Empirische Untersuchung

Datenerhebung

Die Wunschdaten wurden innerhalb der Rostocker Längsschnittstudie (Reis 1997, eine Beschreibung der Studie bei Silbereisen und Vaskovic 1993) erhoben.

Stichprobe

Diese Untersuchung war ursprünglich zum Thema Geburtsrisiken angelegt worden. Für diesen Zweck wurde eine repräsentative Stichprobe des Geburtsjahrganges 1970 im Einzugsbereich der Universitätsklinik Rostock zusammengestellt. Aus dieser Grunderhebung sind im Rahmen eines quasiexperimentellen Designs jeweils 150 Kinder mit leichten Geburtsrisiken und 150 Kinder ohne Geburtsprobleme ausgewählt worden.

In der Anfangsphase stand die prospektive Begleitung der Kinder mit Geburtsrisiken und der Vergleich ihrer Entwicklung mit der Kontrollgruppe im Mittelpunkt. Eine der wichtigsten Aussagen, die in diesem Rahmen getroffen werden kann, ist, daß sich die aus den Geburtsrisiken ergebenen Leistungs- und Persönlichkeitsunterschiede zwischen den Gruppen ab etwa dem Schuleintritt nivellieren. Man erklärt sich dies einerseits mit einem erfolgreichen physiologischen Adaptationsprozeß der Risikogruppe und andererseits der Zunahme sozialer Risikofaktoren bei der Kontrollgruppe. Allerdings ist zu vermuten, daß die Kombination von sozialen und Geburtsrisiken der Kinder bei der aufgetretenen experimentellen Mortalität eine erhebliche Rolle spielte. Für die Fragestellung der Arbeit ist dies jedoch weniger von Belang als die Tatsache, daß die Teilstichproben ab dem Schuleintritt unter psychometrischem Gesichtspunkt praktisch als äquivalent anzusehen sind.

Die gesamte Studie umfaßt bisher sieben Meßzeitpunkte. Zu den letzten drei Meßzeitpunkten, bei denen ein umfangreiches, etwa zwei Stunden dauerndes Interview durchgeführt sowie eine Vielzahl von Fragebogeninstrumenten angewendet worden ist, fragte man am Ende der Sitzung nach den drei Wünschen. (siehe Abbildung 6). Es liegen also vier Datensätze zu den drei Wünschen von Personen vor. Drei der Datensätze stellen eine Meßwiederholung derselben Personen über den Zeitraum der Adoleszenz bis ins frühe Erwachsenenalter dar. Der vierte Datensatz beinhaltet die drei Wünsche von Personen, die in enger Beziehung zu den erstgenannten Probanden stehen, es sind deren Mütter. Somit haben die vorliegenden Daten ein

großes Potential für die Untersuchung entwicklungspsychologischer Fragestellungen. Von diesem soll jedoch in dieser Untersuchung abgesehen werden. Es wird hier vielmehr genutzt, über vier natürliche, nichtsynthetische Datensätze jeweils altershomogener Aggregate verfügen zu können. Unter inhaltlichem Gesichtspunkt sollte zunächst die Struktur von Wünschen zu verschiedenem Altersstufen aufgedeckt werden, bevor man sich entwicklungspsychologisch relevanten Transformationen zwischen diesen Stufen zuwendet. Methodologisch gesehen sind vier Datensätze in der Methodenerprobung günstiger als nur ein Datensatz. Unter ungünstigen Umständen könnte die Datenstruktur in einem einzelnen Datensatz durch Störvariablen verschleiert worden sein. Bei vier Datensätzen liegen dagegen vier Eisen im Feuer und die wiederholte Anwendung der zu erprobenden Methoden mag besseren Aufschluß über deren Eigenschaften liefern als die einmalige Anwendung.

Jahr	1970/71	1972/73	1976/77	1980/81	1984/85	1990/91	1994/95
Alter	Geburt	2 Jahre	6 Jahre	10 Jahre	14 Jahre	20 Jahre	24 Jahre
N	1000	294	279	268	247	199	212
%		100	95	91	84	68	72
Wünsche der Probanden					X	X	X
Wünsche der Mütter							X

Abbildung 6: Stichprobe der ROLS 1970 – 1995
und mit „x“ gekennzeichnete Erhebungszeitpunkte der Wünsche

Instruktion

„Wenn Dir eine gute Fee begegnen würde und sie zu Dir sagte, Du hättest drei Wünsche frei, was würdest Du Dir dann wünschen. Beginne mit dem Wichtigsten.“

Die Antworten wurden von den Interviewern stichwortartig auf einem Antwortbogen notiert. Wurde nur ein Wunsch genannt, fragte der Interviewer nach, um möglichst auch wirklich drei Wünsche für jede Person erheben zu können. Nach Angabe der Interviewer nannten die Probanden in den überwiegenden Fällen ihre drei Wünsche jedoch spontan, schnell und gern.

Insgesamt liegen 2135 Antworten auf die Frage nach den drei Wünschen vor (Tabelle 10).

Tabelle 10: Anzahl der Antworten auf der Frage nach den drei Wünschen zu den Befragungszeitpunkten und unter Berücksichtigung der Antwortreihenfolge.

Anzahl

		Rangplatz			
		1	2	3	Gesamt
ALTER	14	240	233	221	694
	20	191	190	187	568
	25	180	161	128	469
	Mütter	139	137	128	404
Gesamt		750	721	664	2135

Kategorisierung

Bei Meyer-Probst, Teichmann und Engel (1989) sowie Meyer-Probst, Teichmann und Kleinpeter (1989) wurden die Wünsche als Daten im Sinne der IPPNW-Veröffentlichungen kategorisiert (Solantanus, Rimpelä & Taipale, 1984; Chivian & Mack, 1985; Sommers, 1985; Chivian & Bergström, 1986; MacPherson, 1987; Roschtschin & Kabatschenko, 1988). Nur diese Kategorien wurden ursprünglich bei der Datenverarbeitung benutzt.

Für die vorliegende Studie wird auf handschriftlichen Notizen zum Interview in den Erhebungsbögen der ROLS zurückgegriffen. Dazu mußten die entsprechenden Textstücke der Handakte möglichst wortgetreu DV-technisch erfaßt faßt werden. Die Wünsche sind in diesem erneuten Erfassungsvorgang in der Form alphanumerischer Variablen in eine SPSS-Systemdatei angelegt worden. Aus Gründen der Übersichtlichkeit entsprechen unterschiedliche Ausdrucksweisen derselben Sachverhalte als gleiche Texteingabe dem selben Code. Für synonyme Ausdrücke wie beispielsweise „Arbeit“, „Anstellung“ oder „Job“ steht schon ab der Eingabe nur ein Kategoriencode, hier „Arbeit“.

Insgesamt ergeben sich nach diesem Arbeitsschritt bei $r = 2135$ validen Antworten über alle Meßzeitpunkte, alle Kohorten und alle drei Antwortnennungen $k = 764$ Kategorien. Große Teile der Kategorien gehen jedoch auf abweichende Schreibweisen derselben Sachverhalte (teils Fehler der Orthographie, teils uneinheitliche Abkürzungen) zurück. Deshalb werden in einem weiteren Verarbeitungsschritt die verschiedenen Schreibweisen offensichtlich gleicher Inhalte in jeweils eine einzige Kategorie zusammengefaßt. In einem letzten Schritt erfolgt die Fusionierung einzelnen Kategorien nach semantischen Kriterien. Dabei ist die Zuweisung jeder Wunschäußerung zu jeweils genau einer Kategorie vollzogen worden.

Die ganze Verfahrensweise ähnelt dem beschriebenen Vorgehen von Jersild, Markey und Jersild (1933) bei deren Klassifikation von Wunschnennungen, wie auch der Methode von Anderson (1981) bei seinem Schema zur Differenzierung von bildhaften Vorstellungen. Der Unterschied zu diesen beiden Mehtoden liegt jedoch in der Intention. Es wird nicht wie dort die Zusammenfassung der Wunschäußerungen zu Klassen beabsichtigt. Vielmehr soll nur vermieden werden, daß die Abweichungen in der Bezeichnung semantisch offensichtlich gleichwertiger Sachverhalte zu einer unangemessen hohen Anzahl von Kategorien führt. Es resultieren schließlich $m = 49$ antwortnahe Kategorien. Abbildung 11 gibt den Erfassungsablauf bis zur Generierung der Wunschdatenmatrizen wieder.

Grundlage der folgenden Analysen sind also vier Datenmatrizen, jeweils einen für jeden der Erhebungszeitpunkte von Wünschen im Rahmen der ROLS. Die Einträge der Matrizen, Rangplätzen, entsprechen der Reihenfolge, in der die Wünsche genannt wurden. Dabei repräsentieren die Datensätzen (Zeilen) die Personen bezüglich $m = 49$ Wunschkategorien (Spalten).

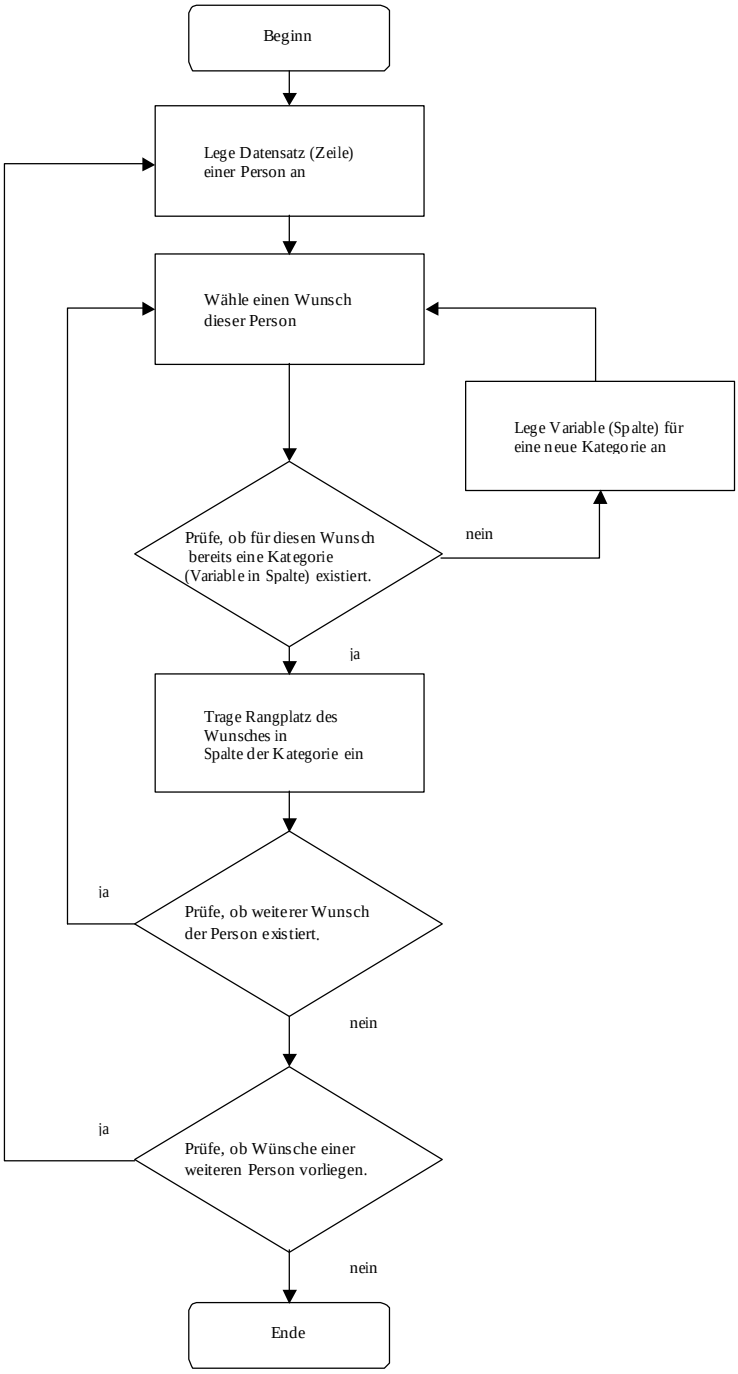


Abbildung 11: Schema zur Erfassung der drei Wünsche aller Personen in einer Datenmatrix

Festzuhalten ist, daß einige der so deklarierten Kategorien sehr marginal besetzt sind und daß die Kategorien, ersichtlich an Histogramm mit nach Frequenz absteigend geordneten Frequenzen, bei allen Stichproben j-Verteilungen bilden. Einige Kategorien treten überhaupt nur

in bestimmten Datensätzen, also nur zu einem bestimmten Lebensalter, auf (Anhang 10 bis Anhang 19). Zusätzlich führt der Umstand, daß jeder Datensatz maximal drei valide Datenpunkte beinhaltet, bei gegebener Anzahl von Kategorien zu einer extremen Datenlage, bei der ca. 94% der Einträge in den Matrizen keine klare Bedeutung haben. Klarheit besteht ausschließlich über die Bedeutung der drei validen Wunschnennungen einer Person. Bei den restlichen 46 möglichen, im Erfassungsschema vorgesehenen Wunschnennungen herrscht Unsicherheit darüber, ob es sich dabei jeweils um einen Wunsch mit zu geringer Wunschstärke handelt, oder ob die Person aktuell über keine Vorstellung der jeweiligen Kategorie verfügt. Eine solche schüttere Datenlage trägt die Bezeichnung „sparse data".

