang der ersten Dimension statt. Vor allem die Kategorien der Gewinde-Variablen THREADN sind auf dieser Dimension exzellent getrennt. Diese Variable liefert damit wesentliche Information über die latente Dimension.
- Die Kategorien von HEADN (Kopfform) werden im zweidimensionalen Raum, also bei Berücksichtigung beider Dimensionen, recht gut separiert. Eine Ausnahme machen nur die Kategorien Rund und

Zylinder, die nicht unterschieden werden können. Damit ist klar, warum HEADN gute Diskriminanzwerte bzgl. beider Dimensionen aufweist.
Recht ähnlich sind die Verhältnisse bei der Variablen INHEADN (Kopfeinkerbung): Gute Separierbarkeit im Raum der beiden ersten Dimensionen spiegelt die beiden hohen Diskriminanzwerte wieder.

- Die Kategorien der Variablen BOTTOM (Fußform) unterscheiden sich deutlich auf der ersten Dimension, wenn auch keine perfekte Trennung möglich ist.
- Die Kategorien der Variablen BRASSN (Messing) können kaum getrennt werden, was aufgrund der oben mitgeteilten Diskriminanzmaße zu erwarten war.
- Bei der Variablen LENGTH (Länge) können mit Hilfe der zweiten Dimension die beiden längsten Objekte vom Rest getrennt werden. Dies bewirkt offenbar das hohe Diskriminanzmaß bezüglich der zweiten Dimension.
- Bei Konzentration auf die erste Dimension sind unter Vernachlässigung des Ausreißers zwei in mehreren Variablen gut getrennte Objekt-Cluster festzustellen. Vielleicht handelt es sich hier um zwei Grundtypen des Eisenhandels. Im ersten Cluster haben alle Objekte ein Gewinde und einen geschlitzten Kopf von runder, zylindrischer oder kegelförmiger Gestalt. Bei der Fußform sind beide Varianten (flach und spitz) vertreten.
  Im zweiten Cluster finden wir perfekte Homogenität bei drei Variablen: Kein Gewinde (THREADN), kein Schlitz im Kopf (INDHEADN), spitzer Fuß (BOTTOMN). Bei der Kopfform (HEAD) gibt es zwar zwei Modelle (flach und tassenförmig), aber keine Form, die auch im ersten Cluster vertreten wäre.

In nächsten Plot wird zu jedem Objekt der (unserem Plastillin-Forscher natürlich nicht bekannte) Name angegeben, um so die Identität der beiden entdeckten Cluster noch weiter zu erhellen. Weil die hochaufgelöste Grafik völlig unleserlich war, wurde die Semigrafik angefordert und leicht modifiziert: Bei Mehrfachbesetzungen wurden alle betroffenen Objekte, bei geringfügigem Genauigkeitsverlust, nebeneinander eingetragen, anstatt positionsgenau ein Fußnotenzeichen zu setzen und unterhalb des Plots zu erklären.



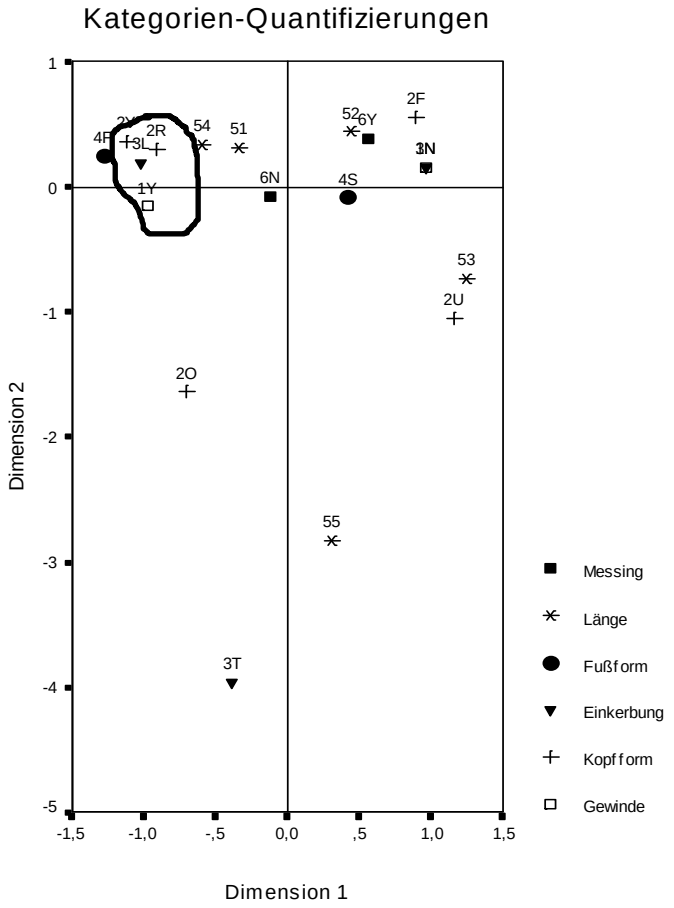
Indem die Plastillin-Forscher mit Hilfe der HOMALS-Methode möglichst viel Information über die Unterschiede diverser Kleineisenteile bzgl. mehrerer kategorialer Variablen in einem zweidimensionalen Raum dargestellt haben, ist ihnen die Entdeckung einer fundamentalen Hardware-Typologie geglückt:

