4. Multivariate Verfahren im Dienst der Testtheorie

4.1 Verfahren zur Optimierung der Kriteriumsvorhersage: Multiple Regression und Diskriminanzanalyse

1. Wie kann man mehrere Tests zu einem Gesamtwert zusammenfassen, um eine möglichst genaue Kriteriumsvorhersage zu bekommen?
2. Wie genau fällt diese Vorhersage aus? Welche Tests können am ehesten weggelassen werden?

Vorstrukturierende Lesehilfe

Die meisten für die Praxis relevanten Kriterien, z.B. Schulerfolg oder Ausbildungserfolg, hängen von einer Vielzahl unterschiedlicher Bedingungen ab, wie Fähigkeiten, Kenntnissen, aber auch Interessen, Einstellungen und Erwartungen. Das Unterfangen, solch ein Kriterium mit einem einzelnen Test vorherzusagen, läßt von vornherein nur begrenzten Erfolg erwarten. Versucht man aber, mit unterschiedlichen Prädiktoren möglichst die gesamte Breite der Bedingungen zu erfassen, so stellt sich die Frage, wie diese unterschiedlichen Informationen relativ zueinander zu gewichten sind. Die Frage wird beantwortet, indem als Gesamtwert eine gewichtete Summe der Prädiktoren gebildet wird. Die Gewichtung wird bei einem quantitativ erfaßten Merkmal durch die multiple Regression, bei einem nicht-quantitativ erfaßbaren Kriterium (Zuordnung zu qualitativ verschiedenen Kategorien) durch die Diskriminanzanalyse bestimmt.

Als Prädiktoren können Informationen unterschiedlicher Art (Tests, Beurteilungen, Schulnoten, Alter u.a.) herangezogen werden. Wenn im folgenden von Tests als Prädiktoren die Rede ist, so ist das als Beispiel, nicht als Einschränkung zu verstehen.

4.1.1 Multiple Regression zur Maximierung der Kriteriumskorrelation

Die multiple Regression kann verwendet werden, um bei bereits feststehender Testauswahl die optimale Gewichtung zu finden, kann aber auch bei der Testauswahl selbst eingesetzt werden.

Wenn bereits feststeht, welche Tests X_1, X_2 . . X_p (z.B. die zehn Untertests des Intelligenz-Struktur-Tests von Amthauer, 1970) zur Vorhersage eines bestimmten Kriteriums Y (z.B. der Schulnote) verwendet werden sollen, so bestimmt die multiple Regressionsrechnung die Gewichte so, daß sich zwischen der gewichteten Summe der Tests und dem Kriterium eine maximale Korrelation ergibt. Man benötigt dazu die

Korrelationen aller Tests untereinander und mit dem Kriterium. Sofern nicht durch eine Standardisierung an der vorliegenden Stichprobe alle Variablen auf gleiche Mittelwerte und gleiche Streuung gebracht werden, benötigt man außerdem die Mittelwerte und Varianzen aller Variablen. Daraus lassen sich die optimalen Gewichte berechnen (das Verfahren kann hier nicht dargestellt werden. Es ist in jedem Lehrbuch über multivariate Statistik beschrieben; Literaturhinweise am Ende dieses Kapitels). Sie heißen *multiple Regressionsgewichte* (Beta-Gewichte). Hat man die Gewichte bestimmt, so wird der Kriteriumswert wie folgt geschätzt:

[4.1] $$Y^* = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \alpha$$

Y^* = geschätzter Kriteriumswert
$ß_1, ß_2, \ldots ß_p$ = multiple Regressionsgewichte (Beta-Gewichte)
α = Regressionskonstante

Die Regressionsgewichte hängen von den Kovarianzen der einzelnen Tests mit dem Kriterium, aber auch von den Kovarianzen der Tests untereinander ab. Jedes Hinzufügen weiterer Tests verändert in der Regel alle Regressionsgewichte. Die Regressionskonstante wird so bestimmt, daß Y und Y* den gleichen Mittelwert haben. Die Korrelation zwischen den Schätzwerten Y* und den tatsächlichen Kriteriumswerten Y heißt *multiple Korrelation* (R).

Berechnet man die multiplen Regressionsgewichte und die multiple Korrelation an derselben Stichprobe, so kommt es - insbesonders bei kleinen Stichproben und vielen Variablen - zu einer systematischen Überschätzung der multiplen Korrelation. Das liegt daran, daß die geschätzten Regressionsgewichte an die spezielle Stichprobe angepaßt werden, also z.B. bei 10 Tests immerhin 10 Parameter. Wenn die Stichprobe nur aus n = 10 Personen besteht, kann man mit 10 im nachhinein angepaßten Parametern in jedem Fall eine perfekte "Vorhersage" der Kriteriumswerte erzielen, selbst dann, wenn in der Grundgesamtheit keinerlei Zusammenhang bestehen sollte. Bei einer Stichprobe von n = 20 Personen und 10 Tests sind zwar mehr vorherzusagende Kriteriumswerte als anzupassende Parameter vorhanden, aber es wird immer noch zu einer deutlichen Überschätzung der multiplen Korrelation kommen. Dieser systematische Schätzfehler stellt allerdings kein grundsätzliches Problem dar, sondern kann mit Hilfe geeigneter Korrekturformeln behoben werden (Näheres dazu siehe Stevens, 1986, Kapitel 3.13).

Wenn noch nicht feststeht, welche Tests endgültig zur Kriteriumsvorhersage verwendet werden sollen, kann die Auswahl der Tests mit Hilfe einer schrittweisen multiplen Regression erfolgen. Bei der sogenannten "Vorwärts-Strategie" wird zunächst der Test gesucht, der die höchste Korrelation zum Kriterium hat. Unter den übrigen wird dann derjenige herausgesucht, der zusammen mit dem ersten die höchste multiple Korrelation (von nunmehr zwei Tests zum Kriterium) ergibt. Dieser Test kommt als zweiter in die Auswahl. Unter den verbleibenden wird wieder derjenige herausgesucht, der als dritter, zusammen mit den bereits ausgewählten Tests die höchste multiple Korrelation ergibt, usw. Das Verfahren wird abgebrochen, wenn der Zuwachs an multipler Korrelation, der sich bei Hinzunahme weiterer Tests erzielen läßt, nicht mehr als lohnend erscheint. Bei der sogenannten "Rückwärts-Strategie" berechnet man zunächst die multiple Korrelation unter Verwendung aller Tests und läßt dann schrittweise immer denjenigen Test weg, dessen Streichung zum geringsten Verlust an multipler Korrelation führt. Beide Strategien haben sich praktisch bewährt, bieten

aber mathematisch gesehen keine Garantie für eine optimale Lösung: Wählt man auf die beschriebene Art fünf von zehn Tests aus, so braucht Vorwärts- und Rückwärtsstrategie nicht zum selben Ergebnis zu führen und keine von beiden kann garantieren, daß es nicht eine noch bessere Fünfer-Kombination gibt. Rechenprogramme, wie z.B. SPSSX, bieten sowohl Vorwärts- als auch Rückwärtsstrategie, als auch gemischte Strategien an.

Wenn anhand von Stichprobendaten eine Untertest-Selektion stattgefunden hat, führt die Berechnung der multiplen Korrelation an denselben Daten der Erwartung nach zu einer systematischen Überschätzung der Güte der Vorhersage in der Population. Diese Überschätzung als Folge der Selektion geht über das hinaus, was durch die Anpassung der Regressionsgewichte bei feststehender Untertestauswahl bedingt ist. Die zu erwartende Überschätzung ist umso stärker, je kleiner die Stichprobe ist und je stärker selegiert wird. Dieses Problem ist nun nicht mehr durch Korrekturformeln lösbar, sondern erfordert eine sogenannte *"Kreuzvalidierung"*. Dafür müssen zwei unabhängige Datensätze zur Verfügung stehen (z.B. durch Zufallsaufteilung der Gesamtdaten in zwei Hälften). An dem einen Datensatz führt man die Untertestauswahl durch und bestimmt die multiplen Regressionsgewichte, an dem zweiten Datensatz wird die so gewonnene Schätzgleichung angewendet und die Korrelation zwischen Schätzwerten und Kriterium bestimmt. Diese kreuzvalidierte multiple Korrelation gibt dann eine unverzerrte Schätzung für die Güte der Vorhersage, die bei Anwendung der Schätzgleichung auf weitere Probanden erreicht werden wird. Beispiel (4.1) illustriert dieses Vorgehen.

Die in [4.1] angegebene Schätzformel geht davon aus, daß das Kriterium aufgrund einer gewichteten Summe der Testwerte vorhergesagt werden soll. Grundsätzlich ist es auch möglich, den in Formel [4.1] angegebenen Ansatz zu erweitern, indem man nichtlineare Ausdrücke (z.B. das Produkt zweier Testwerte, quadratische Funktionen der Testwerte usw.) hinzufügt, was allerdings in der Praxis kaum angewendet wird. Wenn Testwerte und Kriteriumswerte multivariat normalverteilt sind, ist die Regression des Kriteriums Y auf die Tests X_1, X_2 .. X_p linear, und der gemäß Formel [4.1] berechnete Schätzwert Y* liefert die bestmögliche Kriteriumsvorhersage, die aus den Tests zu erstellen ist.