Nicht-Nennungen von Wünschen

Die verwendeten Modelle unterscheiden sich in ihren technischen Möglichkeiten der Verarbeitung von verschiedenen Datenformaten. Die beiden verwendeten Modelle mit euklidischen Abbildungsräumen, NMDS und HOMALS, sind in der Lage, Rangdaten direkt zu verarbeiten. HICLAS und FPA eignen sich im Rahmen der jeweils zur Verfügung stehenden Computerprogramme nicht für Rangdaten, sie erhalten deshalb als Eingabewerte Inzidenzdaten. Aus diesen Unterschieden ergeben sich Konsequenzen für den Umgang mit den nicht genannten Wünschen einer Person. Inzidenzdaten gestatten eine natürliche Behandlung. Die genannten Wünsche werden affirmativ im Sinn von logisch „wahr" mit 1 verkodet, die nicht genannten Wünsche erhalten den Wert 0 im Sinn von logisch „falsch". Dies stellt jedoch nur eine schnelle und einfache technische Lösung bei der Verkodung dar.

Das Bedeutungsproblem für die Vielzahl nicht genannter Wünsche pro Person ist allerdings bei einer Restriktion auf drei Äußerungen nicht gelöst. Für Rangdaten ergeben sich vielmehr zwei Alternativen der Behandlung nicht genannter Wünsche. Einerseits können sie als fehlende Werte aus der Analyse ausgeschlossen werden. Das Modell stützt sich dann ausschließlich auf die wirklich beobachteten Ranginformationen. Andererseits kann man ihnen einen Rangplatz $r > 3$ zuweisen und diese Ränge in das Modell mit aufnehmen.

Im Sinn der oben vorgenommenen Explikation setzt diese Umwandlung der nicht genannten Wünsche in Wünsche unterhalb der Schwellenwerte ρ_i zwei Dinge voraus. Erstens: jede Person hat entsprechende Vorstellungen und zweitens: alle Personen entwickeln für jede dieser Vorstellungen eine Attraktivität $a_i(v) > 0$. Dies sind außerordentlich starke Annahmen. Das diese vorerst nicht überprüft werden könne spricht sehr dafür, nicht genannte Wünsche besser aus dem Modell herauszuhalten und sich nur auf die vorliegenden Rangdaten zu stützen.

Wünsche sollen im Sinne einer Entfaltung modelliert werden. Dabei wird zunächst der traditionelle Weg der Einbindung in das Modell der Nichtmetrischen Multidimensionalen Skalierung eingeschlagen. Ähnlich wie die NMDS führt die Modellierung im Rahmen von Homogenitätsanalysen (HOMALS) zu einer Repräsentation in einem Euklidischen Raum geeigneter Dimensionalität. Die Entfaltungsintention durch eine Repräsentation in einem Vektorraum abzielende Methode zu verfolgen ist jedoch nicht ohne Alternative. Caroll (1976), Harary (1972) und auch Cozzens und Leibowitz (1987) verweisen auf die Möglichkeit, Datenstrukturen nicht nur in einem Euklidischen Raum, sondern als Graphen zu repräsentieren. Dieser Richtung wird mit der Anwendung der Modelle HICLAS und FPA beschritten.

Unfolding im Rahmen Multidimensionaler Skalierung (NMDS)

Die Matrixform von Präferenzwahldaten von n Personen und m Objekten ist rechteckig und abgesehen vom Spezialfall ($n = m$) nicht-quadratisch.

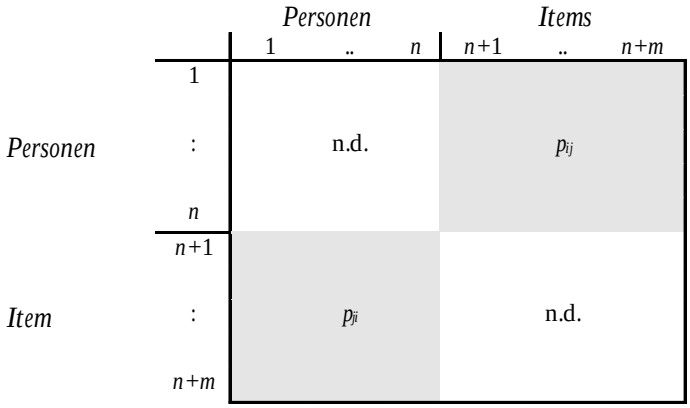


Abbildung 12: Präferenzwahldaten als Teilmatrizen ("off-diagonal corner matrices") einer Supermatrix
Nur die grau unterlegten Bereiche enthalten Daten, die anderen sind nicht definiert (n.d.)

Das übliche Format, das in der NMDS verarbeitet werden kann, entspricht jedoch einer Dreiecksmatrix außerhalb der Hauptdiagonale. Carroll löste das Problem unterschiedlicher Matrixformate, indem die (n x m)-Matrix als Grundlage für Teilmatrizen einer ($n + m$, $n + m$) Matrix dient. Die dabei besetzten "off-diagonal corner matrices" sind jedoch redundant in dem Sinn, daß gilt ($p_{ij} = p_{ji}$) mit p_{ij} unterhalb der Diagonale und p_{ji} oberhalb der Diagonale der Supermatrix (Abbildung 12). Offensichtlich liegen in der Supermatrix Daten ausschließlich für das kartesi-

sche Produkt elementefremder Mengen (between proximities) und nicht für das kartesische Produkt der jeweiligen Menge mit sich selbst (within proximities) vor. In der Supermatrix bleiben die “within proximities“ nicht definiert und können sinnvoll nur als fehlende Werte behandelt werden. Angesichts der großen Menge von nicht definierten Einträgen in der Supermatrix wird diese Einbettung jedoch mit einer erheblichen Auflockerung der Datenlage erkauft. Wie stark die Supermatrix ausgedünnt ist, hängt insbesondere vom Verhältnis der Anzahl der Personen *n* zur Anzahl der Präferenzobjekte *m* ab. Für den Fall, daß beide Anzahlen die gleiche Größe aufweisen, ist die Obergrenze der Bestimmtheit der Supermatrix erreicht. Dann sind 50% der Einträge definiert, vorausgesetzt die (*n* x *m*) Matrix enthält keine fehlenden Werte. Je stärker *n* und *m* jedoch voneinander abweichen, desto ungünstiger ist das Verhältnis der Anzahlen von validen zu nicht definierten Matrixeinträgen. Die Datenlage wird zunehmend schütterer („sparse data“).

Durch das freie Antwortformat bei den verbalen Wunschäußerungen ist aber bereits in der (*n* x *m*) Matrix die übergroße Mehrheit der Einträge nicht definiert. Diese ungünstige Besonderheit der vorliegenden Daten erfährt durch die Formulierung der Supermatrix keine Verbesserung. Liegt der relative Anteil von validen Einträgen eines Zeilenvektors in der (*n* x *m*) Matrix nur bei $\frac{3}{m}$, so steigt dieser relative Anteil aufgrund der geringen Anzahl der validen Datenpunkte $f_{valide}(j)$ innerhalb der die Wünsche repräsentierenden Spaltenvektoren *j* trotz der genannten Redundanz der Supermatrix wegen $\frac{3+f_{valide}(j)}{n+m}$ nicht, sondern zeigt bei Wünschen mit sehr niedrigem Erwartungswert eine Tendenz zu noch stärkerer Verringerung.

Erfahrungen über das Verhalten von NMDS-Algorithmen auf einen derartig hohen Anteil von „Nicht-Information“ in der zugrundeliegenden Datenmatrix liegen bisher nicht vor. Dabei ist von Interesse, ob bei dieser speziellen Datenlage eine größere Gefahr der Degeneration der graphischen Lösungen als bei erprobten Datenkonstellationen zu befürchten ist.Die NMDS-Algorithmen konstruieren die räumliche Repräsentation der Daten nicht, sondern passen eine Ausgangskonfiguration von Punkten an die Ranginformation in den Daten an.Dazu werden die Unähnlichkeitsdaten p_{ij} (Dissimilaritäten) von zwei Objekten *i* und *j* einer Transformation unterzogen, um als Zielwerte der Anpassung der Distanzen d_{ij} der korrespondierenden Punkte dienen zu können.

$$F(p_{ij}) = d_{ij} + e$$

Dabei wird der Fehler *e* als Funktion der Abweichung von transformierten Dissimilaritäten $F(p_{ij})$ zu den Distanzen d_{ij}, minimiert. Eine geeignete Dissimilaritätstransformation F ist durch

den Anwender zu wählen. Besitzen die Dissimilaritäten bereits eine Metrik (Intervalskalenniveau), kann man sich für eine lineare Transformation entscheiden, das Verfahren wird dann als metrische MDS bezeichnet. Beinhalten die Daten lediglich Ranginformationen (Ordinalskalenniveau), ist eine monotone Transformation angezeigt. Diese Klasse von Verfahren heißt nonmetrische MDS (NMDS). Für die Wünsche wird keine Metrik, dagegen aber eine Ordnungsstruktur angenommen. Dementsprechend fällt die Wahl der hier anzuwendenden Modelle auf NMDS-Modelle.

Anforderungen der NMDS an die Daten

Um eine NMDS-Lösung sinnvoll interpretieren zu können, müssen die Daten den drei Eigenschaften einer Metrik genügen:

1) Die Distanz eines Punktes zu sich selbst ist $d_{ii}=0$

2) Symmetrie: $d_{ij} = d_{ji}$

3) Dreiecksungleichung: $d_{ik} = d_{ij} + d_{jk}$

Für die Wünsche und Personen kann man aus der Merkmal 1) die Forderung nach Konstanz der zu skalierenden Objekte ableiten, die hier auch angenommen wird. Die Symmetrieforderung 2) lässt bei gegebender Datenerhebung keine Prüfung zu, da mit der Nennung des Wunsches durch die Person nur eine Relationsrichtung definiert ist. Nur die Aussage „eine Person besitzt einen Wunsch“ ist in diesem Kontext sinnvoll. Dahingegen ist die Aussage „ein Wunsch besitzt eine Person“ lediglich ein sprachliches Bild und beschreibt keine Relation im Sinn der geforderten Symmetrie. Durch den Ansatz der „off-diagonal corner matrix“ (siehe S. 11) wird aber die geforderte Symmetrie in den Daten implizit formal gesichert.