Schrauben und Bolzen vs. Nägel und Stifte

Diese "latente kategoriale Variable" hat offenbar für die Homogenität der manifesten kategorialen Variablen gesorgt, welche durch die erste Dimension mit dem beträchtlichen Eigenwert 0,62 reflektiert wird.
Ob die vom HOMALS-Verfahren gelieferten metrischen Dimensionen als „echte“ latente Eigenschaften der auf manifester Ebene nur kategorial beschriebenen Objekte betrachtet werden sollen oder lediglich als diagnostische Hilfsmittel zur approximativen Abbildung komplexer Muster in einen niedrigdimensionalen Raum, der unseren Sehgewohnheiten und Verständnismöglichkeiten entgegenkommt, sei dahingestellt.

4.7.2 Plots der Kategorien-Quantifizierungen

Die Lokalisation der Variablen-Kategorien in der Dimensionsebene und Ähnlichkeitsbeziehungen zwischen Kategorien zeigt der folgende Plot der Kategorien-Quantifizierungen. Durch Änderung der Werte-Etiketten wurde dafür gesorgt, daß jede Kategorie durch ihren Kennbuchstaben (siehe Tabelle in Abschnitt 4.1) und die Nummer der Variablen markiert wird, damit möglichst wenige Überlappungen auftreten.



Anmerkung zur Interpretation:

- In der linken oberen Ecke stehen die Kategorien 2Y (zylindrischer Kopf), 2R (runder Kopf), 4F (flacher Fuß), 3L (geschlitzter Kopf) und 1Y (Gewinde) recht nah beieinander. Diese Ansammlung repräsentiert offenbar die klassische Metallschraube.
- Da eine Kategorien-Quantifizierung gerade der mittlere Score der zugehörigen Objekte ist, liegen zwei Kategorienpunkte um so näher, je mehr Objekte sie gemeinsam haben.

4.8 HOMALS als Analyse der uni- und bivariaten Randverteilungen

In Abschnitt 2 wurde darauf hingewiesen, daß sich die Gifi-Methoden bei der multivariaten Analyse auf die ein- und zweidimensionalen Randverteilungen beschränken. Dies zeigt sich besonders prägnant bei der folgenden Untersuchung der HOMALS-Lösung mit Hilfe der Matrixalgebra. Man kann nämlich für den Lösungsvektor **y** zur ersten Dimension zeigen:

$$\mathbf{D}^{-1}\mathbf{G}'\mathbf{G}\mathbf{y} = \frac{\mathbf{y}'\mathbf{D}\mathbf{y}}{n}\mathbf{y} \tag{39}$$