Mit Hilfe der multiplen Regression scheint das Problem der Kriteriumsvorhersage optimal gelöst zu sein. Wenn es mit psychologischen Testbatterien gelänge, in einem Anwendungsbereich die wesentlichen Grunddimensionen individueller Unterschiede zu erfassen (z.B. für den Bereich der Schulleistungen die wesentlichen Intelligenzfaktoren, Interessens- und Einstellungsdimensionen), so könnte man die unterschiedlichen Kriterien (z.B. Noten in den einzelnen Schulfächern, Erfolg in verschiedenen Ausbildungsgängen) aus einer einheitlichen Testbatterie (allgemeiner: einem festen Satz von Prädiktoren) unter Verwendung der jeweils optimalen Gewichtung vorhersagen. Demgegenüber erscheint es zunächst überraschend, daß die multiple Regression in der Praxis so wenig genutzt wird: Kaum ein Testmanual enthält Berichte über Multiple-Regressions-Studien oder empfiehlt die Anwendung bestimmter Regressionsgewichte; lediglich einige Test-Kurzformen, die auf multiplen Regressions-Schätzungen des Gesamttestwerts beruhen, erfreuen sich größerer Verbreitung (z.B. WIP nach Dahl, 1972; WIPKI nach Baumett, 1973). Dafür dürften folgende Gründe verantwortlich sein:

Beispiel 4.1: Verwendung der multiplen Regression zur Vorhersage der Gesamttestleistung aus einer Kurzform

Der Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK) besteht aus 10 Untertests. Baumett (1973) setzte sich zum Ziel, daraus eine Kurzform zu entwikkeln, die mit dem aus dem Gesamttest errechneten IQ möglichst hoch korrelieren soll. Daraus ergeben sich die Fragen, (a) welche der 10 Untertests verwendet werden sollen und (b) wie diese Untertests gewichtet werden sollen. Faßt man das Gesamttestergebnis als Kriterium Y auf und die Untertests als die Prädiktoren X_I bis X_{10}, so läßt sich diese Fragestellung mittels multipler Regression bearbeiten. Die folgende Darstellung des Vorgehens von Baumert (1973) ist vereinfacht und bezieht nur einen Teil der dort durchgeführten Analysen mit ein.

Als Daten standen die Testprotokolle von 614 Kindern zur Verfügung, die den ganzen Test bearbeitet hatten. In Hinblick auf die geplante Kreuzvalidierung wurde zunächst die Gesamtstichprobe nach dem Zufall in zwei Teilstichproben zu je 307 Testprotokollen aufgeteilt. Es wurde an jeder der beiden Teilstichproben getrennt eine schrittweise multiple Regression nach der Vorwärtsstrategie durchgeführt. Dabei ergab sich in beiden Stichproben nach Auswahl von vier Untertests eine hohe multiple Korrelation (.94 und .95).

Die in die Auswahl aufgenommenen Tests waren aber nicht genau dieselben: Nur drei der vier Tests (AW=Allgemeines Wissen, GF=Gemeinsamkeiten finden, BO=Bilder ordnen) waren in beiden Fällen in der Auswahl, als vierter Test tauchte einmal FL(=Figurenlegen), einmal MT(=Mosaiktest) auf. Aufgrund weiterer Gesichtspunkte, u.a. aufgrund der höheren Reliabilität von MT im Vergleich zu FL, wurde dann die Kombination AW,GF,BO,MT als Kurzform festgelegt.

Danach wurde die Schätzgleichung aufgestellt und kreuzvalidiert. Die Regressionsgewichte wurden zunächst an der einen Datenhälfte bestimmt, und dann an der anderen Datenhälfte angewendet, um die kreuzvalidierte Korrelation zu berechnen. Dabei zeigt sich nur eine minimale Schrumpfung der kreuzvalidierten gegenüber der an derselben Teilstichprobe berechneten multiplen Korrelation. Diese geringe Schrumpfung ist dem großen Stichprobenumfang von 2 mal 307 Personen zu verdanken.

Nach dieser Absicherung wurde als beste Schätzung der in der Population gültigen Regressionsgleichung die Regressionsgleichung aus den Gesamtdaten (n=614) berechnet. Sie lautet:

$$IQ^* = 33 + 1.84\,AW + 1.35\,GF + 1.41\,BO + 1.66\,MT$$

Für die Leistungen in den einzelnen Untertests sind dabei die jeweils erzielten Punkte (sog.“Wertpunkte”, die aus den Antworten des Probanden gemäß Testhandanweisung altersspezifisch zu bestimmen sind) einzusetzen. Die angegebene Formel schätzt dann aus den vier Untertests den IQ, den der Proband bei Vorgabe des ganzen Tests erhalten hätte.

Als Anmerkung kann man feststellen, daß sich die Regressionsgewichte für die vier Untertests nicht sehr stark unterscheiden. Das legt die Vermutung nahe, daß eine einfache ungewichtete Addition mit anschließender Transformation auf IQ-Einheiten keine wesentlich schlechteren Ergebnisse gebracht hätte.

(1.) Ein Hinzufügen oder Wegnehmen von Tests verändert in der Regel alle Regressionsgewichte. Eine multiple Regressionsschätzung ist also nur möglich, wenn genau die angegebene Testbatterie verwendet wird.

(2.) Die multiplen Regressionsgewichte ändern sich von Population zu Population. Alles, was auf die Korrelationen der Tests untereinander und die Korrelationen der Tests mit dem Kriterium Einfluß hat (insbesondere Selektionseinflüsse aller Art), beeinflußt auch die Regressionsgewichte. Eine multiple Regressionsgleichung ist also nur dann anzuwenden, wenn die zu beratenden Probanden aus derselben Population stammen, für die die Regressionsgewichte bestimmt wurden. Das aber erscheint vielfach als fraglich, zumal wenn die Regressionsstudie zeitlich und örtlich unter recht speziellen Bedingungen durchgeführt wurde.

(3.) Wenn die Tests gleich standardisiert sind und untereinander und mit dem Kriterium positiv korrelieren, liegt die multiple Korrelation nur wenig über dem Wert, den man bei einer einfachen gleichgewichtenden Addition erreicht (Wainer, 1976). Die einfache Addition hat aber Vorteile, wenn man daran denkt, daß das Testergebnis dem Ratsuchenden vermittelt werden muß: Das Abschneiden in den einzelnen Untertests sowie ein aus den Untertests gleich gewichtend errechneter Gesamtwert ist dem Probanden leicht verständlich zu machen. Die Gewichte der multiplen Regression können für den Probanden unplausibel sein und zu einer Ablehnung des darauf gegründeten Rates führen. Diese Gründe, zusammen mit dem erheblichen Datenaufwand, der mit dem Erstellen einer multiplen Regressionsgleichung verbunden ist, dürften wohl dafür verantwortlich sein, daß die multiple Regression in der Praxis nicht stärker zum Einsatz kommt.

4.1.2 Diskriminanzanalyse zur optimalen Trennung von Kriteriumsgruppen

Wenn das Kriterium nicht quantitativ erfaßt ist (wie z.B. Ausbildungserfolg, gemessen an den Abschlußnoten), sondern zwischen qualitativ verschiedenen Gruppen unterschieden werden soll (z.B. zwischen erfolgreichen Vertretern unterschiedlicher Berufsgruppen: zwischen mehreren klinischen Gruppen, o.ä.), kann eine Diskriminanzanalyse eingesetzt werden. Aus der Testbatterie wird dann - ähnlich wie bei der multiplen Regression - eine gewichtete Summe gebildet, wobei die Gewichte so gewählt werden, daß sich die Gruppen im Summenwert möglichst gut unterscheiden: Die Mittelwertsunterschiede zwischen den Gruppen sollen möglichst groß, die Varianz innerhalb der Gruppen möglichst klein sein. Die entsprechenden Gewichtszahlen heißen *Diskriminanzgewichte,* die mit den Diskriminanzgewichten aus den Testwerten gebildete gewichtete Summenvariable heißt *Diskriminanzfunktion.* Die Werte der einzelnen Probanden auf der Diskriminanzfunktion heißen *Diskriminanzwerte.* Die Diskriminanzgewichte hängen von den Mittelwerten der Gruppen in den Tests ab, aber auch von den Varianzen und den Kovarianzen der Tests untereinander, sowie von den relativen Anteilen, mit denen Vertreter der einzelnen Gruppen in der Stichprobe repräsentiert sind (bei einer Diskriminanzanalyse zur Unterscheidung zwischen Berufsgruppen vom Anteil der einzelnen Berufe an der Gesamtstichprobe).

Bei mehr als zwei Gruppen können mehrere Diskriminanzfunktionen gebildet werden, bei k Gruppen maximal k-l. Die erste wird so gewählt, daß sie eine bestmögli-

che Trennung der Gruppen (gemessen als Varianz zwischen den Gruppenmittelwerten relativ zur Varianz innerhalb der Gruppen) ermöglicht. Die Gewichte für die zweite Diskriminanzfunktion werden so gewählt, daß der resultierende Summenwert (zweite Diskriminanzfunktion) mit dem ersten unkorreliert ist. Unter dieser Restriktion wird wieder nach einer Gewichtung gesucht, die die Gruppen bestmöglich trennt. Die dritte Diskriminanzfunktion muß mit jeder der ersten beiden unkorreliert sein, usw. (zur rechnerischen Durchführung sowie zur Erweiterung des Ansatzes auf nichtlineare Funktionen sei auf die am Ende des Kapitels angeführten Lehrbücher verwiesen).