Spezifikation der NMDS

Des weiteren steht die Wahl einer Minkowski-Metrik offen. Dazu ist ein Parameter r festzulegen für:

$$d_{ij} = \left[\sum_{d=1}^{k} \left| x_{di} - x_{dj} \right|^{r} \right]^{\frac{1}{r}}$$

Der Parameter k bezeichnet die Anzahl der gewählten Dimensionen, $\mathbf{x}_i$ und $\mathbf{x}_j$ sind Punkte des Raumes und d_{ij} ihre Distanz. Gebräuchliche Metriken sind die Cityblock mit (r=1) und Euklid mit (r=2).

Neben der Wahl der Metrik des Abbildungsraumes ist auch dessen Dimensionalität festzulegen. Je geringer die Dimensionalität ist, um so leichter kann die gefundene Lösung interpre-

tiert werden. Insbesondere Lösungen mit einer Dimension k > 3 entziehen sich der Anschauung. Andererseits steigt die Anpassungsgüte mit der Dimensionszahl. Zwischen Anpassungsgüte und Interpretierbarkeit muß also ein Kompromiß gefunden werden.

Shepard (1973, S. 2) gibt seine Erfahrung hinsichtlich der zu wählenden Dimension wider: ”Still, it is a fact of decisive significance that most applications for multidimensional scaling have yielded and sometimes even enlightened representations in no more than three and, indeed quite often, in only two spatial dimensions.“

Es liegen drei Familien von verwendbaren Algorithmen vor, die Methoden nach Kruskal (MDScal), Guttman und Lingoes (SSA II) und Young (ALSCAL). Die Unterschiede liegen in verschieden spezifizierten Verlustfunktionen zur Optimierung der Passung der Punkte im Raum an die Daten. Wilkinson (1995) empfiehlt die Verlustfunktionen nach Kruskal unter dem Aspekt der günstigsten Laufzeit. Der Algorithmus der Smallest Space Analysis nach Guttman und Lingoes scheint die geringste Neigung zu besitzen, vorschnell aufgrund lokaler Minima abzubrechen. Abgeraten wird von der Methode nach Young, da die jeweils größeren Distanzen eine stärkere Berücksichtigung bei der Optimierung finden als kleinere Distanzen. Die Problematik wird mit Verweis auf Weinberg und Menil (1980) deutlich, die feststellen, daß bei Dissimilaritäten eine Korrelation von der Fehlervarianz mit dem Mittelwert zu erwarten ist. Die stärkere Betonung größerer Distanzen im Anpassungsalgorithmus ist gegenüber einem Fehlereinfluß offener als ein dazu alternatives Vorgehen.

Probleme bei der NMDS-Modellierung

Anpassungsgüte

Die Passung der Relationen der Punkte in der NMDS-Lösung an die Relationen, unter denen Unähnlichkeiten bestehen, wird in der Regel durch zwei Gütemaße beschrieben: Stress und RMSQ.

Die Maßzahl Stress (STandardized REsidual Sum of Squares) entspricht dem Minimierungsparameter des Kruskalschen Algorithmus und ist im Rahmen diesen Vorgehens ein natürliches Gütemaß. Dabei werden die Abweichungsquadrate von Punktabständen und transformierten Dissimilaritäten in das Verhältnis zur Quadratsumme der Punktabstände gesetzt.

$$stress = S = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} d_{ij}^{\ 2}}}$$

Kruskal (1964, S. 3) schlägt zur Interpretation der numerischen Ausprägung von Stress eine allgemein akzeptierte Konvention vor (Tabelle 11).

Tabelle 11: Vorschlag Kruskals (1964) zur Beurteilung numerischer Stressmaße

Stress	Goodness of fit
20%	poor
10%	fair
5%	good
2½%	excellent
0%	perfect

Je geringer der Stress-Wert einer Lösung ausfällt, desto mehr Vertrauen kann bei der Interpretation in die Repräsentation gesetzt werden, weil dann die Lösung die Daten hinreichend gut wiedergibt.

Bei einem Unfolding-Modell ergibt sich eine Möglichkeit, die Anpassungsgüte des Gesamtmodells zu verbessern, indem die Minimierung von Stress nicht für alle Personen simultan (matrixkonditional), sondern vielmehr für jede Person separat (zeilenkonditional) vorgenommen wird. Das Standardmodell sieht eine einheitliche Transformationsfunktion der Dissimilaritäten für alle Personen vor. Dies beruht auf einer starken Annahme einheitlicher Urteilsprozesse bei allen Personen und führt dazu, daß eventuell existierende interindividuelle Unterschiede in der Verankerung der Einschätzung übergangen werden. Bei der sogenannten zeilenkonditionalen Optimierung, üblicherweise repräsentieren die Zeilen der Datenmatrix die Personen, wird für jede einzelne Person *i* eine eigene monotone Transformationsfunktion F_i formuliert. Auf diese Weise lassen sich die individuellen Ranginformationen mit geringerer Abweichung in die Gesamtlösung einbringen und das gesamte Stressmaß verringert sich unter Umständen gegenüber der matrixkonditionalen Optimierung.

Lokale Minima im Ablauf der Optimierung

Ein Einfluß der Startkonfiguration auf die resultierende Lösung kann nicht ausgeschlossen werden, obwohl natürlich ein robuster Algorithmus immer auf dasselbe Ergebnis konvergieren sollte. Tatsächlich ist eine gewisse Abhängigkeit von Startkonfigurationen nachgewiesen (Cox und Cox 2001). Führen unterschiedliche Startkonfigurationen zu verschiedener Anpassungsgüte, so ist dies ein ernsthafter Hinweis auf ein Problem lokaler Minima. In diesem Fall bricht der Algorithmus vorzeitig die Optimierung ab, da scheinbar die formalen Kriterien zum

Abbruch erfüllt wurden, obwohl das globale Minimum der Abweichung von Modell und Daten noch nicht erreicht ist.

Borg (1984) schlägt zur Vermeidung lokaler Minima einer Alternierung von monotoner und linearer Transformation der Dissimilaritäten pro Iterationsschritt im Anpassungsalgorithmus vor. Eine solche Spezifikation des Ablaufs des Algorithmus kann jedoch meines Wissens nur in KYST vorgenommen werden und setzt eine hohe Vertrautheit mit dieser Software voraus.

Heiser und Groenen (1997) entwickelten einen Majorisationsalgorithmus zur Vermeidung lokaler Minima, der in der Prozedur PROXSCAL in SPSS ab Version 10.0 Anwendung findet (Groenen 1999). Allerdings kann PROXSCAL leider vorerst noch kein Unfolding modellieren. Somit scheint es zur Zeit die zugänglichste Form der Problembehandlung lokaler Minima zu sein, einen multiplen Programmstart mit verschiedenen Startwerten vorzunehmen, wie es SPSS und auch SYSTAT anbieten, und die Lösung mit den günstigsten Anpassungswerten zur Interpretation heranzuziehen.

Unvollständige Daten

MacCallum (1978) untersuchte zumindest für den ALSCAL-Algorithmus den Einfluß unvollständiger Datenpunkte (20%-60% fehlende Werte) auf die Güte einer NMDS-Lösung bei gleichzeitiger Berücksichtigung anderer Einflußfaktoren wie Zufallsfehler, Anzahl der Urteiler und Gewichtung der Urteile. Die Anzahl beurteilter Stimuli und die Dimensionszahl ($k = 3$) waren dabei gegeben und konstant. Es wurde sichtbar, daß eine Gefährdung der Rekonstruktion der ursprünglichen Lösung bei 60% nach Zufallsprinzip gelöschten Datenpunkten (sparse data) nur unter der weiteren Bedingungen auftritt, daß eine geringe Anzahl von Urteilern vorliegt. Es scheint aber so zu sein, daß durch eine geeignet große Anzahl von Urteilern diese Verzerrung kompensiert werden kann.

Homogenitätsanalyse

Die Homogenitätsanalyse (HOMALS, Gifi, 1981, 1990; Meulman, 1982, 1986; de Leeuw, 1986; Matschinger, 1989) beruht auf zwei Kerngedanken. Einer bezieht sich auf Homogenität, ein weiterer auf optimale Kategorienscores.

Homogenität bezieht sich auf das Verhältnis latenter Skalen $\mathbf{x}_k$, als Objektvektoren bezeichnet, zu manifesten Skalen $\mathbf{h}_j$, die als Beobachtungsvektoren, also Variablen dienen. Die Vektoren sind durch dieselben Datenquellen – Personen - indiziert.

Im eindimensionalen Fall heißt diese Indizierung maximal homogen, wenn die einzelnen Beobachtungsvektoren $\mathbf{h}_j$ einem latenten Objektvektor $\mathbf{x}$ entsprechen.

$$\mathbf{h}_j = \mathbf{x} \qquad \forall \mathbf{h}_j \quad \text{mit } j = \{1,...,m\}$$

Daraus läßt sich eine Forderung minimaler Abstände und eine Verlustfunktion $\mathbf{L}(\mathbf{x})$ ableiten, mit deren Hilfe sich latente Skalenvektoren aus einer Menge von Beobachtungsvektoren schätzen lassen.

$$\mathrm{L}(\mathbf{x}) = m^{-1} \sum_{j=1}^{m} (\mathbf{x} - \mathbf{h}_j)'(\mathbf{x} - \mathbf{h}_j)$$

Wenn die Kategorisierung der Ursprungsvariablen per fiat vorgenommen wurde, kann nicht mit Sicherheit davon ausgegangen werden, daß die vorgenommene Kodierung den Gegebenheiten der Daten entspricht. Kategorisierungen per fiat sind nicht eindeutig, das heißt, es gibt verschiedene Möglichkeiten, den Kategorien Zahlen zuzuordnen. Diese Etikettierung berührt das Maß an Homogenität, wie aus ihrer Definition als minimale Differenz hervorgeht. Am selben Datensatz ergeben sich unter Umständen verschiedene Homogenitätswerte in Abhängigkeit von der jeweils gewählten Kodierung. Optimal heißen in diesem Zusammenhang Kategorien, unter deren Anwendung die Homogenität maximal ist. Eine per fiat vorgenommene Kategorisierung kann nur als vorläufig angesehen werden, solange nicht belegt wurde, daß sie optimal ist. Was liegt also näher, als optimale Kategorienscores zu bestimmen und die willkürlichen, per fiat eingesetzten Werte zu ersetzen?

Dazu wird die Matrix **H** der Beobachtungsvektoren zunächst mittels einer Art Dummy-Kodierung in eine sogenannte Indikator- oder Burtmatrix **G** ohne Verlust an Information umge-

wandelt. Wenn für einen Beobachtungsvektor $\mathbf{h}_j$ die Anzahl der Kategorien mit $c(\mathbf{h}_j)$ bezeichnet sei, so ist **G** eine (n x $\sum_{j=1}^{m} c(\mathbf{h}_j)$)-Matrix, wenn **H** eine (n x m)-Matrix ist.

Durch die Vektormultiplikation von **G** mit einem geeignetem Vektor von Kategorienscores **y** ergibt sich wiederum die Koinzidenzmatrix **H**.

$$\mathbf{H} = \mathbf{G}\mathbf{y}$$

Ein Beispiel für die Beziehung einer Datenmatrix zu einer Indikatormatrix liefern die Tabellen 11 und 12.