Die Matrix **G'G** enthält gerade die ein- und zweidimensionalen Randverteilungen zu **H**, was im Fall von Hartigans Hardware-Daten ergibt:

$$\mathbf{G'G} = \begin{bmatrix} 1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0 \\ 0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&1&0&0&0&1 \\ & & & & & & & & & & & . & . & . & . & & & & & & & & & \\ & & & & & & & & & & & . & . & . & . & & & & & & & & & \\ 1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \end{bmatrix} \begin{bmatrix} 1&0 \\ 1&0 \\ 1&0 \\ 1&0 \\ 1&0 \\ 1&0 \\ 1&0 \\ 1&0 \\ 1&0 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 0&1 \\ 1&0 \\ 1&0 \\ 1&0 \\ 0&1 \end{bmatrix}$$

$$= \begin{bmatrix} 12&0&9&0&0&3&0&0&12&0&0&12&4&4&2&1&1&9&3 \\ 0&12&0&3&3&0&6&11&0&1&6&6&6&2&0&3&1&11&1 \\ 9&0&9&0&0&0&0&0&9&0&0&9&4&4&0&1&0&6&3 \\ 0&3&0&3&0&0&0&2&0&1&1&2&2&0&0&0&1&2&1 \\ 0&3&0&0&3&0&0&3&0&0&1&2&0&1&0&2&0&3&0 \\ 3&0&0&0&0&3&0&0&3&0&0&3&0&0&2&0&1&3&0 \\ 0&6&0&0&0&0&6&6&0&0&4&2&4&1&0&1&0&6&0 \\ 0&11&0&2&3&0&6&11&0&0&6&5&6&2&0&3&0&10&1 \\ & & & & & & & . & . & . & . & & & & & & & & \\ & & & & & & & . & . & . & . & & & & & & & & \\ & & & & & & & . & . & . & . & & & & & & & & \\ & & & & & & & . & . & . & . & & & & & & & & \\ & & & & & & & . & . & . & . & & & & & & & & \\ & & & & & & & . & . & . & . & & & & & & & & \\ 9&11&6&2&3&3&6&10&9&1&6&14&6&6&2&4&2&20&0 \\ 3&1&3&1&0&0&0&1&3&0&0&4&4&0&0&0&0&0&4 \end{bmatrix}$$

D wurde schon in Abschnitt 4.3 erklärt als die Diagonalmatrix mit den Hauptdiagonalelementen von **G'G**.

Der Lösungsvektor **y** ist also ein Eigenvektor der Matrix $\mathbf{D}^{-1}\mathbf{G'G}$ 45, wobei der zugehörige Eigenwert $\frac{\mathbf{y'Dy}}{n}$ 46bis auf den Vorfaktor 1/*m* mit dem ersten HOMALS-Eigenwert identisch ist (vgl. Abschnitt 4.6.3).

Damit ist klar, daß die HOMALS-Lösung auf Informationen aus den ein- und zweidimensionalen Randverteilungen basiert.

4.9 Ein Simulationsexperiment mit HOMALS

Damit wir etwas Routine im Umgang mit HOMALS gewinnen, soll noch ein weiterer Datensatz analysiert werden[2]. Im Unterschied zum ersten Beispiel werden dabei künstliche Daten mit bekannten Eigenschaften verwendet, um die Reaktion von HOMALS zu studieren. Es handelt sich um zwei Familien bestehend aus 2 bzw. 3 untereinander sehr homogenen, dichotomen Variablen, wobei die beiden Familien relativ unabhängig voneinander sind. Die Verhältnisse lassen sich gut durch die Matrix der Phi-Korrelationen zwischen den Variablen beschreiben:

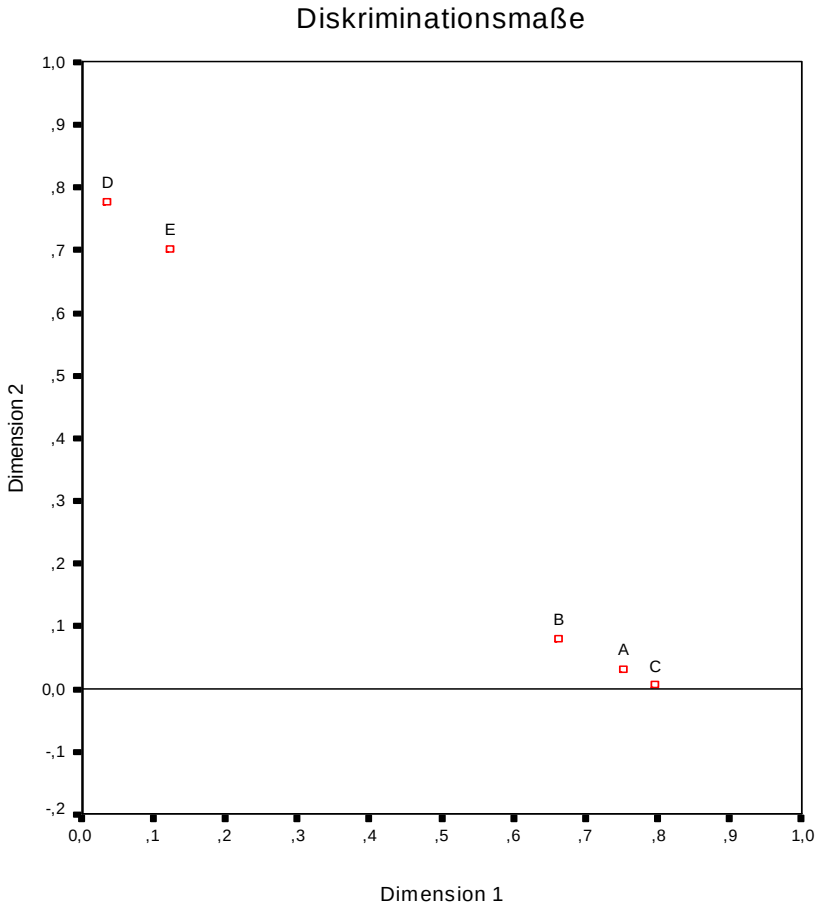
```
- -  Correlation Coefficients  - -

              A          B          C          D          E

A           1.0000      .6319**    .7033**   -.0385     -.1160
B            .6319**   1.0000      .6319**    .0385     -.0331
C            .7033**    .6319**   1.0000     -.0385     -.2652
D           -.0385      .0385     -.0385     1.0000      .6298**
E           -.1160     -.0331     -.2652      .6298**   1.0000
* - Signif. LE .05     ** - Signif. LE .01     (2-tailed)
```

Die fett dargestellten Binnenkorrelationen sind hoch, die Korrelationen zwischen den Familien sind geringfügig. Wir wollen sehen, ob HOMALS diese zweidimensionale Struktur aufdeckt.
Eine Inspektion der Diskriminanzmaße zeigt, daß dies sehr gut gelingt:

```
Dimension   Eigenvalue
---------   ----------
    1          ,4744
    2          ,3200

Discrimination measures per variable per dimension
==================================================

Variable  Dimension

               1       2

A            ,753    ,032
B            ,663    ,081
C            ,797    ,007
D            ,035    ,777
E            ,124    ,703
```

Der zugehörige Plot:

[2] Sie finden dieses Programm in der Datei **KATPCA.SPS** an der im Vorwort vereinbarten Stelle. Es ist außerdem im Anhang dieses Manuskripts abgedruckt.



Das Ergebnis entspricht genau unseren Erwartungen: Die erste Dimension wird von den Variablen A, B und C markiert, die zweite Dimension von den Variablen D und E.