Kennt man ‘die Testwerte eines Probanden, so können daraus seine Werte in den Diskriminanzfunktionen berechnet werden. Wenn bestimmte Voraussetzungen erfüllt sind (die Testwerte sind in jeder Kriteriumsgruppe multivariat normalverteilt; die Kovarianzmatrizen sind gleich; die Grundraten, d. h. die Anteile, die die einzelnen Kriteriumsgruppen an der Gesamtpopulation ausmachen, sind bekannt), kann man daraus die bedingten Wahrscheinlichkeiten für die Zugehörigkeit zu den einzelnen Kriteriumsgruppen berechnen. Unter schwächeren Voraussetzungen kann man globale Ähnlichkeitsmaße verwenden, die die Nähe des Probanden zu den einzelnen Kriteriumsgruppen ausdrücken. Darauf aufbauend können verschiedene diagnostische Entscheidungsstrategien gewählt werden, nach denen die Probanden den Kriteriumsklassen zugeordnet werden: Man kann die Entscheidungsregel so wählen, daß einfach die Gesamtzahl richtig Klassifizierter maximiert wird, oder man kann verschiedene Arten von Fehlklassifikationen unterschiedlich stark gewichten und ein daraus abgeleitetes Nützlichkeitsmaß maximieren (Näheres dazu findet man bei Kallus & Janke, 1988).

Ähnlich wie bei der multiplen Regression kann man auch bei der Diskriminanzanalyse versuchen, durch schrittweises Hinzufügen von Tests eine möglichst sparsame Testbatterie zusammenzustellen, die eine möglichst gute Trennung der Kriteriumsgruppen erlaubt. Statt schrittweise hinzuzufügen (Vorwärtsselektion), kann man auch von einer gegebenen Testbatterie ausgehend schrittweise jeweils denjenigen Test weglassen, der am wenigsten zur Unterscheidung der Gruppen beiträgt (Rückwärtsselektion).

Bezüglich der Verallgemeinerbarkeit der Ergebnisse aus einer Diskriminanzanalyse sind dieselben Einschränkungen zu machen, wie bei einer multiplen Regression:

(1.) Ein Hinzufügen oder Wegnehmen von Tests verändert in der Regel alle Diskriminanzgewichte.

(2.) Ein Hinzunehmen oder Wegnehmen von Gruppen oder Verschiebungen in den relativen Anteilen der Gruppen an der Gesamtpopulation verändert in der Regel die Diskriminanzgewichte.

(3.) Wenn die Berechnung der Diskriminanzfunktionen und die Bestimmung der Vorhersagegenauigkeit (Prozent richtig klassifizierter Probanden) an derselben Stichprobe erfolgen, kommt es zu einer Überschätzung der Güte der Vorhersage. Das gilt in verstärktem Maß, wenn anhand derselben Daten eine Variablenselektion (s. oben) stattgefunden hat. Die Überprüfung der Vorhersagegenauigkeit sollte deshalb an einem neuen, unabhängigen Datenmaterial erfolgen (Kreuzvalidierung).

Eines der größten Forschungsprojekte im Bereich der angewandten Diagnostik, bei dem die Diskriminanzanalyse eingesetzt wird, dürfte die Entwicklung der maschinellen Auswertung der Testbogen in der Berufsberatung sein (Engelbrecht 1975; 1978). Bei der Bundesanstalt für Arbeit liegen aufgrund langjähriger Datensammlung inzwi-

schen für eine Vielzahl von Berufen Testwerte ehemaliger Ratsuchender vor, die inzwischen ihren Beruf erfolgreich ausüben. Aufgrund einer diskriminanzanalytischen Auswertung, die mit EDV.-Einsatz realisiert wird, ist es möglich, für jeden neuen Ratsuchenden die globale Ähnlichkeit zu den Vertretern der einzelnen Berufsgruppen als bedingte Wahrscheinlichkeit der Berufsgruppenzugehörigkeit anzugeben. Im Beratungsgespräch stellt sich allerdings das Problem, wie das Testergebnis an den Ratsuchenden zu vermitteln ist, so daß das Zustandekommen einer Empfehlung für den Probanden nachvollziehbar ist. Dazu sind Werte auf Diskriminanzfunktionen wenig geeignet. Deshalb wird zusätzlich für jede Berufsgruppe angegeben, in welchen Einzeltests (Leistungstests, Interessentests) der Proband relativ zu dieser Berufsgruppe sehr hohe oder sehr niedrige Werte aufweist, also vom durchschnittlichen Vertreter dieser Berufsgruppe stark abweicht. Sowohl Abweichungen nach oben (hohe Fähigkeiten oder Interessen in Bereichen, die für den Beruf nicht typisch sind) als auch nach unten (geringe Ausprägung von berufstypischen Interessen und Fähigkeiten) können auf Probleme hinweisen und Gegenstand des weiteren Beratungsgesprächs sein.

Zusammenfassung

Die Frage, wie mehrere Prädiktoren zu gewichten sind, um ein Kriterium bestmöglich vorherzusagen, wird bei quantitativ erfaßbaren Kriterien durch die multiple Regression, bei qualitativ et-faßbaren Kriterien durch die Diskriminanzanalyse beantwortet. Bei der multiplen Regression wird aus der gewichteten Summe der Prädiktoren ein geschätzter Kriteriumswert berechnet; Maß für die Güte der Vorhersage ist die multiple Korrelation.

Bei der Diskriminanzanalyse werden aus den Prädiktoren zunächst Werte auf Diskriminanzfunktionen berechnet; aus diesen wieder können bedingte Wahrscheinlichkeiten für die Zugehörigkeit zu den einzelnen Kriteriumsgruppen (oder andere Maße, die die Nähe des Probanden zu den einzelnen Kriteriumsgruppen ausdrucken) berechnet werden. Maß für die Güte der Vorhersage ist der Anteil richtig klassifizierter Probanden oder ein darauf aufbauendes Nützlichkeitsmaß. Die Ergebnisse sowohl einer multiplen Regression als auch einer Diskriminanzanalyse sind für die Personenpopulation und die spezielle Prädiktorenauswahl spezifisch und können in der Regel nicht darüber hinaus verallgemeinert werden.

Einführende Literatur:

Lehrbücher über multivariate statistische Verfahren:

Fahrmeir, L. & Harnerle, A. (Hrsg.) (1984). *Multivariate statistische Verfuhren.* Berlin: De Gruyter.

Hartung, J. & Elpelt, B. (1984). *Multivariate Statistik.* München: Oldenbourg.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale: Lawrence Erlbaum.

Weiterführende Literatur:

Zur multiplen Regression:

Schubö, W., Haagen, K. & Oberhofer, W.(1983). Regressions- und kanonische Analyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S.207-292). Göttingen: Hogrefe.

Zur Diskriminanzanalyse:

Krauth, J. (1983). Diskriminanzanalyse. In J.Bredenkamp & H.Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S.293-350). Göttingen: Hogrefe.

Zu Klassifikations- und Entscheidungsstrategien in der Diagnostik:

Cronbach,L.J. & Gleser,G.C.(1965). *Psychological tests and personne1 decisions.* Urbana: University of Illinois Press.

Kallus, K.W. & Janke, W. (1988). Klassenzuordnung. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik* (S. 131- 145). München: Psychologie Verlags Union.

Noack, H. und Petermann, F. (1988). Entscheidungstheorie. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik* (S.241-253). München: Psychologie Verlags Union.

4.2 Faktorenanalyse zur Untersuchung der Konstruktvalidität

1) Von welchen Grundannahmen geht die klassische Faktorenanalyse aus?
2) Warum kommt es zwischen faktorenanalytischen Theorien zu keiner Entscheidung? Warum werden die Ergebnisse von Faktorenanalysen nicht mehr als funktional erklärende Theorien betrachtet?
3) Was unterscheidet konfirmatorische Faktoranalysen von exploratorischen? Welche Anwendungsmöglichkeiten bieten sich in der Testtheorie?

Vorstrukturierende Lesehilfe

Die Faktorenanalyse ist ein multivariates Verfahren, das mit der Geschichte der Theorienbildung im Bereich der Intelligenz- und Persönlichkeitsforschung und mit der Entwicklung psychologischer Tests besonders eng verbunden ist. Die Ideen der grossen Faktorentheoretiker wie Spearman, Guilford, Thurstone haben die Konzepte unserer heutigen Tests in Inhalt und Aufbau nachhaltig beeinflußt. Die heute im Gebrauch befindlichen psychometrischen Tests gehen überwiegend auf diese Vorbilder zurück.

Der ursprünglich hohe Anspruch, mit der Faktorenanalyse eine funktionale Analyse leisten zu können, z. B. die Grundfähigkeiten zu entdecken und zu messen, aus denen die menschlichen Intelligenzleistungen erklärbar würden, wird heute allerdings nicht mehr erhoben. Vielmehr betrachtet man heute die Faktorenanalyse als eine Methode, die geeignet ist, Korrelationsmuster überschaubarer zu machen und Interpretationsmöglichkeiten aufzuzeigen. Auch mit diesem reduzierten Anspruch kann sie zur Beantwortung der Frage nach der Validität eines Tests wertvolle Beiträge leisten. Im folgenden wird zunächst der Grundansatz der Faktorenanalyse dargestellt, dann werden die Hauptkritikpunkte wiedergegeben, die zu der erwähnten Rücknahme des Anspruchs geführt haben. Schließlich soll noch die konfirmatorische Faktorenanalyse als Weiterentwicklung der klassischen Faktorenanalyse bezüglich ihrer Anwendung auf testpsychologische Fragestellungen diskutiert werden.

4.2.1 Grundannahmen der Faktorenanalyse

Die folgende Darstellung orientiert sich am Modell mehrerer gemeinsamer Faktoren als dem allgemeinsten Ansatz. Andere Modelle lassen sich als Spezialfalle auffassen, die aus diesem Ansatz durch Zusatzannahmen hervorgehen (z. B. das Ein-Faktor-Modell durch die Annahme, es gebe nur einen gemeinsamen Faktor; das Hauptkomponenten-Modell durch die Annahme, die gesamte Testvarianz gehe auf gemeinsame Faktoren zurück). Eine umfassende Darstellung, die auch historische Aspekte miteinbezieht, gibt Pawlik (1971).