Tabelle 11: Eine (n x 2)-Datenmatrix **H** mit zwei Beobachtungsvektoren $\mathbf{h}_j$ und *n* Personen

	$c(\mathbf{h}_1)=3$	$c(\mathbf{h}_2)=2$
Person 1	1	2
Person 2	3	1
:	:	:
Person n	2	2

Tabelle 12: Die zur Tabelle gehörige (n x 5)-Indikatormatrix **G**, gegeben **y**' = [1,2,3,1,2]

	bezgl. $\mathbf{h}_1$			bezgl. $\mathbf{h}_2$	
	g_{11},	g_{12},	g_{13},	g_{21},	g_{22},
Person 1	1	0	0	0	1
Person 2	0	0	1	1	0
:	:	:	:	:	:
Person n	0	1	0	0	1

Dementsprechend ergibt sich ein Beobachtungsvektor $\mathbf{h}_j$ als Linearkombination einer Teilindikatormatrix $\mathbf{G}_j$ mit einem Teilvektor von Kategorienscores $\mathbf{y}_j$.

$$\mathbf{h}_j = \mathbf{G}_j \mathbf{y}_j$$

Der direkten Schätzung optimaler Kategorienscores steht jedoch im Wege, daß dafür die latenten Skalenwerte unter der Bedingung maximaler Homogenität bekannt sein müssen. Dies ist von vornherein jedoch eben nicht der Fall. Zur Bestimmung von latenten Skalenwerten

unter der Bedingung der Homogenität sind wiederum optimale Kategorienscores nötig. Die Verlustfunktion der Homogenität wird somit auf die optimalen Kategorienscores erweitert.

$$\mathrm{L}(\mathbf{x};\mathbf{y}) = m^{-1}\sum_{j=1}^{m}(\mathbf{x}-\mathbf{G}_j\mathbf{y}_j)'(\mathbf{x}-\mathbf{G}_j\mathbf{y}_j)$$

Gelöst werden kann dieses Problem durch einen iterativen Algorithmus, genannt "alternating least square" (ALS). Dabei wird der Zusammenhang zwischen den beiden unbekannten Vektoren ausgenutzt. Beginnend mit geeigneten Startwerten behandelt der Algorithmus jeweils einen der zu schätzenden Vektoren als konstant und schätzt den anderen. Im darauffolgenden Schritt wird dieser geschätzte Wert konstant gehalten und auf dieser Basis ein Wert für die erste Größe ermittelt. Dieser Wechsel im Konstanthalten und Schätzen wird solange durchgeführt, bis ein Konvergenzkriterium erfüllt ist.

Es resultieren zwei Mengen von synthetischen Variablen. Die eine Menge bezieht sich auf die *n* Objekte, damit sind die Personen bezeichnet, die andere Menge bezieht sich auf die c ($\mathbf{h}_j$) Kategorien von Variablen. Die Besonderheit des Verfahrens besteht also darin, daß nicht direkt die beobachteten Messungen als Punkte im Bildraum dargestellt werden, sondern die jeweiligen Ausprägungen der Messung. Die Spanne, die diese optimalen Scores umfassen, gilt als Maß der Beteiligung der beobachteten Messung an der Ausdehnung des latenten Bildraumes und damit als Maß der Bedeutsamkeit dieser Variable für die HOMALS-Lösung. Diese Spanne wird als Diskrimination einer Variable bezeichnet.

Es gibt eine Verwandtschaft der Diskrimination im Rahmen von HOMALS mit dem Konzept der Ladung einer Variable auf einen Faktor innerhalb eines faktoranalytischen Modells. Ähnlich wie bei Ladungen ist es gerechtfertigt, Variablen mit numerisch hoher Diskrimination als Marker zur Interpretation der konstruierten Dimensionen der Lösung heranzuziehen.

Das Verfahren hat starke Parallelen zur Multiplen Korrespondenzanalyse (MCA, Greenacre 1978, 1984). Beide Verfahren bilden eine Koinzidenzmatrix in einen niedrigdimensionalen Raum ab. Bei der MCA geht man algorithmisch allerdings den Weg einer Single-Value-Dekomposition. Abhängig von der a priori durch den Anwender festzulegenden Dimensionszahl *k* werden Eigenvektoren für die *k* größten Single Values der Indikatormatrix, die man als *k* Eigenwerte des Bildraumes benutzt, sowohl für die Zeilen als auch für die Spalten der Koinzidenzmatrix bestimmt. Die Indikatormatrix wird somit linear in zwei niedrig-dimensionale Räume, einen Zeilenbildraum und einen Spaltenbildraum, unter der Bedingung projiziert, daß die Abstände zwischen den Paaren jeweils Zeilen respektive Spalten darstellenden Punkte in linearer Funktion des χ^2 -Wertes der beiden Häufigkeitsprofile von Zeilen respektive Spalten stehen.

Das Problem der MCA besteht in erster Linie in der Verbindung der beiden separaten Bildräume. Im sogenannten “French plot“, bei dem der Zeilen- mit dem Spaltenbildraum in eine Abbildung gebracht wird, warnen eine ganze Reihe von Autoren (dazu Greenacre 1988) vor der Interpretation der Distanzen zwischen Zeilen- und Spaltenpunkten. Allerdings zeigten Caroll, Green und Schaffer (1986, 1987, 1988), daß eine sinnvolle Interpretation in erster Linie durch ein numerisches Gefälle zwischen den Single Values gefährdet wird. Durch eine geeignete Transformation, wie die nach den Autorennamen bezeichnete CGS-Skalierung, läßt sich jedoch ein solcher Malus ausgleichen.

Als Maß der Modellgüte wird für HOMALS ein Maß des Unterraumes benutzt. Da als Bildraum der Standardraum mit einer Determinante $d = 1$ gewählt wird, kann als Referenz die Summe der Eigenwerte genutzt werden. Liegt die Summe ebenfalls bei 1, so diskriminieren die Items maximal und die Passung von Daten und Bildraum ist perfekt. Bei totaler Degeneration, wenn also Personen- und Bildpunkt auf denselben Punkt abgebildet werden, liegt die Summe der Eigenwerte bei Null. Das Modell hätte dann minimale Passung an die Daten. Günstig ist ein Wert, der numerisch nahe bei Null liegt. Da als Gesamtfitmaß ein additiver Term gewählt wird, ist es möglich, den Beitrag jeder einzelnen Dimension an der Gesamtanpassung zu identifizieren. Darüber hinaus kann der Beitrag jedes Items zum Modell anhand des Diskriminationsmaßes eingeschätzt werden, das in Funktion der Spanne der optimalen Kategorienscores steht. Die genannten Gütemaße lassen sich analog zu den Parametern einer Hauptkomponentenanalyse (PCA) verstehen.

Tabelle 12: Analogie von Parametern einer Hauptkomponentenanalyse und den Gütemaßen von HOMALS

PCA	HOMALS
durch Faktoren aufgeklärte Gesamtvarianz der Items	Summe der Eigenwerte
Kommunalität eines Faktors	Eigenwert einer Dimension
Ladung eines Items	Diskriminationsmaß eines Items

Hierarchische Klassenanalyse

Mit HICLAS (Hierarchical Classes) stellen De Boeck und Rosenberg (1988) ein diskretes, kategoriales Modell vor. De Boeck und Rosenberg (1988) gehen davon aus, daß die einfachste Beschreibung von *n* Objekten (beispielsweise Personen) mittels einer Liste von *m* Attributen vorgenommen werden kann. Wenn das Objekt *i* ein entsprechendes Attribut *j* besitzt, so liegt es nahe, genau dann einen Eintrag (m_{ij}=1) für alle $i \in \{1,...,n\}$ und alle $j \in \{1,...,m\}$ in einer (*n* x *m*)-Matrix **B**, vorzunehmen, anderenfalls den Eintrag (m_{ij}=0). Die resultierende (0,1)-Matrix **B** kann als Element der Boolschen Matrizen aufgefaßt werden. Boolsche Matrizen haben als Einträge Elemente einer Boolschen Algebra. Das heißt, daß folgende Eigenschaften für die Menge der Codierungen C= {0,1} vorausgesetzt werden:

C ist nicht leer und zwei binäre Operationen (+, ·) sind auf C definiert:

Tabelle 15: Schema der binären Boolschen Operationen (+) und (·)

+	0	1
0	0	1
1	1	1

.	0	1
0	0	0
1	0	1

1) Beide Operationen sind kommutativ
   $(a + b = b + a)$ und $(a \cdot b = b \cdot a)$ für $\forall a,b \in \{0,1\}$

2) Für beide Operationen gilt Distributivität
   $a + (b \cdot c) = (a + b) \cdot (a + c)$ und $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ für $\forall a,b \in \{0,1\}$

3) Für jede Operation existiert ein eigenes neutrales Element mit
   $(a + 0 = a)$ und $(a \cdot 1 = a)$ für $\forall a \in \{0,1\}$

4) Für $\forall a \in \{0,1\}$ existiert ein komplementäres Element a^c mit
   $(a + a^c = 1)$, $(a \cdot a^c = 0)$

Da mit 0 das neutrale Element der Operation „+“ und mit 1 das neutrale Element der Operation „·“ bezeichnet ist, ergibt sich für beide Operationen aus der Eigenschaft 4) ein essentieller Unterschied einer Boolschen Algebra zur Struktur einer mathematischen Gruppe. Boolsche Matrizen sind nicht über einen Körper, sondern lediglich über einen Semiring definiert und dürfen somit nicht als ein Vektorraum, wie beispielsweise die im Rahmen des ALM verwendeten Matrizen mit Einträgen aus dem Körper der reellen Zahlen, verstanden werden. Kon-

zepte der linearen Algebra, die bei der Analyse solcher Vektorräume, wie sie in der Psychologie im Rahmen des ALM verwendet werden, sind auf dieses System nicht direkt übertragbar.

Dies gilt insbesondere für das wichtige Konzept des Ranges einer Matrix. $\rho_r(\mathbf{B})$, der Zeilen - oder der Spaltenrang, $\rho_c(\mathbf{B})$, einer Boolschen Matrix **B** soll eine Untergrenze besitzen, den sogenannten Schein-Rang $\rho_S(\mathbf{B})$, mit

$$\rho_S(\mathbf{B}) \leq \rho_r(\mathbf{B}) \leq \rho_{\max}(\mathbf{B}); \rho_S(\mathbf{B}) \leq \rho_c(\mathbf{B}) \leq \rho_{\max}(\mathbf{B})$$

und $\rho_{\max}(\mathbf{B})$ als maximalen Rang der Matrix.

Eine Boolsche Matrix **B** sei als komponentenweise Summe von Kreuzvektoren $\mathbf{K}_l$ darstellbar:

$$\mathbf{B} = \bigoplus_{l=1}^{k} \mathbf{K}_l .$$

Kreuzvektoren **K** ergeben sich für geeignete Boolsche Vektoren **v**, **w** aus

$\mathbf{K} = \mathbf{v}\mathbf{w}^{\mathbf{T}}$ durch die Boolsch-multiplikative Operation $k_{ij} = v_i \cdot w_j$

mit $i \in \{1,...,n\}$ und $j \in \{1,...,m\}$ und sind somit Boolsche (n x m)-Matrizen.