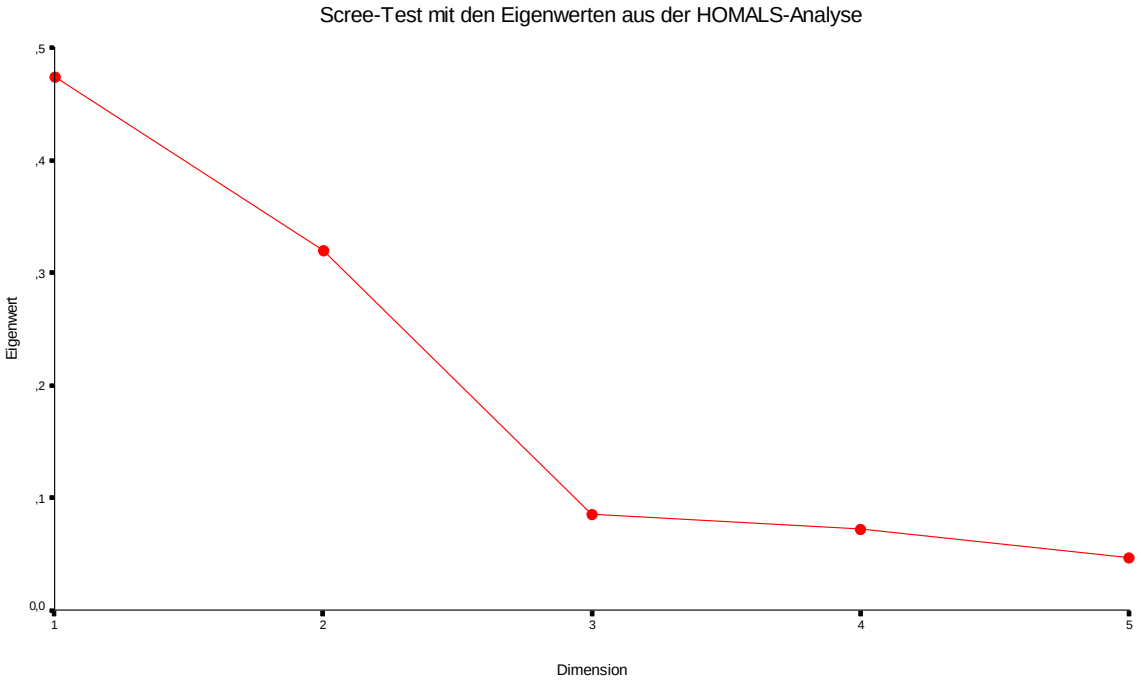
Exakt dieselben Ergebnisse liefert im hier vorliegenden Spezialfall von dichotomen kategorialen Variablen übrigens die lineare Hauptkomponentenanalyse. Wir erhalten für denselben Datensatz als **unrotierte** Faktormatrix:

```
FACTOR MATRIX:

                FACTOR  1      FACTOR  2

A                 .86763         .17859
B                 .81403         .28524
C                 .89293         .08263
D                -.18811         .88152
E                -.35198         .83841
```

Wie in Abschnitt 4.6.3 erwähnt, können die Diskriminanzmaße η^2_{js} 47 analog zu quadrierten Ladungen interpretiert werden. Hier erhalten wir durch Quadrieren der Ladung einer Variablen auf einem unrotierten Faktor der linearen Hauptkomponentenanalyse sogar exakt ihr Diskriminanzmaß für die entsprechende HOMALS-Dimension.
Der Eigenwerte einer unrotierten Hauptkomponente ist gerade identisch sind mit der Summe ihrer quadrierten Ladungen. Ganz analog gilt für einen HOMALS-Eigenwerte und die zugehörigen Diskriminanzmaße:

$$EW_s := \frac{1}{m} \sum_{j=1}^{m} \eta^2_{js}$$

Folglich stimmen in unserer Situation die Eigenwerte aus der Hauptkomponenten- und der HOMALS-Analyse bis auf den Vorfaktor $1/m$ überein. Folglich erhält man auch beim Scree-Test mit den HOMALS-Eigenwerten die richtige Empfehlung einer zweifaktoriellen Lösung



Eine so weitgehende Entsprechung von Hauptkomponenten- und HOMALS-Analyse findet sich aber nur im Spezialfall zweiwertiger Ausgangsvariablen. Hier ist offenbar die im Optimierungsverfahren der Hauptkomponentenanalyse enthaltene lineare Transformation der Ausgangsvariablen allgemein genug (vgl. Abschnitt 3.2).

5 Literatur

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

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

Heuser, H. (1986). *Lehrbuch der Analysis, Teil 1*.Stuttgart: Teubner.

Jöreskog, K. G. & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago, IL: SPSS.

Sonnemann, E. (1992). *Exploratorische Datenanalyse*. Mitschrift der Vorlesung im Sommersemester 1992.

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

Stoer, J. & Bulirsch, R. (1978). *Einführung in die numerische Mathematik II*. Berlin: Springer.