4.2.1.1 Die Grundgleichungen

Als Beispiel wollen wir annehmen, wir hätten die Korrelationen zwischen einer Vielzahl von Leistungstests (Intelligenztests, Schulleistungstests usw.) vorliegen. Es liegt

nahe, anzunehmen, daß diese Korrelationen dadurch zustande kommen, daß die Tests sich in ihren Anforderungen überschneiden, d.h. z.T. dieselben Fähigkeiten beanspruchen. Ziel der Faktorenanalyse ist es nun, solche mehreren Tests gemeinsamen Fähigkeiten zu definieren und ihr relatives Gewicht für die einzelnen Tests zu bestimmen. Allgemeiner gesprochen, besteht das Ziel der Faktorenanalyse darin, Korrelationen zwischen Variablen (hier: Leistungstests) auf gemeinsame Faktoren (= Dimensionen individueller Unterschiede, hier: Fähigkeiten) zurückzuführen und damit eine sparsame Interpretation der Korrelationen anzubieten.

Gemäß den Annahmen der Faktorenanalyse sind also für jede Testleistung mehrere Fähigkeiten (= Faktoren) erforderlich (z.B. zum Lösen eingekleideter Rechenaufgaben: Textverständnis, schlußfolgerndes Denken, Rechenfertigkeit), die sich mit unterschiedlichen Gewichten auf die Testleistung auswirken. Fähigkeiten (Faktoren), die von mehreren Tests (einer in einer Faktorenanalyse gemeinsam analysierten Testgruppe) beansprucht werden, heißen gemeinsame Faktoren, solche die nur in einem einzigen Test vorkommen, spezifische Faktoren. Darüber hinaus enthält jeder Test Meßfehler.

Gleichung [4.2] gibt an, wie die Testleistung einer Person in einem Test gemäß den Grundannahmen der Faktorenanalyse zustande kommt:

[4.2] $$z_{iv} = a_{i1} f_{1v} + a_{i2} f_{2v} + ...+ u_{iv}$$

Z_{iv} = Testwert der Person v im Test i, ausgedruckt in z-Werten, d.h. standardisiert auf Mittelwert 0 und Varianz 1.

a_{i1} = Gewicht, mit dem Faktor 1 die Testleistung im Test i bestimmt = Faktorladung des Tests i in Faktor 1.

a_{i2} = Faktorladung des Tests i in Faktor 2. Weitere Faktorladungen sind analog definiert.

f_{1v} = Faktorwert der Person v im ersten Faktor (individuelle Fähigkeitsausprägung in Faktor 1).

f_{2v} = Faktorwert der Person v im zweiten Faktor.

Weitere Faktorwerte für Person v sind analog definiert.

Die Faktorwerte sind für jeden Faktor auf den Mittelwert 0 und die Varianz 1 standardisiert.

U_{iv} = Durch die gemeinsamen Faktoren nicht erklärter Restanteil (englisch: uniqueness). Er enthält Einflüsse spezifischer Faktoren und Meßfehler und wird als von den gemeinsamen Faktoren unabhängig vorausgesetzt.

Aus dieser Grundgleichung ergibt sich, wie die Korrelation zwischen zwei Tests i und j zustande kommt: Die Leistung der Person v im Test j kann analog zerlegt werden (Gleichung [4.2a]):

[4.2a] $$z_{jv} = a_{j1} f_{1v} + a_{j2} f_{2v} + ...+ u_{jv}.$$

Betrachtet man nun die Kovarianz (= Korrelation, weil die Tests z-standardisiert sind) zwischen Test i und j, so sieht man, daß sie einerseits von den Gewichten (= Faktorladungen) abhängt, die die gemeinsamen Fähigkeiten für die beiden Tests haben, andererseits von den Korrelationen der Fähigkeiten (= Faktorwerte) untereinander.

In einer obliquen Faktorenanalyse werden die Fähigkeiten als beliebig korreliert gedacht, in der orthogonalen Faktoranalyse werden sie als unabhängig vorausgesetzt

bzw. definiert. Die Annahme unabhängiger Faktoren führt zu einigen mathematischen Vereinfachungen.

So ergibt sich in der orthogonalen Faktorenanalyse die Korrelation zwischen zwei Tests i und j allein aus den Ladungen in den gemeinsamen Faktoren, wie in [4.3] angegeben:

[4.3] $$r_{ij} = a_{i1}\, a_{j1} + a_{i2}\, a_{j2} + \ldots$$

Darüber hinaus läßt sich in der orthogonalen Faktorenanalyse die Ladung zugleich als die Korrelation des Tests mit den Faktorwerten in diesem Faktor interpretieren. Weiterhin läßt sich bei orthogonalen Faktoren die beobachtbare Testvarianz in additive Anteile aufspalten, die auf die einzelnen Faktoren zurückgehen (Gleichung [4.4]):

[4.4] $$\sigma^2(z_i) = a^2_{i1} + a^2_{i2} + \ldots. + \sigma^2(u_i)$$

Das Quadrat der Ladung (a^2) gibt somit den Anteil an der Testvarianz an, der auf den entsprechenden Faktor (auf individuelle Unterschiede in der entsprechenden Fähigkeit) zurückzuführen ist. Die Summe der Ladungsquadrate für einen Test heißt Kommunalität und gibt an, zu welchem Anteil die Varianz dieses Tests durch die gemeinsamen Faktoren "aufgeklärt" wird. Sie wird gewöhnlich mit h^2 bezeichnet. Zur globalen Charakterisierung, inwieweit in einer Faktoranalyse die Varianz aller Variablen durch die gemeinsamen Faktoren aufgeklärt wird, kann man die durchschnittliche Kommunalität angeben. Zur Charakterisierung dessen, wieviel jeder einzelne Faktor zur aufgeklärten Varianz aller Variablen beiträgt, kann man die Summe der Ladungsquadrate dieses Faktors (summiert über die Variablen) zur Summe der Kommunalitäten in Beziehung setzen.

4.2.1.2 Geometrische Darstellung, Rotationsproblem, Kommunalitätenproblem

Die rechnerische Aufgabe der Faktorenanalyse besteht darin, aus den Korrelationen aller Tests untereinander die Faktorladungen zu bestimmen. Dabei ist man bestrebt, mit möglichst wenigen Faktoren auszukommen und dabei die Faktorladungen so zu bestimmen, daß man aus ihnen die beobachteten Korrelationen möglichst genau reproduzieren kann. Hat man es z.B. mit 20 Tests zu tun, deren Korrelationen aus 3 Faktoren erklärt werden sollen, so müssen sich die 190 beobachteten Korrelationen aus nur 20 x 3 = 60 Faktorladungen jeweils gemäß Gleichung [4.3] ergeben.

Wie man rechnerisch vorgeht, um dies in bestmöglicher Näherung zu erreichen und wie man entscheidet, ob weitere Faktoren notwendig sind, kann hier nicht dargestellt werden. Statt dessen sollen hypothetische Ausgangsdaten und das Ergebnis einer orthogonalen Faktorenanalyse die Grundgleichungen an einem Zahlenbeispiel illustrieren. An diesem Beispiel soll dann auch die geometrische Darstellung und das Rotationsproblem erläutert werden.

Tabelle 4.1a enthält die Korrelationen zwischen 6 Tests (fingierte Daten), Tabelle 4.1b die Faktorladungen in zwei gemeinsamen Faktoren und Tabelle 4.lc die aus den Faktorladungen gemäß Gleichung (4.3) rekonstruierten Korrelationen. Die Abweichungen zwischen den Korrelationen in den Daten und den rekonstruierten Korrelationen (= Residuen) sind hier so gering, daß man zwei Faktoren als zur Erklärung der Korrelationen ausreichend ansehen wird (Tabelle 4.ld).

Tabelle 4.1a: Korrelationen zwischen 6 Tests (fingierte Daten)

	Test					
	1	2	3	4	5	6
1	-	.41	.43	.30	.33	.19
2	-	-	.70	.50	.56	.30
3	-	-	-	.54	.63	.36
4	-	-	-	-	.68	.39
5	-	-	-	-	-	.44
6	-	-	-	-	-	-

Tabelle 4.1b: Faktorladungen der 6 Tests in den 2 Faktoren und Kommunalitäten der Tests

		Faktoren		Kommunalität
		I	II	h^2
Tests	1	.46	.20	.21
	2	.74	.32	.55
	3	.83	.36	.69
	4	.35	.70	.61
	5	.44	.74	.74
	6	.25	.44	.26

Tabelle 4.1c: Aus den in Tabelle 4.1b angegebenen Faktorladungen gemäß Gleichung [4.3] rekonstruierte Korrelationen zwischen den 6 Tests

	Test					
	1	2	3	4	5	6
1	-	.40	.45	.30	.35	.20
2	-	-	.72	.48	.56	.32
3	-	-	-	.54	.63	.36
4	-	-	-	-	.67	.39
5	-	-	-	-	-	.43
6	-	-	-	-	-	-

Tabelle 4.1d: Residuen
Differenzen zwischen den Ausgangskorrelationen in Tabelle 4.la und den aus den Faktorladungen rekonstruierten Korrelationen in Tabelle 4.lc

	Test					
	1	2	3	4	5	6
1	-	.Ol	-.02	.00	-.02	-.0l
2	-	-	-.02	+.02	.00	-.02
3	-	-	-	.00	.00	.00
4	-	-	-		+.01	.00
5	-	-	-	-	-	.01
6	-	-	-	-	-	-

Geometrische Darstellung

Bei nur zwei Faktoren lassen sich die Ergebnisse einer Faktorenanalyse leicht graphisch veranschaulichen. In Abbildung 4.1 sind die Faktoren I und II als Achsen eines Koordinatensystems dargestellt und die Tests sind gemäß ihren Ladungen eingetragen.