Der Schein-Rang ist als die minimale Anzahl von Summanden $k \rightarrow \min$ bezüglicher dieser komponentenweisen Addition definiert

$$\rho_S(\mathbf{B}) = k.$$

Eine Boolsche Matrix kann somit auf der Basis komponentenweiser Boolscher Addition in eine minimale Anzahl von k Matrizen dekomponiert werden, die als Kreuzvektoren definiert wiederum auf der Basis Boolscher Multiplikation in k Paare von Boolschen Vektoren dekomponierbar sind

$$\mathbf{B} = \bigoplus_{l=1}^{k} (v_{il} \cdot w_{jl})$$

Die Kreuzvektoren werden im Rahmen von HICLAS in der psychologischen Anwendung beispielsweise durch geordnete Paare von personenbezogenen und kategorienbezogenen Boolschen Vektoren konstituiert:

$\mathbf{K}_l = \mathbf{p}_l \mathbf{s}_l^{\mathbf{T}}$ mit der Anzahl der untersuchten Personen n - $i \in \{1,...,n\}$- und der Anzahl der Kategorien oder Attribute m - $j \in \{1,...,m\}$. Der Lösungsraum für solche Dekompositionen wird in drei Hinsichten restringiert:

1) Assoziation
2) Äquivalenz
3) hierarchische Ordnung

Personen und Kategorien sind genau dann assoziiert, wenn die jeweils *q*-te Stelle der Personen- und Kategorienvektoren einen 1-Eintrag aufweist.

$$p_{ql} = s_{ql} = 1$$

Personen *h* und *i* sind genau dann äquivalent, wenn die mit *h* und *i* assoziierten Mengen von Kategorien identisch sind. Vice versa spricht man von äquivalenten Kategorien *g* und *j* genau dann, wenn die Menge der mit *g* respektive *j* assoziierten Mengen von Personen identisch sind. Personen *h* und *i* stehen in einer hierarchischen Ordnung genau dann, wenn die mit *h* assoziierte Menge von Kategorien eine echte Untermenge der mit *i* assoziierten Kategorienmenge darstellt. Dies gilt analog auch für hierarchisch geordnete Kategorien (van Mechelen und de Boeck, 1995). Existieren Personen- und Kategorienvektoren mit den genannten Bedingungen, so liegt eine HICLAS-Lösung vor. Diese kann eindeutig in der Form eines spezifischen Graphen wiedergegeben werden. Der Graph besteht aus zwei Teilgraphen, jeweils für die Personen- und die Kategorienmenge. Die Knoten, die für assoziierte Personen und Kategorien stehen, sind durch eine Kurve verbunden. Eine Menge äquivalenter Personen wird durch genau einen Knoten dargestellt. Dasselbe gilt entsprechend für äquivalente Kategorien. Eine hierarchische Ordnung ist an den Ebenen des Graphen und den Kantenzügen zwischen Knoten der in Ordnungsrelation stehenden Personen oder Kategorien ersichtlich.

Zur Erläuterung sei hier ein Beispiel nach van Mechelen und de Boeck (1995) dargestellt. Ausgehend von einer Rohdatenmatrix (Tabelle 13) ergibt sich bei einem Schein-Rang $k = 4$ der Graph in Abbildung 13.

Tabelle 13: Fiktive binäre Rohdaten für eine disjunktive HICLAS-Lösung (nach van Mechelen und de Boeck, 1995, S. 511)

	Kategorien			
Personen	a	b	c	d
1	1	0	0	1
2	0	1	0	1
3	1	1	1	1
4	0	0	0	1

Jede der vier Personen ist mindestens mit einer Kategorie assoziiert. Wenn das nicht der Fall wäre, würde die entsprechende Person im Graphen in einem Knoten ohne Verbindung repräsentiert werden. Dies gilt in analoger Weise für die Kategorien.

Im Beispielgraphen in Abbildung 13 gibt es keine äquivalenten Personen oder Kategorien. Wenn der Knoten also mit einem beobachteten Objekt besetzt ist, so ausschließlich singulär.

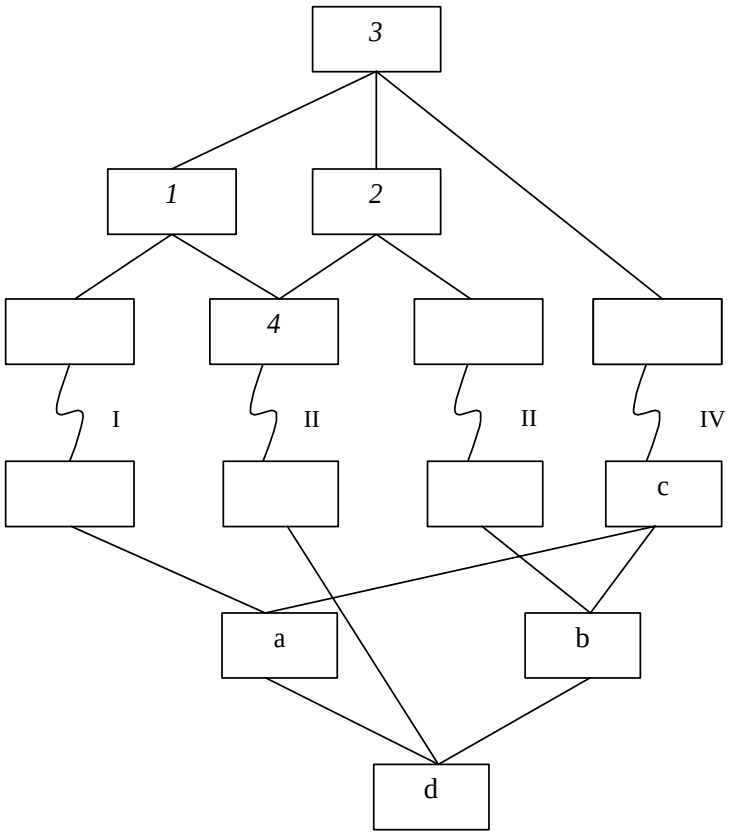


Abbildung 13: Beispiel eines HICLAS-Graphen mit einem Schein-Rang $k = 4$ (nach van Mechelen und de Boeck, 1995, S. 512)

Die durch Kurven zwischen Person- und Kategorienknoten verdeutlichten Assoziationen werden durch genau einen Kreuzvektor generiert. Dieser Bereich des Graphen wird als Bündelung (bundle) bezeichnet und mit römischen Ziffern beschriftet. Die Anzahl der Bündelungen entspricht dem Schein-Rang *k*. Es existiert hier eine hierarchische Ordnung sowohl bei den Personen als auch bei den Kategorien. Die Person *3* ist den Personen *1* und *2* übergeordnet. Auf der untersten Hierarchieebene befindet sich die Person *4*. In ähnlicher Weise ist die Kategorie *d* den Kategorien *a* und *b* vorgeordnet, während *c* auf der untersten Ebene steht. Ob eine hierarchisch übergeordnete Person oder Kategorie existiert, stellt einen empirischen Befund dar und ist keine Modellkomponente. Leere Knoten werden nur dann berücksichtigt, wenn ihnen ein Knoten mit einem beobachteten Element übergeordnet ist. Diese übergeordneten Personen oder Kategorien liefern den empirischen Beleg für die spezifische Vorhersage, da in ihren Assoziationen diejenigen Assoziationen der durch die leeren Knoten prädizierten Personen oder Merkmale als Untermenge enthalten sind.

Sechs Knoten im Graphen von Abbildung 13 sind unbesetzt. Unter Geltung der HICLAS-Lösung werden diese bestimmten Antwortmuster der Personen wie auch die entsprechenden Kategorienmustern durch das Modell vorhergesagt. Die Bedeutung der leeren Knoten kann über die Differenz der Menge von Assoziationen des übergeordneten Knotens und der Menge

der Assoziationen aller Knoten auf der Hierarchieebene, denen derselbe Knoten übergeordnet ist, erklärt werden.

Das entscheidende Problem bei der Konstruktion einer HICLAS-Lösung liegt darin, den Schein-Rang zu bestimmen. Da bei empirischen, also fehlerbehafteten Daten in der Regel nicht erwartet werden kann, daß mit einem Schein-Rang von ($\rho_S(\mathbf{B}) \leq 5$) die Rohdatenmatrix vollständig im Rahmen einer HICLAS-Lösung reproduziert werden kann, wird die Bedeutung des Schein-Ranges dementsprechend modifiziert. Der Schein-Rang einer HICLAS-Lösung soll die minimale Anzahl von Kreuzvektoren sein, die eine hinreichende Anpassung an die Rohdatenmatrix gewährt. Hinreichend sei die Anpassung genau dann bei ($\rho_S(\mathbf{B}) = k$) und ($k \geq 1$), wenn bei mehrfacher Analyse mit inkrementiertem Schein-Rang bei Schritt $k + 1$ kein wesentlicher Anpassungszuwachs im Vergleich zur Lösung mit ($\rho_S(\mathbf{B}) = k$) erzielt werden kann. Graphisch läßt sich die Entscheidung über den gewählten Schein-Rang anhand des Screeplots eines geeigneten Gütemaßes darstellen. Als Gütemaß bietet sich der Jaccard-Index an, es kann aber auch die Anzahl der Übereinstimmungen von Daten und Lösung benutzt werden.

Anlaß zur Formulierung von HICLAS sind die Vorarbeiten von Gara und Rosenberg (1979) und Gara (1985), die sich mit einem Problem der Personenwahrnehmung befassen. Exemplarisch für eine erfolgreiche Modellierung in HICLAS sollen hier kurz die Ergebnisse einer Reanalyse durch de Boeck und Rosenberg (1988) wiedergegeben werden. Ausgangspunkt dieses Ansatzes ist das Phänomen, daß Personen bei frei wählbaren Kategorien zur Beschreibung ihrer sozialen Umwelt einer Minderheit von Personen quasi den vollen Satz von verwendeten Kategorien zuschreiben, während für die Mehrzahl der umgebenden Personen nur eine Teilmenge der verwendeten Kategorien benutzt wird.

Daraus ergeben sich zwei gekoppelte Fragestellungen:

1) Gibt es einige Merkmale, die andere Merkmale implizieren?

2) Gibt es Prototypen von Persönlichkeiten, das heißt eine Teilmenge von Personen, die mit einem Satz von Merkmalen beschrieben werden kann, der auch für alle übrigen Personen relevant ist?

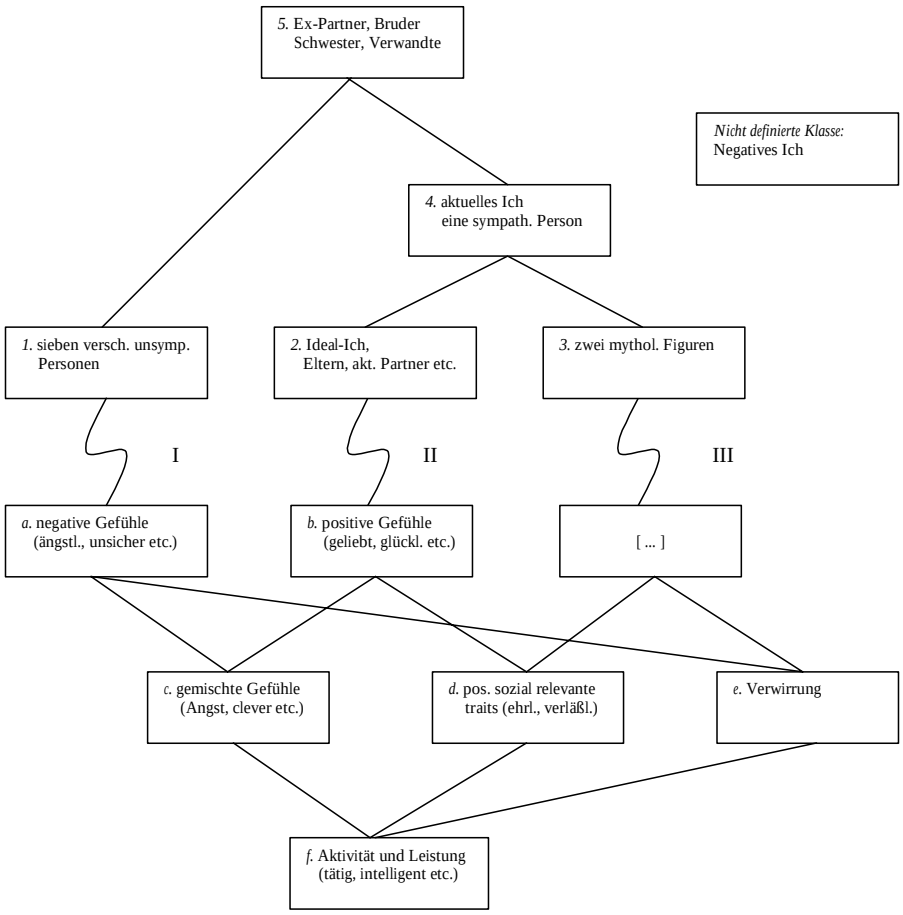


Abbildung 14: HICLAS-Graph der Reanalyse von Personenbeschreibungsdaten (Gara und Rosenberg, 1979; nach de Boeck und Rosenberg, 1988, S. 376)