6 Anhang

Das SPSS-Programm zur Analyse von Hartigans Hardware-Daten:

```
* SPSS Inc. Categories (1990) Figure 7.1a The Hartigan hardware data

DATA LIST / THREAD 1(A) HEAD 3(A) INDHEAD 5(A)
 BOTTOM 7(A) LENGTH 9 BRASS 11(A) OBJECT 13-18(A).
VARIABLE LABELS HEAD 'Head form'
 INDHEAD 'Indentation of head'
 BOTTOM 'Bottom shape'
 LENGTH 'Length in half inches'.
VALUE LABELS
 THREAD 'Y' 'Yes Thread' 'N' 'No_Thread'
 /HEAD 'F' 'FLAT' 'U' 'CUP' 'O' 'CONE' 'R' 'ROUND' 'Y' 'CYLINDER'
 /INDHEAD 'N' 'NONE' 'T' 'STAR' 'L' 'SLIT'
 /BOTTOM 'S' 'sharp' 'F' 'flat'
 /BRASS 'Y' 'YesBr' 'N' 'NotBr'
 /LENGTH 1 '1/2 in' 2 '1 in' 3 '1 1/2 in'
         4 '2_in' 5 '2_1/2 in'.
BEGIN DATA
N F N S 1 N TACK
N F N S 4 N NAILl
N F N S 2 N NAIL2
N F N S 2 N NAIL3
N F N S 2 N NAIL4
N F N S 2 N NAIL5
N U N S 5 N NAIL6
N U N S 3 N NAIL7
N U N S 3 N NAIL8
Y O T S 5 N SCREWl
Y R L S 4 N SCREW2
Y Y L S 4 N SCREW3
Y R L S 2 N SCREW4
Y Y L S 2 N SCREW5
Y R L F 4 N BOLTl
Y O L F 1 N BOLT2
Y Y L F 1 N BOLT3
Y Y L F 1 N BOLT4
Y Y L F 1 N BOLT5
Y Y L F 1 N BOLT6
N F N S 1 Y TACKl
N F N S 1 Y TACK2
N F N S 1 Y NAILB
Y O L S 1 Y SCREWB
END DATA.
AUTORECODE VARIABLES=THREAD,HEAD,INDHEAD,BOTTOM,BRASS,OBJECT
 /INTO THREADN,HEADN,INDHEADN,BOTTOMN,BRASSN,OBJECTN.

HOMALS VARIABLES=THREADN(2) HEADN(5) INDHEADN(3)
 BOTTOMN(2) LENGTH(5) BRASSN(2) OBJECTN(24)
 /ANALYSIS=THREADN HEADN INDHEADN BOTTOMN LENGTH BRASSN
 /DIMENSION=2
 /PRINT=DEFAULT OBJECT
 /PLOT=DISCRIM OBJECT(THREADN HEADN INDHEADN BOTTOMN LENGTH
  BRASSN OBJECTN) QUANT
 /SAVE=SC.
```

Das SPSS-Programm zum Simulationsexperiment in Abschnitt 4.9:

```
* Kuenstlicher Datensatz zur Demonstration der kategorialen
  Hauptkomponentenanalyse mit HOMALS.
data list free /a b c d e.
begin data.
1 1 1 1 1
1 1 1 2 2
1 1 1 2 2
1 1 1 1 1
1 1 1 2 2
1 1 1 1 1
1 1 1 1 1
1 1 1 2 2
1 1 1 1 2
1 1 1 2 1
1 1 2 1 1
1 2 1 2 2
1 2 2 1 1
1 2 1 2 2
2 1 1 1 1
2 1 2 2 2
2 2 1 1 2
2 2 2 2 1
2 2 2 2 2
2 2 2 1 1
2 2 2 2 2
2 2 2 1 1
2 2 2 2 2
2 2 2 1 1
2 2 2 2 1
2 2 2 1 1
2 2 2 1 1
end data.

value labels a 1 'a1' 2 'a2'
 /b 1 'b1' 2 'b2' /c 1 'c1' 2 'c2'
 /d 1 'd1' 2 'd2' /e 1 'e1' 2 'e2'.

homals var = a b c d e(2)
 /analysis = a b c d e
 /print = all
 /plot = discrim quant object(a b c d e).
```

7 Stichwortverzeichnis