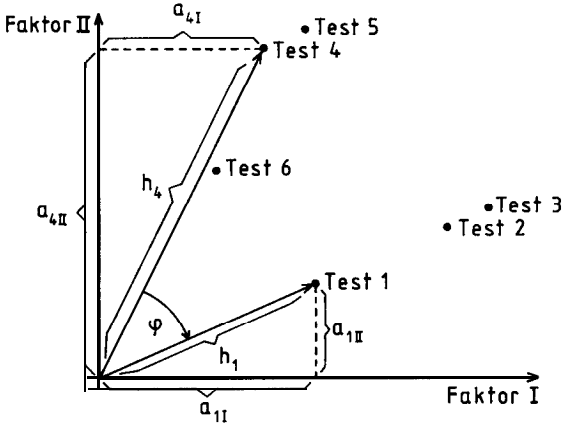


Abbildung 4.1: Darstellung von 6 Tests im zweidimensionalen Faktorraum. Ladungen gemäß Tabelle 4.1b

Es läßt sich zeigen (pythagoräischer Lehrsatz), daß die Wurzel aus der Kommunalität der Länge des Vektors eines Tests (graphisch als Pfeil vom Nullpunkt des Koordinatensystems zum Test hin dargestellt) entspricht. Weiter ergibt sich die Korrelation zweier Tests aus der Länge ihrer Vektoren und dem eingeschlossenen Winkel

(ableitbar aus dem Cosinus-Satz der Geometrie), wie in Gleichung [4.5] angegeben.

[4.5] $$r_{ij} = a_{i1}\, a_{j1} + a_{i2}\, a_{j2} = h_i\, h_j \cos \varphi$$

Diese Beziehungen gelten bei mehr als zwei Faktoren im mehrdimensionalen Raum entsprechend.

Das Rotationsproblem

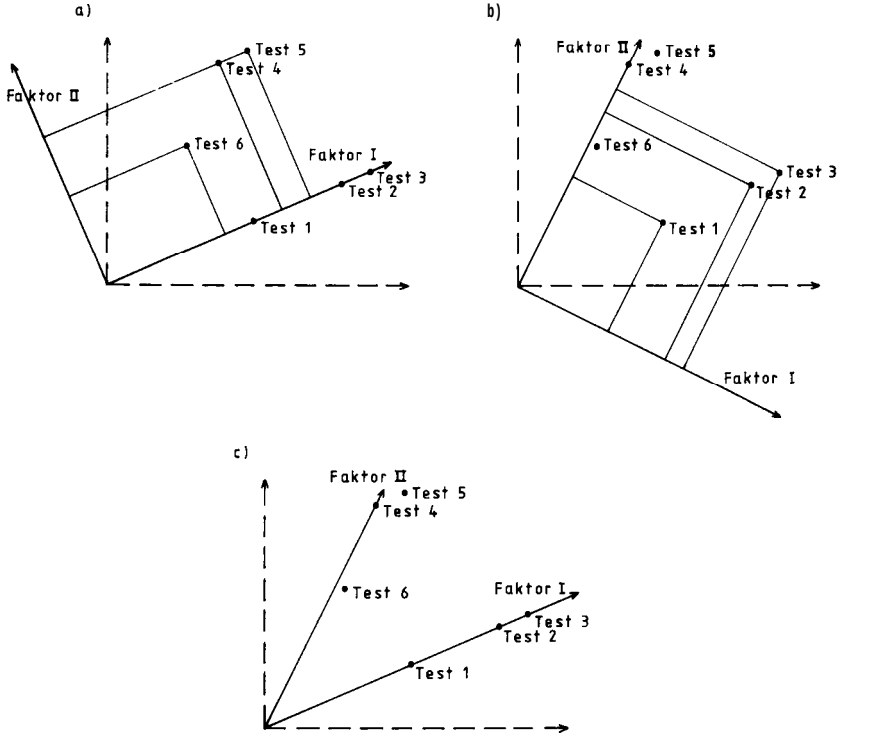
Wenn man eine Faktorlösung gefunden hat, so kann man dazu beliebig viele weitere konstruieren, die zu genau denselben rekonstruierten Korrelationen führen, also ganz genauso gut auf die Daten passen: Wie in Gleichung [4.5] angegeben, hängen die Korrelationen zwischen den Tests nur von der Länge ihrer Vektoren und den eingeschlossenen Winkeln ab. Wenn nun das Koordinatensystem beliebig gedreht wird, so ändert sich an diesen Winkeln und Längen nichts (d.h. die reproduzierten Korrelationen bleiben gleich), wohl aber an den Koordinaten der Tests (den Faktorladungen), die nun am neuen Koordinatensystem abzulesen sind. Indem man das Koordinatensystem beliebig dreht, kann man somit beliebig viele mathematisch gleichwertige Lösungen für die Faktorladungen produzieren. Diese können inhaltlich recht unterschiedliche Deutungen nahelegen: Bei der in Abbildung 4.1 bzw. Tabelle 4.1b dargestellten Lösung kommen die Korrelationen zwischen den Tests durch zwei Faktoren zustande, auf denen alle Tests Ladungen haben. Durch eine Drehung des Koordinatensystems um ca. 24 Grad nach links erhält man die in Abbildung 4.2a angegebene Lösung mit einem Generalfaktor und einem Gruppenfaktor. Die Tests 1,2,3 laden nur im Generalfaktor während die Tests 4, 5, 6 einen zusätzlichen Faktor gemeinsam haben. Durch eine Drehung des Koordinatensystems um ca. 26 Grad nach rechts entsteht das in Abbildung 4.2b dargestellte Bild: Nun scheinen die Tests 4,5 und 6 nahezu nur einen Generalfaktor zu erfassen, während die Tests 1,2,3 einen zusätzlichen Faktor gemeinsam haben. Das Problem, zwischen solchen mathematisch gleichwertigen, aber inhaltlich verschiedenen Lösungen zu entscheiden, ist als Rotationsproblem der Faktorenanalyse bekannt. Über mathematische Versuche zu definieren, was eine "einfache" und damit gut interpretierbare Lösung ist, und über die rechnerische Durchführung der Rotation bei mehr als zwei Faktoren soll hier nicht berichtet werden. Das Thema ist in den am Ende des Kapitels genannten Lehrbüchern behandelt.

Schiefwinkelige Rotation und Faktoren zweiter Ordnung

Bei den in Abbildung 4.1 und 4.2a,b dargestellten Lösungen stehen die als Koordinaten gezeichneten zwei Faktoren jeweils im rechten Winkel aufeinander. Dem entspricht die Annahme, daß die beiden Faktoren unkorreliert (= orthogonal) sind. Faktor I könnte z.B. Rechenfähigkeit, Faktor II Wortflüssigkeit sein. Die Orthogonalität bedeutet dann, daß die beiden Fähigkeiten in der Personenpopulation nicht korrelieren. Wenn es gelänge, zwei Tests zu konstruieren, von denen der eine ausschließlich Rechenfähigkeit mißt (also auf der Ordinate liegt) und der andere ausschließlich Wortflüssigkeit mißt (also auf der Abszisse liegt), so wäre der Winkel zwischen den Vektoren dieser beiden Tests 90 Grad und gemäß Formel [4.5] müßten auch die beobachteten Testwerte zu Null korrelieren.

Geht man von der Forderung unkorrelierter Faktoren ab, so kann man weitere Lösungen produzieren, indem man die Koordination verschiedene Winkel bilden läßt

Abbildung 4.2: Rotation. Orthogonale Lösungen (a) und (b) und eine nicht-orthogonale Lösung (c)



Gegenüber der in Abbildung 4.1 dargestellten Lösung ist das Koordinatensystem in (a) um 24 Grad nach links, in (b) um 26 Grad nach rechts gedreht. In (c) bilden die Koordinaten einen Winkel von 40 Grad und entsprechen einer Faktorkorrelation von 0.77.

und jeweils die Koordinaten der Punkte in diesem schiefwinkeligen Koordinatensystem berechnet. Eine solche Lösung ist in Abbildung 4.2c angegeben. Die Koordinaten bilden hier einen Winkel von 41 Grad, was einer Korrelation der Faktoren von cos (41°) = 0.76 entspricht. Sowohl die Tests 1, 2, 3 als auch 4,5,6 liegen fast genau auf einer Koordinatenachse. Die Tests messen also jeweils (fast) nur eine der beiden Fähigkeiten, die beiden Fähigkeiten sind aber miteinander korreliert. Auch dies wäre eine inhaltlich plausible Deutung des Korrelationsmusters.

Wenn man es nicht nur mit zwei, sondern mit mehreren Faktoren zu tun hat und sich für eine Lösung mit korrelierenden Faktoren entschieden hat, so kann man weiter fragen, wie denn die Korrelationen zwischen den Faktoren zustande kommen. Nimmt man die Korrelationen zwischen den Faktoren als "Daten" und unterzieht sie ihrerseits einer Faktoranalyse, so nennt man die daraus resultierenden Faktoren "Faktoren zweiter Ordnung", und es entsteht ein hierarchisches Modell: Die Testleistungen werden aus gewichteten Summen von Fähigkeiten erklärt, die Korrelationen zwischen den Fähigkeiten aus Faktoren zweiter Ordnung (z.B. Fähigkeiten wie Wort-

schatz, Worteinfallgeschwindigkeit, Erkennen verbaler Beziehungen aus einem allgemeineren verbalen Faktor und spezifischeren Komponenten). Auch Faktorenanalysen zweiter Ordnung lassen ihrerseits wieder Spielraum für Rotation.