Zur Beantwortung beider Fragestellungen legten Gara und Rosenberg (1979) 14 Probanden eine Liste von 36 sozialen Rollen vor. Zunächst identifizierten die Versuchspersonen jede dieser Rollen mit einer konkreten Person. Dann beschrieben sie diese Personen verbal in freiem Antwortformat. Die Beschreibungen wurden kategorisiert. Es traten dabei auf den ersten Blick zwei Gruppen von Beschreibungen auf, zum einen Persönlichkeitsmerkmale im Sinne von Traits und zum anderen Gefühle, die durch die jeweilige Person bei den Probanden hervorgerufen wurden. Der HICLAS-Graph des am besten passenden Modells mit einem Schein-Rang $k = 3$ der Reanalyse durch de Boeck und Rosenberg findet sich in Abbildung 14. Werden zunächst die zugeschriebenen Merkmale der Personen betrachtet, so ist die Äquivalenzklasse *f* mit Leistungsmerkmalen wie Intelligenz und Aktivität die allgemeinste. Soziale Kompetenzen, Klasse *d*, sind dieser untergeordnet und damit als Kriterien ausgewiesen, mit denen nur eine bestimmte Teilmenge der Personen beschrieben wird. Nach diesen hierarchisch geordneten Trait-Klassen findet sich Klasse *b* mit positiven Gefühlsäußerungen, die wiederum auf einen noch eingeschränkteren Personenkreis angewendet werden. Ähnlich restriktiv ist die Klasse *a*, die negative Emotionen beinhaltet. Dieser sind die allgemeineren Klassen *c* und *e* vorgeordnet. Die beiden Klassen beziehen sich auf Vorbehalte in der Personenbeschreibung, die, ebenfalls wie auch die Klassen *d* und *b*, von den Aspekten der Leistungsbe-

schreibung von Klasse *f* dominiert werden. Klasse *c* vereint Kriterien, die zusammen eine gewisse Ambivalenz ausdrücken, Klasse *e* ist eher Ausdruck der emotionalen Schwierigkeit, sich ein Bild von einer Person zu verschaffen.

Auf der Personenseite des Graphen läßt sich eine Klasse nachweisen, für die alle verwendeten Kriterien der Beschreibung herangezogen werden (Klasse *5*). Ihr nachgeordnet findet sich eine Klasse mit sieben allgemein als unsympathisch empfundenen Personen (Klasse *1*). Diese Klasse von Personen ist mit der Klasse *a* der Beschreibungsmerkmale, den negativen Gefühlen assoziiert. Ebenfalls der allgemeinen Klasse *5* ist jene Klasse nachgeordnet die unter anderem das Real-Selbstbild beinhaltet (Klasse *4*). Sie dominiert Klasse *2* mit dem Ideal-Selbstbild und den Personen des nächsten sozialen Umfeldes wie Partner, Eltern und Freunde sowie Klasse *3* mit zwei gewählten mythologischen Figuren. Klasse *2* ist mit Merkmalsklasse *b* der positiven Gefühlen assoziiert. Die Klasse *3* mit den mythologischen Figuren steht mit einer leeren Klasse von Merkmalen in Assoziation, die allerdings von der Klasse *d*, die der positiven sozialen Traits, und der Klasse *e*, die Verwirrung im Urteil ausdrückt, bestimmt wird. Ein wichtiges Ergebnis liegt auch in der undefinierten Klasse, die das negative Selbstbild enthält. Offensichtlich finden sich zur Beschreibung dieser imaginären Person keine Kriterien, die zur Beschreibung anderer Personen benutzt werden.

Die wichtigen Personen des nahen sozialen Umfeldes werden also in der Regel nicht an den negativ konnotierten Kriterien der Klasse *a* gemessen, so wie auch unsympathische Personen nicht anhand des Auftretens positiver Gefühle beurteilt werden. Personen, bei denen anzunehmen ist, daß sowohl angenehme als auch unangenehme Erfahrung (bsw. Trennung vom Ex-Partner, Geschwisterrivalitäten) vorliegen, werden allerdings mit Hilfe beider Kriterienklassen beschrieben.

Zu vermuten ist, daß ein Schein-Rang von $k = 2$ hinreichend sein könnte, wenn nicht simultan zu realen Personen auch noch mythologische Gestalten zu beschreiben gewesen wären. Würde das Bündel mit der entsprechenden Klasse nicht mehr Bestandteil des Modells sein, gäbe es nur noch zwei Bündel für sympathische und unsympathische Personen. So würde noch deutlicher ersichtlich werden, daß eine starke Interaktion von der Wahl emotionalen Kriterien bei der Beschreibung von Personen mit der allgemeinen Sympathie gegenüber den zu beschreibenden Personen vorliegt.

Feature Pattern Analysis

Der Entwicklung der Feature Pattern Analysis (FPA, Merkmalsmusteranalyse; Feger, 1988) liegen zwei Anliegen zugrunde. Einerseits geht es darum, geometrische Skalierverfahren wie zum Beispiel eine Guttman-Skala oder eine Parallelogrammanalyse datentheoretisch zu fundieren. Zum anderen soll eine Alternative in der bisher unbefriedigenden methodologischen Situation hinsichtlich der Klassifikation von Objekten geboten werden. Ein Problem stellt unter anderem das zweistufige Vorgehen bei clusteranalytischen Verfahren dar. Dabei werden zunächst paarweise Zusammenhangsparameter zwischen den Objekten geschätzt und dann diese Koeffizienten einem Gruppierungsalgorithmus übergeben. Problematisch ist in erster Linie die Auswahl des Zusammenhangsmaßes. Hier gibt es einen erheblichen Umfang an Freiheitsgraden in der numerischen Kodierung der Zusammenhänge, die letztlich immer auf eine Reduktion der vorliegenden Information hinausläuft. Ein Teil der Information wird zum Fehler deklariert und spielt ab dieser Deklaration für die Weiterverarbeitung im zweiten Schritt keine Rolle mehr.

Ungünstigenfalls kann dieser deklarierte Fehleranteil strukturrelevant sein. Durch den unglücklichen und unbemerkten Ausschluß von möglicherweise relevanten Datenaspekten im ersten Schritt des Verfahrens sind die Ergebnisse des zweiten, des Gruppierungsschrittes von fragwürdigem Wert. Um dies erkennen zu können, müßte die Struktur der Daten im vorhinein bekannt sein. Da aber die Struktur den Gegenstand von Klassifikationsverfahren bildet, haftet der Entscheidung für ein bestimmtes Zusammenhangsmaß wegen der damit verbundenen Unwägbarkeiten etwas Willkürliches an.

Vorteilhafter ist es, wenn die Zusammenhangsinformationen direkt in das zu spezifizierende Modell eingehen. Bei diesem Vorgehen ist zu erwarten, daß in erheblichem Umfang redundante Information anfällt. Dieser zunächst hinderlich erscheinende Umstand kann sich jedoch als sehr förderlich bei der Identifikation der zu untersuchenden Datenstruktur erweisen.

Die zentrale Idee der FPA liegt in der Identifikation einer Variable (Merkmal, Item) mit einem Unterraum der Dimension dim = (k - 1) in einem Repräsentationsraum mit der Dimension dim = k.. Eine erfolgreiche Repräsentation ist genau dann möglich, wenn die Kontingenzen der beteiligten Items zwei Kriterien, dem Nullzellen- und dem Konsistenzkriterium, genügen. In diesem Fall spricht man von einer Lösung.

Eindimensionale Lösungen

Die einfachste Repräsentation im Raum mit der Dimension $k = 1$ basieren auf Zusammenhängen von jeweils zwei bivariaten Items. Hier lassen sich die Merkmale als Punkte im Sinn von Intervallgrenzen auf einer Geraden darstellen.

Zur Illustration seien drei dichotome Items oder Merkmale A, B, C mit den Kategorien oder Ausprägungen $a, b, c \in \{0,1\}$ gegeben. Kontingenztafeln dieser drei Items finden sich in Tabelle 14.

Es sind vier Muster der zwei Ausprägungen zweier Merkmale kombinatorisch möglich. Auf einer Geraden mit zwei Intervallgrenzen lassen sich jedoch nur drei Muster repräsentieren. Dabei lassen sich vier Fälle unterscheiden (Abbildung 15).

Mögliche Anordnungen zweier Items in $k = 1$		Nicht repräsentierbares Muster *ab*
1)	A: 0 | 1 — B: 0 | 1	*01*
2)	A: 0 | 1 — B: 1 | 0	*00*
3)	A: 1 | 0 — B: 0 | 1	*11*
4)	A: 1 | 0 — B: 1 | 0	*10*

Abbildung 15: Alle vier möglichen Anordnungen zweier bivariater Items auf einer Dimension: jeweils ein Itemmuster *ab* ist nicht darstellbar.

Das nicht darstellbare Muster jeder Anordnung bestimmt die Orientierung der beiden Items im Raum bis auf die Leserichtung der Itemgrenzen (von links oder von rechts). Das nicht darstellbare Muster wird anhand der Zwei-Weg-Kontingenztafel der betreffenden Items bestimmt. Eine der Kategorienkombinationen muß die Häufigkeitsbesetzung von Null aufweisen. Dieses nicht beobachtete Item- oder Merkmalsmuster soll Nullzelle heißen.

Tabelle 14: Alle drei möglichen Kontingenztafeln von drei bivariaten Items A, B, C mit fiktiven Zellbesetzungen.

A x B

f_{ab}		Item B	
		0	*1*
Item A	*0*	25	**0**
	1	50	25

A x C

f_{ac}		Item C	
		0	*1*
Item A	*0*	**0**	25
	1	25	50

B x C

f_{bc}		Item C	
		0	*1*
Item B	*0*	25	25
	1	**0**	50

Beispielsweise muß für eine Zelle der Kontingenztafel der Items A und B gelten ($f_{ab} = 0$). Ist dies für alle untersuchten Kontingenzen der Fall, so ist das sogenannte Nullzellenkriterium der FPA erfüllt, und die beiden jeweiligen Items dieser Kontingenzen sind im selben Raum der Dimension $k = 1$ repräsentierbar. Allgemein werden zur Prüfung des Nullzellenkriteriums (k + 1)-Weg-Kontingenzen über alle jeweils betrachteten m Items herangezogen. Die Prüfung bezieht sich somit auf $\binom{m}{k+1}$ Kontingenzen.

Ist das Nullzellenkriterium erfüllt, wird in einem nächsten Schritt das zweite, das Konsistenzkriterium der FPA geprüft. Allgemein bezieht sich die Überprüfung auf die durch die (k +1) - Weg - Kontingenzen definierten Nullzellen von jeweils (k + 2) Items. Im eindimensionalen Beispiel sei also für jedes Itempaar das Nullzellenkriterium erfüllt, das heißt, es existiert eine Nullzelle in der Kontingenztafel für jedes der drei Itempaare. Diese drei Nullzellen sind genau dann konsistent, wenn zwei der an der Bezeichnung der Nullzellen beteiligten Merkmalskategorien konstant bleiben und eine Bezeichnung alterniert. In Tabelle 15 sind die Merkmalskategorien *a* und *b* konstant, während *c* in der Ausprägung geändert ist. Sind die Nullzellen in A x B (*01*) und B x C (*10*), so fordert die Erfüllung des Konsistenzkriteriums für A x C (*00*).

Tabelle 15: Die Nullzellen von Zweierkontingenzen für die drei Items aus Tabelle 14 erfüllen das Konsistenzkriterium einer FPA-Lösung im eindimensionalen Raum.

	Itemkategorie		
zwei-Weg-Kontingenz	*a*	*b*	*c*
A x B	0	1	-
B x C	-	1	0
A x C	0	-	1

Daraus folgt, daß die Itemorientierungen widerspruchsfrei im Sinn einer linearen Ordnung sind und daß die Items sich somit im Raum der Dimension $k = 1$ darstellen lassen (Abbildung 16).

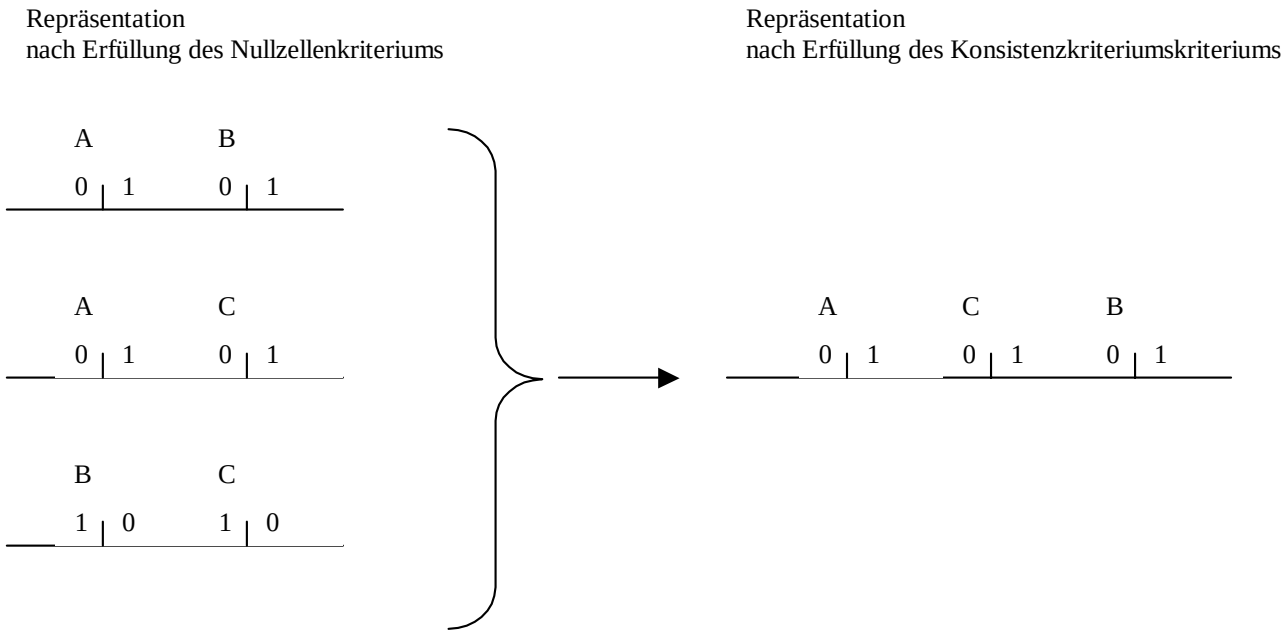


Abbildung 16: Verdichtung der Repräsentation nach Erfüllung des Nullzellenkriteriums durch die gleichzeitig Geltung des Konsistenzkriteriums - ein FPA-Lösung
(am Beispiel der Items aus Tabelle 15)

Zweidimensionale Lösungen

Genügen Items A, B, C nicht beiden Kriterien der FPA, wird die Dimension des Abbildraumes erhöht. Erhöht man die Dimensionszahl *k* auf den Wert k = 2, so wird der Darstellungsraum zur Ebene. Untersucht werden nicht mehr zwei-Weg-Kontingenzen (2-Tupel), sondern vielmehr drei-Weg-Kontingenzen (3-Tupel). Mit der Wahl der Ordnung der zugrundeliegenden Kontingenz ist die Dimension des Repräsentationsraumes eineindeutig bestimmt. Die Items werden in der Ebene durch eindimensionale Unterräume (Pseudogeraden) repräsentiert. Diese teilen den Darstellungsraum in zwei mit den Itemkategorien eindeutig bezeichneten Halbräume. Die Items einer FPA-Lösung schneiden sich gegenseitig in der Ebene genau einmal (Feger, 1994). Innerhalb einer Kontingenztafel zweiter Ordnung von drei dichotome Items sind $2^3 = 8$ Muster logisch möglich. In zwei Dimensionen können jedoch nur sieben dieser Muster repräsentiert werden (Abbildung 17).

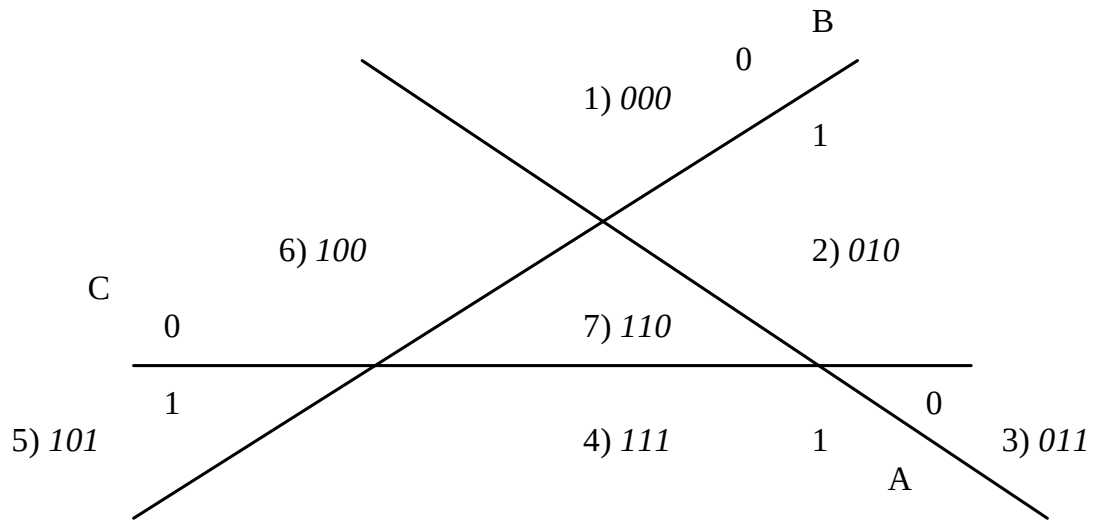


Abbildung 17: In der Ebene sind sieben Muster dreier dichotomer Items darstellbar, hier kann beispielsweise das Muster (*abc = 001*) nicht repräsentiert werden.

Analog zum eindimensionalen Fall ist das Nullzellenkriterium genau dann erfüllt, wenn ein Muster *abc* der Drei-Weg-Kontingenztafel über die Items A, B und C mit der Häufigkeit ($f_{abc} = 0$) besetzt ist. Das Konsistenzkriterium untersucht die Kategorienmuster der Nullzellen über vier Items. Es muß also ein weiteres Item D mit den Merkmalsausprägungen $d \in \{0,1\}$ vorliegen. Zur Erfüllung des Kontingenzkriteriums im zweidimensionalen Fall müssen für jeweils zwei Items bei den drei Nullzellen, an denen diese Items beteiligt sind, dieselben Kategorien auftreten. Bei den beiden anderen Items alternieren die Merkmalsausprägungen, die eine entsprechenden Nullzellen bezeichnen, genau einmal (Tabelle 16).

Brehm (1995, 2001a, 2001b) zeigt, daß Nullzellen- und Konsistenzkriterium gemeinsam notwendig und hinreichend für eine FPA-Lösung im zweidimensionalen Fall sind.

Tabelle 16: Beispiel für Nullzellen aus Dreierkontingenzen für vier Items, die das Konsistenzkriterium einer zweidimensionalen FPA-Lösung erfüllen (siehe Abbildung 18).

	Itemkategorie			
Itemtripel	*a*	*b*	*c*	*d*
A x B x C	*0*	*0*	*1*	-
A x B x D	*0*	*1*	-	*0*
A x C x D	*0*	-	*1*	*0*
B x C x D	-	*0*	*1*	*1*

Durch gegebene konsistente Nullzellen wie in Tabelle 16 ist ein Pseudogeradenarrangement eindeutig bestimmt (Abbildung 18). Die sich schneidenden Pseudogeraden umschreiben Poly-

gone, die eindeutig durch die Itemorientierungen bezeichnet werden können. Die Polygone bilden isotone Regionen bezüglich der Itemorientierungen.

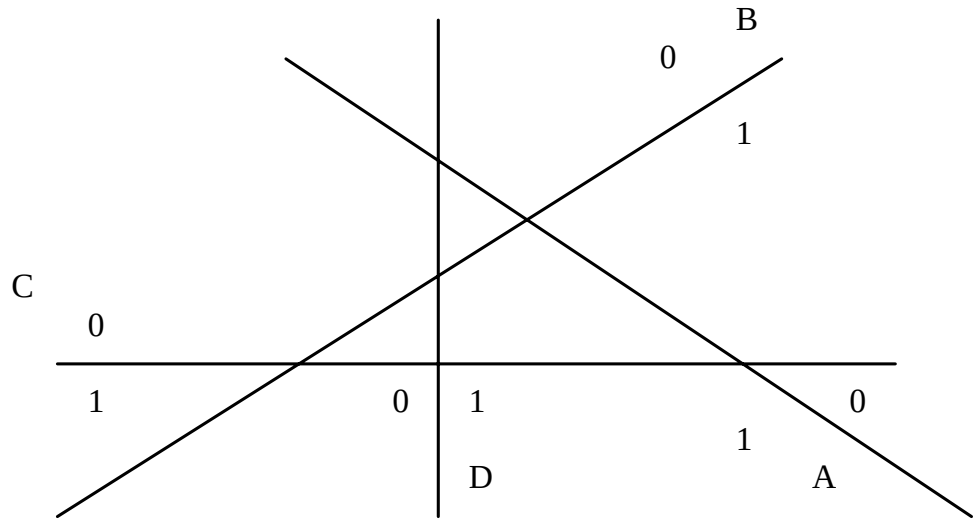


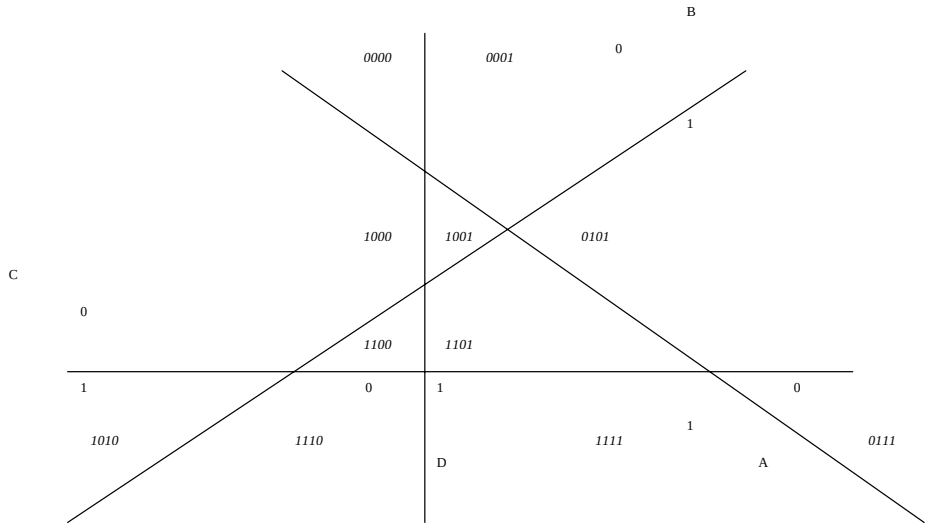
Abbildung 18: Repräsentation von vier Items in der Ebene mit Nullzellen wie in Tabelle 16.