Aufgrund der dargestellten Vielfalt äquivalenter Modelle wird man wohl kaum den Anspruch erheben, mittels Faktorenanalyse eine bestimmte Lösung als die richtige ausweisen zu können. So z.B. lassen die positiven Korrelationen zwischen Intelligenztests, die man in aller Regel findet (selbst wenn sich der Testautor um Unabhängigkeit der Einzeltests bemüht hat, z.B. Intelligenz-Struktur-Test von Amthauer, 1953, 1970; Leistungsprüfsystem von Horn, 1962), verschiedene inhaltliche Deutungen zu:

(a) Es gibt einen Faktor der allgemeinen Intelligenz, der in alle Testleistungen mehr oder weniger stark eingeht.

(b) Es gibt keinen allgemeinen Faktor, sondern Intelligenz besteht aus mehreren unabhängigen Einzelfähigkeiten. Es ist aber nicht möglich, faktoriell reine Tests zu konstruieren, sondern jeder Test beansprucht mehrere Fähigkeiten.

(c) Es gibt keinen allgemeinen Faktor, sondern mehrere Einzelfähigkeiten. Die Tests (oder Testgruppen) erfassen jeweils nur eine dieser Fähigkeiten, die Fähigkeiten sind miteinander korreliert.

Jede dieser inhaltlichen Deutungen läßt sich in ein faktorenanalytisches Modell umsetzen, jedes ist mit den Daten vereinbar. Diese Unentscheidbarkeit, die als Rotationsproblem schon im Modell-Ansatz enthalten ist, ist ein Grund dafür, daß man von einer Faktoranalyse keine abschließenden Aussagen über die einer Testleistung zugrundeliegenden Funktionen und Prozesse erwarten kann.

Praktisch bevorzugt werden möglichst einfache, gut interpretierbare Lösungen. Bei den meisten Arbeiten, die Faktorenanalysen anwenden, werden orthogonale Lösungen gewählt, und es wird zu einem mathematisch definierten Einfachheitskriterium rotiert. Das bekannteste ist das Varimax-Kriterium: Pro Faktor wird die Varianz der quadrierten Ladungen berechnet; die Summe dieser Varianzen ist das Kriterium, das maximiert wird. Die Varianz der quadrierten Ladungen wird groß, wenn sowohl Null-Ladungen als auch dem Betrag nach hohe Ladungen vorhanden sind. Ein in diesem Sinn prägnantes Ladungsmuster läßt sich im allgemeinen leichter inhaltlich deuten als ein Ladungsmuster mit vielen mittleren Ladungen auf allen Variablen.

Das Problem der Kommunalitätenschätzung

Die rechnerische Durchführung der Faktorenanalyse setzt nicht nur die Kenntnis der Korrelationen, sondern auch die Kenntnis der Kommunalitäten (zum Begriff der Kommunalität siehe Abschnitt 4.2.1.1) voraus. Diese aber ergeben sich erst aus den zunächst noch unbekannten Faktorladungen. Es gibt zwar verschiedene Möglichkeiten, die Kommunalitäten schon vorher zu schätzen, doch kann das Ergebnis einer Faktorenanalyse nicht nur bezüglich der Höhe der Ladungen, sondern auch bezüglich der Zahl der benötigten Faktoren von der Wahl des Kommunalitäten-Schätzverfahrens abhängen. Dieses Problem, das ebenfalls schon im mathematischen Ansatz der Faktorenanalyse steckt, trägt zur weiteren Uneindeutigkeit faktorenanalytischer Lösungen bei.

Die Hauptkomponenten-Analyse (englisch: principal component analysis) unterscheidet sich vom klassischen Ansatz der Faktorenanalyse, wie er in Formel [4.2] angegeben ist, dadurch, daß keine Uniqueness vorgesehen ist und alle Kommunalitäten

gleich Eins sind. Da jeder Test Meßfehler enthält, ist dieser Ansatz mit einer funktionalen Interpretation von vornherein nicht vereinbar. Ziel ist hier lediglich, die Vielzahl von Testvariablen auf einige wenige Faktoren zu reduzieren, die die in den Testvariablen enthaltene Information möglichst gut repräsentieren. Näheres zur Beziehung zwischen klassischer Faktoranalyse und Hauptkomponentenanalyse findet man bei Snook & Garsuch (1989) und Velicer & Jackson (1990).

4.2.2 Haupteinwände gegen die Faktorenanalyse als erklärende Theorie

Bereits im vorangehenden Kapitel wurde deutlich, daß die Ergebnisse der Faktorenanalyse mathematisch nicht eindeutig sind, sondern dem Forscher einen erheblichen Interpretationsspielraum lassen. Das betrifft sowohl die Anzahl der Faktoren, die je nach dem gewählten Kommunalitäten-Schätzverfahren und je nach Abbruchkriterium für die Faktorenextraktion unterschiedlich ausfallen kann, als auch die Festlegung der Rotation. Allein diese Unbestimmtheit mag die Faktorenanalyse als "weiche" Methode erscheinen lassen, wenig geeignet für eine stringente Überprüfung von Theorien.

Die Haupteinwände dagegen, daß man mittels Faktorenanalysen die Grundfähigkeiten entdecken und das Zustandekommen von Testleistungen erklären, also die Grundgleichung allgemeinpsychologisch auffassen und funktional interpretieren könnte, sind jedoch nicht nur in dieser mathematischen Unterbestimmtheit zu sehen, sondern vor allem in einer Reihe von Kritikpunkten, die Ende der Sechzigerjahre von verschiedener Seite (Fischer, 1968; 1974; Kallina, 1967; Kalveram, 1965; 1970a, b; Merz & Kalveram, 1965) vorgetragen wurden: Die prinzipielle Unüberprüfbarkeit des Ansatzes, die Populationsabhängigkeit der Ergebnisse, die Entstehung von Artefakten durch simultane Überlagerung oder Selektionseffekte.

Zur Unüberprüfbarkeit des Ansatzes

In der Grundgleichung (Gleichung [4.2]) wird angenommen, daß die Testleistung aufgrund einer gewichteten Summe von Fähigkeiten zustande kommt, wobei
(a) die Gewichtung für alle Personen gleich ist und
(b) die Fähigkeiten einander beliebig kompensieren können.

Als Ausgangsdaten für eine Faktorenanalyse stehen aber nur die Korrelationen zwischen den Tests zur Verfügung. Egal wie diese zustande gekommen sind - ob gemäß den in der Grundgleichung ausgedrückten Annahmen oder ganz anders -,jede Korrelationsmatrix kann faktorisiert werden, und es ist dem Ergebnis der Faktorenanalyse nicht anzusehen, ob die Annahmen der Grundgleichung zutreffen oder nicht.

Die Populationsabhängigkeit des Ergebnisses

Korrelationen beschreiben Merkmalszusammenhänge in Populationen. Sie können in unterschiedlichen Populationen (definiert nach Alter, Geschlecht, Schulbildung usw.) unterschiedlich ausfallen. Dementsprechend wird auch das Ergebnis einer Faktorenanalyse derselben Tests, sowohl was die Anzahl der Faktoren als auch was die Ladungen anbelangt, von Population zu Population unterschiedlich sein. Andererseits gehört aber jede einzelne Person mehreren Populationen zugleich an: eine 13jährige

Oberschülerin z.B. der Population der 13jährigen, der Population der Mädchen, der Population der Oberschülerinnen. Interpretiert man das Ergebnis einer Faktorenanalyse auf individueller Ebene als Aussage darüber, wieviele und welche Fähigkeiten eine Person für die Lösung des Tests einsetzt, so gerät man sehr bald in Widersprüche. Derselben Person wäre je nachdem, welcher Population man sie gerade zurechnet, eine andere Fähigkeitsstruktur zuzuschreiben.

Artefakte durch simultane Überlagerung und Selektionseffekte

Selbst wenn die Testleistung in einer Population bei jedem Einzelindividuum so zustande kommt, wie in der Grundgleichung angenommen, ist nicht gewährleistet, daß man als Ergebnis der Faktorenanalyse die richtige Zahl von Faktoren und richtigen Ladungen erhält. Das haben Merz & Kalveram (1965) am Beispiel der Differenzierungshypothese der Intelligenz eindrucksvoll gezeigt:

Gemäß der Differenzierungshypothese ändert sich die Intelligenz in der Entwicklung vom älteren Kind **zum** Erwachsenen vor allem qualitativ durch Differenzierung. Dementsprechend wird mit dem Alter ein Absinken der Korrelationen zwischen den Tests, eine Zunahme der Zahl unabhängiger Fähigkeiten und eine Abnahme der Bedeutung des Generalfaktors erwartet. Merz & Kalveram (1965) konnten zeigen, daß dasselbe Ergebnis zu erwarten ist, wenn die Intelligenzstruktur, was Anzahl und Gewicht der zur Lösung eingesetzten Faktoren anbelangt, gleichbleibt, auf den einzelnen Altersstufen aber unterschiedlich starke individuelle Differenzen im allgemeinen Entwicklungsstand bestehen. Besonders auf den unteren Altersstufen, wo **das** Entwicklungstempo noch rasch ist, werden manche Kinder gegenüber den Gleichaltrigen einen alle Fähigkeiten mehr oder weniger stark betreffenden Entwicklungsvorsprung, andere einen Entwicklungsrückstand haben. Wenn alle Testleistungen eines Probanden zugleich (simultan) in dieselbe Richtung beeinflußt werden, steigen die Korrelationen zwischen den Tests. Merz & Kalveram (1965) sprechen von "simultaner Überlagerung" der Korrelationsstruktur durch Kovarianz, die auf Unterschiede im Entwicklungsstand zurückgeht. Im Erwachsenenalter da**gegen,** wenn die Entwicklung praktisch abgeschlossen ist, spielen diese Entwicklungsunterschiede keine Rolle mehr, und die Korrelationen fallen niedriger aus. Als Ergebnis von orthogonalen Faktoranalysen erhält man bei den Jüngeren höhere Kommunalitäten, einen stärkeren Generalfaktor, geringere Ladungen in den weiteren Faktoren und -je nach Abbruchkriterium - eine geringere Gesamtzahl von Faktoren. Insgesamt entsteht also ein Bild, das voll den Erwartungen aufgrund der Differenzierungshypothese gleicht. Weitere Beispiele für Artefakte durch simultane Überlagerung sind in derselben Arbeit und bei Kalveram (1965) zu finden.