Hassediagramm

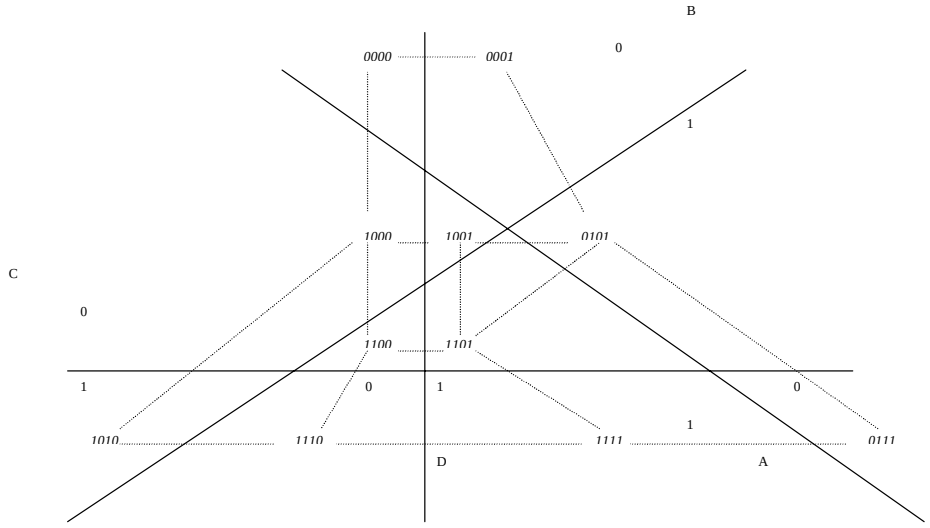
Ein Hassediagramm ist der Dualgraph des Pseudogeradenarrangements. Das bedeutet, jeder Knoten des Hassediagramms liegt in genau einer isotonen Region, in einem Polygon des Pseudogeradenarrangements und jede Kante des Hassediagramms schneidet die entsprechende Pseudogerade genau einmal. In drei Schritten läßt sich das Hassediagramm aus einem gegebenen Pseudogeradenarrangement ableiten (Abbildung 19).

Brehm (1995, 1998, 2001a, 2001b) liefert über die Form des gewählten Beweises einen Algorithmus zur Konstruktion einer Lösung für die Dimensionen $k \in \{0,1\}$. Dieser Algorithmus ist der Kern eines Computerprogramms zur Schätzung von FPA-Lösungen (Metzner, 2001).

1. Pseudogeradenarrangement mit Eintrag der durch die Regionen repräsentierten Muster *abcd*

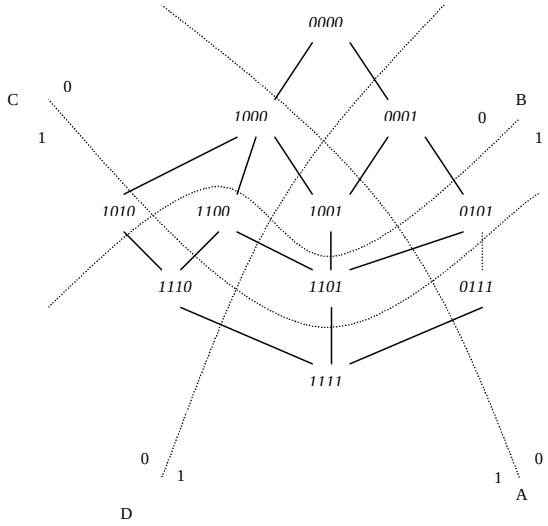


2. Konstruktion eines Dualgraphs in das Pseudogeradenarrangement



3. Vertikale Orientierung des Dualgraphs nach der Anzahl der 1 in *abcd*

4. Hassediagramm der Muster *abcd* – eine FPA-Lösung



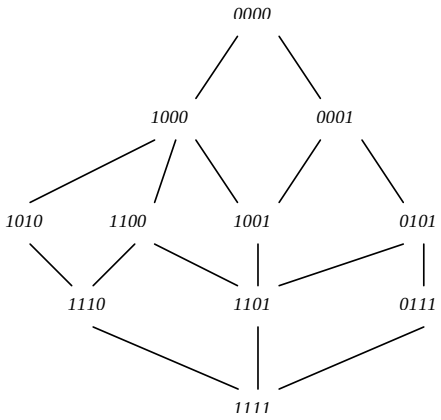


Abbildung 19: Vier Schritte zur Ermittlung des Hassediagramms aus einem Pseudogeradenarrangements

Approximative Lösungen

Das FPA-Modell setzt voraus, daß das Nullzellenkriterium für jede Kontingenz über alle Items eindeutig erfüllt ist. Bei empirischen Daten ist dies allerdings leider nicht zu erwarten.

Dabei lassen sich zwei Formen der Verletzung des Nullzellenkriteriums vorstellen:

1. alle möglichen Muster der Kontingenz sind mit Frequenzen besetzt
2. mehr als ein Muster wurde nicht beobachtet

Der letztere Verstoß ist im eindimensionale Fall jedoch immer auf eine triviale Datenlage zurückzuführen. Bei mehr als einer Nullzelle in einer Zweierkontingenz dichotomer Items ist entweder mindestens eines der beiden Items invariant oder sie sind bei unter Umständen vorzunehmender Umpolung eines der Items äquivalent. Eine Analyse macht dann jeweils keinen Sinn.Der Verstoß der ersten Form läßt keine FPA-Lösung zu. Allerdings kann man vermuten, daß die Daten eine Lösung eigentlich zulassen würden, wenn nicht ein Fehler zur irrtümlichen Besetzung der eigentlichen Nullzelle geführt hätte. Unter Zuhilfenahme von Annahmen über diesen mutmaßlichen Fehler läßt sich jedoch eine approximative FPA-Lösung ermitteln. Auf der Basis dieser Annahmen kann eine Schwellenwert definiert werden. Man deklariert die Muster, deren Frequenzen unterhalb der so festgelegten Schwelle liegen, zur Nullzelle. Durch die Wahl einer geeigneten Fehlerschwelle e kann das Nullzellenkriterium unter Umständen auf diese Weise erfüllt werden. Dies geschieht trotz der Kosten, daß alle diejenigen Gesamtmuster (über alle m Items), in denen die Muster der deklarierten Nullzellen enthalten sind, im Hassediagramm und damit in der Lösung nicht repräsentiert werden können.

Darüber hinaus kann dieses Vorgehen zu einer Transformation des Verstoßes vom Typ 1 in einen Verstoß vom Typ 2 führen. Erweist sich mehr als ein Muster als nicht besetzt oder ergibt es sich, daß mehr als ein Muster durch die Festlegung der Fehlerschwelle e zur Nullzelle deklarierte wird, so gibt es keine eindeutige Nullzelle, sondern zwei oder mehr Nullzellenkandidaten. Bei der Prüfung des zweiten, des Konsistenzkriteriums, ergeben sich daraus unter Umständen erhebliche algorithmische Probleme. Da der Prüfschritt für jeden Nullzellenkandidaten vollzogen werden muß, kann es schnell zu einem exponentiellen Anstieg der Anzahl der damit verbundenen Operationen kommen. Dies führt zu einer nicht mehr praktikablen Laufzeit eines entsprechenden Algorithmus. Im Gegensatz zu dem vorhergehenden Programm (Bäßler, 1995) tritt dieses Problem bei der aktuell vorliegenden Software in Abhängigkeit von der absoluten Anzahl der Nullzellenkandidaten erst bei einer Itemzahl von $m \approx 12$ auf.

Dabei wird im zweidimensionalen Fall noch vor der Anwendung des Konstruktionsalgorithmus nach Brehm (1995) ausgenutzt, daß die Nullzellenkandidaten, wenn sie Bestandteil einer Gesamtlösung sind, nur an konsistenten Partiallösungen über vier Items beteiligt sein dürfen (Metzner, 2001).

Im allgemeinen ist aber zu erwarten, daß keine eindeutige Lösung gefunden wird, sondern eine Menge von Lösungen. Mit der Anzahl von Nullzellenkandidaten steigt der Umfang der Lösungsmenge. Eine akzeptables Kriterium für die Auswahl unter dieser Vielfalt approximativer Lösungen besteht in der Modellgüte. Zwei Aspekte der Güte, personen- und itemorientiert, können berücksichtig werden. Der Anteil der in der Lösung repräsentierten Muster relativ zur Anzahl der Inputmuster legt das Augenmerk mehr auf den Itemaspekt. Wenn dieser relative Anteil maximal ist, wurden sämtliche empirisch vorliegenden Zusammenhänge unter den Items berücksichtigt.

Eher auf den Personenaspekt fokussiert das Verhältnis von in der FPA-Lösung repräsentierten Frequenzen zur Gesamtfallzahl (ähnlich Guttmans Reproduktionskoeffizient REP). Insbesondere wenn die Muster mit jeweils unterschiedlicher Fallzahl besetzt sind, weichen beide Gütemaße voneinander ab. Unter psychologischem Anwendungsaspekt einer FPA-Lösung als Skala erscheint ein personenorientiertes Gütemaß wie REP[4] nützlicher als ein itemorientiertes, wenn für möglichst alle Personen eine Aussage getroffen werden soll. Für Verwendungsziele mit möglichst feiner Differenzierung des Gegenstandsbereiches ist wiederum ein itemorientiertes Gütemaß zweckdienlicher.

FPA und Rangdaten

Feger und von Hecker (1999) schlagen explizit ein Modell zur zweidimensionalen Entfaltung auf der Basis der FPA vor. Dabei werden die individuellen Rangreihen von Präferenzaussagen in Paarvergleiche zerlegt. Man identifiziert ein Paar von Objekten A und B jeweils mit einer Variable $A|B$, indem der Variable genau dann der Wert $(A|B) = 1$ zugewiesen wird, wenn das Präferenzobjekt A einen höheren Rangplatz innehat als B. Vice versa gilt $(A|B) = 0$, wenn dies nicht der Fall ist. Durch diese Abbildung überführt man die Ranginformationen vermittels Paarvergleichen in eine binäre Variable und macht sie so der Analyse im Rahmen der FPA zugänglich.

[4] Der Nutzen von REP liegt in erster Linie in seiner Anschaulichkeit, seit Festinger (1954) weiß man, daß REP kein wohldefinierter Index ist.

Allerdings wird dieser Vorteil durch einen höheren Aufwand bei der Analyse erkauft, da sich mit dem Wechsel von der Untersuchungseinheit Objekt zu Paaren von Objekten bei einer ursprünglichen *n* x *m* Datenmatrix nun eine bedeutend umfangreichere *n* x $\binom{m}{2}$-Matrix ergibt.

Bei vollständigen Rangreihen scheint das Problem jedoch effizient lösbar, da mit geeigneten Algorithmen die durch die Transformation erhöhte Redundanz ausgenutzt werden kann.

In der Anwendung auf die hier vorliegende schüttere Datenlage der Wünsche führt dieses Vorgehen jedoch zur noch stärkeren Verdünnung der validen Information. Nach der Umformung bleiben nicht mehr nur (*m*-3) Einträge pro Zeile der Datenmatrix ohne Definition, sondern es ergibt sich vielmehr eine überproportionale Erhöhung auf $\binom{m}{2}-\binom{3}{2}$ undefinierte Einträge pro Zeile.

Die nicht definierten Einträge könnten für alle *m* Präferenzobjekte mit dem Wert (A | B) = 0 belegt werden. Offensichtlich würde dies jedoch dazu führen, daß sich das *000*-Muster in allen Drei-Weg-Kontingenzen als am frequenzstärksten erweist. Somit kann es keine Nullzelle bezeichnen und würde in der Lösung zwangsläufig repräsentiert werden. Diese Datenlage würde zu der paradoxen Konstellation führen, daß der Präferenzraum stets genau das Muster enthält, keines der Objekte zu präferieren. Eine substanzwissenschaftliche Interpretation würde durch einen solchen Artefakt nicht gerade erleichtert.

Diese Gefahr besteht bei der direkten Verarbeitung der Wünsche als Variablen nicht im selben Ausmaß. Deshalb werden aufgrund der vielen fehlenden Werte die direkten Kontingenzen der Wunschnennungen als Datenbasis für die Analysen benutzt.