Eine weitere Quelle von Artefakten, die die Korrelationen zwischen den Tests so verändern können, daß selbst dann, wenn die Grundgleichung als Annahme über den Lösungsprozeß bei jedem einzelnen Probanden zutrifft, die Faktorenanalyse als Ergebnis weder die richtige Faktorenzahl noch die richtigen Ladungen liefert, sind Selektionseffekte. Kalveram (1969) demonstriert an einem Beispiel mit Intelligenztestdaten, daß schon eine mäßige Selektion nach der Punktsumme (Weglassen der Probanden mit den höchsten und niedrigsten Werten für den Gesamt-IQ) deutliche Effekte auf die lnterkorrelationen der Tests hat: Extreme Summenwerte kommen zustande, wenn Probanden in allen Tests gut oder in allen Tests schlecht abgeschnitten haben. Ein Weglassen dieser Fälle muß zu einer Reduktion der Korrelationen führen. In einem so selegierten Datenmaterial sind dann weder die gemeinsamen Faktoren voneinander unabhängig, wie das in der orthogonalen Faktorenanalyse vorausgesetzt wird, noch **auch** die spezifischen von dem gemeinsamen (eine Voraus-

setzung, die auch in der obliquen Faktorenanalyse gemacht wird), und das Ergebnis der Faktorenanalyse wird in die Irre führen.

Dem kann man nun entgegenhalten, daß eine explizite Selektion an den Daten ja in der Regel nicht erfolgt. Andererseits kann auch in "natürlichen" Populationen, wie z.B. Schülern einer bestimmten Schulart mit mittlerem Anforderungsniveau, faktisch eine Selektion nach dem Durchschnittsniveau eines Schülers (Mittel über seine Fähigkeiten) stattgefunden hat, indem Extremfälle positiver wie negativer Art die Schule häufiger verlassen haben. In diesem Falle gelten die obigen Argumente entsprechend.

Aufgrund der genannten Argumente wurde der Anspruch aufgegeben, die Ergebnisse von Faktorenanalysen könnten als für jeden einzelnen Probanden gültige Aussage über das Zustandekommen von Testleistungen interpretiert werden.

Ungeachtet dessen bleibt das Problem bestehen, daß man bei der Beurteilung der Validität eines Tests weitgehend auf Korrelationen angewiesen ist und größere Mengen von Korrelationen konsistent interpretieren möchte. Da die Faktorenanalyse solche Interpretationsmöglichkeiten aufzeigen kann, wurde sie trotz des Vorwurfs der Nicht-Falsifizierbarkeit als Theorie, als heuristisches Instrument auch in Zeiten stärkster Kritik unvermindert zum Einsatz gebracht. Sie wird dann als eine datenexplorierende Technik aufgefaßt, die mit dem Korrelationsmuster vereinbare Deutungen anbietet, wobei man freilich zunächst nicht weiß, ob eine davon richtig ist und welche. Die Entscheidung darüber, welche Hypothesen weiter verfolgt werden sollen, ist dann nur aufgrund zusätzlicher Information aus inhaltlichen Gründen möglich.

Eine noch entschiedenere Abkehr vom ursprünglichen Anspruch vollzieht man, wenn man die Faktoranalyse als eine Methode auffaßt, die für eine bestimmte Population Dimensionen individueller Unterschiede beschreibt. Da Beschreibungsdimensionen nur nach Gesichtspunkten der Zweckmäßigkeit, Ökonomie und Ergiebigkeit zu beurteilen sind, nicht aber nach "wahr" oder "falsch", stellt sich die Frage nach der Falsifizierbarkeit erst gar nicht. Daß bei Populationen, die sich in Art und Ausmaß individueller Unterschiede unterscheiden, jeweils andere Beschreibungsdimensionen in den Vordergrund treten, erscheint dann als selbstverständlich und sachlich begründet und nicht als Mangel der Methode. Auch das Problem der Artefakte, z.B. durch simultane Überlagerung oder Selektionseffekte, stellt sich erst, wenn man über die Definition von Beschreibungsdimensionen hinausgeht und nach den Gründen fragt, warum z.B. in der einen Altersklasse ein Generalfaktor den größten Teil der Varianz abschöpft, in der anderen nicht. Die Beschreibung des Sachverhalts läßt mehrere Deutungen (Differenzierungshypothese, simultane Überlagerung) zu, zwischen denen erst durch zusätzliches Wissen (hier über Entwicklungskurven und das Ausmaß individueller Unterschiede auf den einzelnen Altersstufen) zu entscheiden ist.

Eine typische Anwendung dieser Art, bei der die Faktorenanalyse von vornherein nur mit dem Ziel eingesetzt wird, eine Vielzahl von Variablen auf eine oder einige wenige Beschreibungsdimensionen zu reduzieren, die die wesentliche Information enthalten, liegt z.B. vor, wenn aus einer Vielzahl von Intelligenztests ein Gesamtwert gebildet werden soll, der dann in der weiteren Auswertung anstelle der vielen Einzeltests die Intelligenz repräsentieren soll. Hier liegt es nahe, aus einer Faktorenanalyse nach der Hauptkomponentenmethode die erste Hauptkomponente (den Faktor, der die meiste Varianz abschöpft) zu verwenden. Eine weitere Anwendung, die mit einer Deutung der Faktorenanalyse als Methode zur bloß deskriptiven Dimensionsanalyse auskommt, ist die Faktorenanalyse von Testitems mit dem Ziel, ltemgruppen zu Skalen zusammenzustellen, die für diese Population (!) eine hohe innere Konsistenz der

Skalen erwarten lassen. Auch bei völliger Rücknahme des Anspruchs auf ein bloßes Datenreduktionsverfahren lassen sich also sinnvolle Anwendungen für die Faktorenanalyse finden.

Eine wesentliche Weiterentwicklung der klassischen Faktorenanalyse, die inzwischen oft auch als "exploratorische" Faktorenanalyse bezeichnet wird, stellt die konfirmatorische Faktorenanalyse dar. Sie geht von inhaltlichen Hypothesen aus und macht falsifizierbare Aussagen über die Struktur der Korrelations- oder Kovarianzmatrix. Einige Einsatzmöglichkeiten im Rahmen der Testtheorie sollen im folgenden an Beispielen dargestellt werden. Dabei wird sich freilich auch zeigen, daß auch eine falsifizierbare Theorie, wenn sie auf die Daten paßt, deshalb noch lange nicht die einzige mögliche Erklärung sein braucht. Wenn die Vorhersagen der Theorie sehr strikt sind, wird es allerdings sehr schwer werden, plausible Alternativerklärungen für dieselben Daten zu finden.

4.2.3 Einsatzmöglichkeiten und Grenzen der konfirmatorischen Faktorenanalyse

In der klassischen Faktorenanalyse braucht der Forscher kein Vorwissen über die Anzahl der Faktoren oder über das zu erwartende Ladungsmuster zu besitzen. Es werden so lange Faktoren extrahiert, bis die Korrelationsmatrix aus den Faktorladungen hinreichend genau reproduzierbar ist. Es wird dann durch Rotation (nach mathematischen oder inhaltlichen Kriterien) eine gut interpretierbare Lösung gesucht. In der konfirmatorischen Faktorenanalyse, die - ausgehend von den Arbeiten von Jöreskog (1967; 1969) -vor allem in den siebziger Jahren entwickelt wurde, muß der Forscher schon vor Eintritt in das Verfahren eine Hypothese über die Zahl der Faktoren und das Ladungsmuster haben. Bei der Hypothese über das Ladungsmuster handelt es sich in der Regel um Annahmen darüber, daß einzelne Tests auf bestimmten Faktoren nicht laden (vorgeschriebene Null-Ladungen), oder um Annahmen über Gleichheit bestimmter Ladungen (Gleichheits-Restriktionen). Darüber hinaus können über die Korrelationen der Faktorwerte einschränkende Annahmen gemacht werden (z.B. daß alle oder auch nur bestimmte Faktoren unkorreliert sind) und bezüglich der Residuen (testspezifische Faktoren und Meßfehler) Festlegungen getroffen werden (z.B. Gleichheit der Residualvarianzen bei bestimmten Tests). Insgesamt müssen die gesetzten Restriktionen ausreichen, um die Lösung mathematisch eindeutig zu machen, insbesondere also auch die Rotation festzulegen.

Ausgangsdaten können Korrelations- oder Kovarianzmatrizen sein. Die Parameter des Modells (Faktorladungen, Korrelationen der Faktoren, Residualvarianzen) werden dann so geschätzt, daß sie (a) den durch die Hypothese gesetzten Restriktionen genügen und (b) die empirischen Korrelationen (oder Kovarianzen) zwischen den Tests so gut, wie unter den gesetzten Restriktionen möglich, reproduzieren. Anhand der erreichten Anpassung (Übereinstimmung der aus den Ladungen reproduzierten Korrelationen mit den aus den Daten errechneten) wird beurteilt, ob die Hypothese mit den Daten vereinbar ist oder nicht.

Zur Schätzung der Parameter und zur Beurteilung der Anpassung stehen eine Reihe theoretisch unterschiedlich begründeter Verfahren zur Verfügung (eine neuere Übersicht findet man bei Anderson & Gerbing, 1988). Das am stärksten verbreitete Computer-Programm dürfte nach wie vor das Programm LISREL (zur Zeit neueste Version: LISREL 7, Jöreskog & Sörbom, 1989) sein, an zweiter Stelle dürfte das Pro-

gramm EQS (Bentler, 1985) stehen. Beide Programme umfassen einen weiten Bereich von linearen Strukturgleichungsmodellen und enthalten die konfirmatorische Faktorenanalyse als Spezialfall.

Im folgenden soll an vier unterschiedlichen Fragestellungen gezeigt werden, wie Problemstellungen aus der Testtheorie mit Hilfe konfirmatorischer Faktorenanalysen bearbeitet werden können.

Beispiel 1 : Überprüfung der Parallelität von Tests

Zwei oder mehr Tests sind parallel im Sinne der klassischen Testtheorie, wenn sie dieselben wahren Werte und gleiche Meßfehlervarianzen haben. Daraus folgt u.a., daß ihre Varianzen gleich sind, daß die Kovarianzen der Parallelformen untereinander gleich sind und daß die Kovarianzen der Parallelformen zu einem beliebigen Außenkriterium gleich sind. Diese Struktur der Kovarianzmatrix kann in einer konfirmatorischen Faktorenanalyse überprüft werden. Die Parallelitätshypothese wird dabei ausgedrückt, indem für die Tests festgelegt wird, (a) daß sie auf einem gemeinsamen Faktor laden, (b) daß die Ladungen auf diesem Faktor gleich sind und (c) daß die Residualvarianzen gleich sind. Abbildung 4.3 zeigt ein hypothetisches Beispiel:

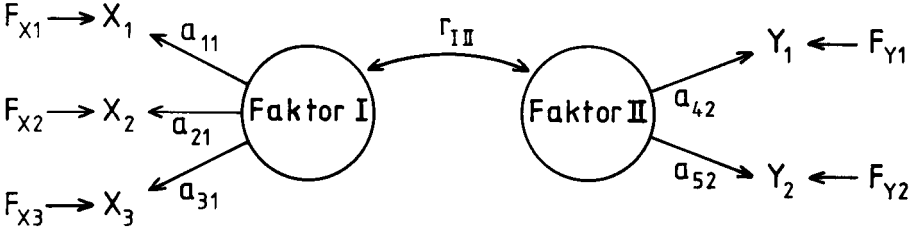


Abbildung 4.3: Konfirmatorische Faktoranalyse zur Prüfung der Parallelität der Tests X_I, X_2, X_3 und Y_1, Y_2. Parameterspezifikation und Restriktionen s. Tabelle 4.2

Die Tests X_1, X_2, X_3 sollen drei Parallelformen eines Wortschatztests sein, die Tests Y_1 und Y_2 zwei Parallelformen eines Rechentests. Der Anteil, den Faktor I an der Varianz eines Wortschatztests ausmacht, entspricht der wahren Varianz; die Restvarianz ist die Fehlervarianz. Entsprechendes gilt für Faktor II und die beiden Rechentests. Die Korrelation der Faktoren I und II ist die Korrelation der wahren Werte von Wortschatztest und Rechentest. Will man das Modell prüfen, so hat man die Parametermatrizen zu spezifizieren und die Restriktionen zu setzen, wie in Tabelle 4.2 angegeben.

Wenn das Modell nicht paßt, kann eine schwächere Hypothese geprüft werden: Beispielsweise könnten die Tests X_1, X_2, X_3 dasselbe messen, aber mit unterschiedlicher Reliabilität. Will man unterschiedliche wahre Varianzen zulassen, so ist die Gleichheitsrestriktion für Faktor I im Ladungsmuster aufzuheben; will man zusätzlich unterschiedliche Fehlervarianzen zulassen, so entfällt die entsprechende Restriktion bezüglich der Residualvarianzen.

Beispiele, in denen verschieden streng gefaßte Modelle an realen Daten (Intelligenz- und Schulleistungstest) vergleichend geprüft wurden, findet man in der inzwischen als klassisch anzusehenden Arbeit von Jöreskog (1978).

Tabelle 4.2: Parameterspezifikation zu dem in Abbildung 4.3 dargestellten Modell einer konfirmatorischen Faktoranalyse zur Prüfung der Parallelität der Tests X_1, X_2, X_3 und Y_1, Y_2.

	Ladungsmatrix Faktoren	
	I	II
Tests		
X_1	a_{11}	0
X_2	a_{21}	0
X_3	a_{31}	0
Y_1	0	a_{42}
Y_2	0	a_{52}

	Kovarianzmatrix der Faktoren	
	I	II
II	1	
II	$r_{1.11}$	1

Gleichheitsrestriktionen für die
(a) Ladungen:
$a_{11} = a_{21} = a_{31}$
$a_{42} = a_{52}$

(b) Fehlervarianzen:
$\sigma^2(F_{X1}) = \sigma^2(F_{X2}) = \sigma^2(F_{X3})$
$\sigma^2(F_{Y1}) = \sigma^2(F_{Y2})$

Beispiel 2: Überprüfung von Hypothesen über die Gleichheit von Ladungsmustern in verschiedenen Populationen
Wenn man die Anwendungen der klassischen Faktorenanalyse überblickt, so findet man ganz überwiegend orthogonale Faktorenlösungen. Werden analoge Analysen für unterschiedliche Personenstichproben durchgeführt, so wird in der Regel jede Faktorenanalyse für sich gerechnet und die Ergebnisse hinterher vergleichend diskutiert. Methoden zur Ähnlichkeitsrotation wurden zwar vorgeschlagen (z.B. Fischer & Roppert, 1964), aber kaum angewendet.

Andererseits ist es alles andere als plausibel anzunehmen, die Faktoren seien in den unterschiedlichsten Populationen immer wieder unkorreliert. Wenn die Faktoren in den Populationen unterschiedlich korreliert sind, wird eine von der Methode her gesetzte Orthogonalitätsrestriktion zu von Population zu Population unterschiedlichen Faktorlösungen führen - auch dann, wenn die Tests in allen Populationen dasselbe messen.

Ein Vorteil des Programms LISREL (Jöreskog & Sörbom, 1989) besteht darin, daß es die Möglichkeit bietet, an mehrere Datensätze simultan eine konfirmatorische Faktorenanalyse anzupassen. Dabei kann festgelegt werden, daß bestimmte Parameter (z.B. Faktorladungen) für alle Datensätze gleich sein sollen, während andere (z.B. Varianzen und Kovarianzen der Faktoren) von Stichprobe zu Stichprobe variieren können. Man kann also z.B. der Reihe nach folgende, zunehmend restriktive Modelle testen:

1. Dasselbe Ladungsmuster (Zuordnung der Tests zu den Faktoren und entsprechend vorgeschriebene Null-Ladungen) paßt in allen Populationen. Die Ladungen können aber in den einzelnen Populationen unterschiedlich hoch sein und die Faktoren können in den einzelnen Populationen unterschiedlich korreliert sein.
2. Ladungsmuster und Ladungen müssen in allen Populationen übereinstimmen, Faktorvarianzen und Faktorkorrelationen können aber von Population zu Population unterschiedlich sein.
3. Die Lösungen stimmen völlig überein, d.h. die Korrelationsmatrix ist in allen Populationen gleich.

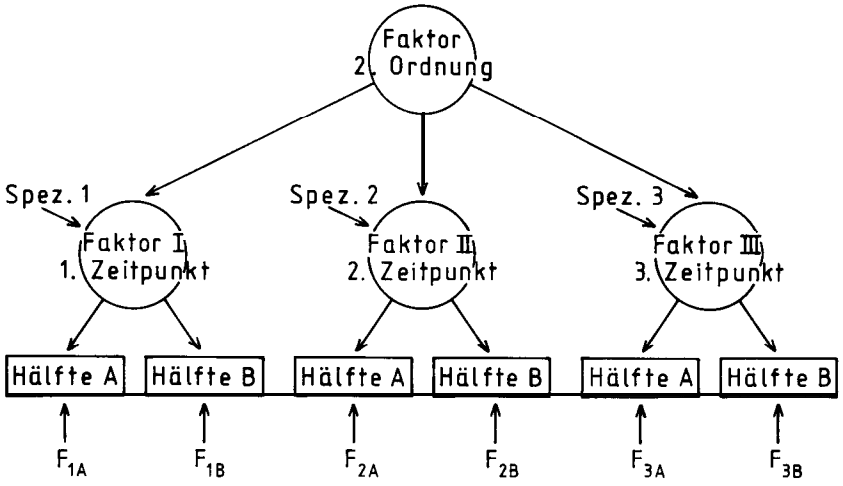
Ein Beispiel für eine solche schrittweise Anpassung einer gemeinsamen faktoranalytischen Lösung für 5 Datensätze findet man bei Schmidt (1983). Er untersuchte an 5 Altersgruppen die faktorielle Struktur eines Fragebogens über Arbeitsorientierungen. Angenommen wurde schließlich ein Modell mit 4 Faktoren und derselben Zuordnung der Items zu den Faktoren (gleiches Ladungsmuster), wobei aber die Höhe der Ladungen in den einzelnen Altersgruppen unterschiedlich war. Ein komplexeres Beispiel, bei dem 6 Modellvarianten verglichen wurden, findet man bei Jöreskog & Sörbom (1985, Kapitel V 3). Stelzl (1987) illustriert an einem Beispiel mit hypothetischen Daten die Vorteile einer simultan konfirmatorischen Faktorenanalyse über alle Datensätze gegenüber getrennten klassischen Faktorenanalysen mit Rotation zur orthogonalen oder auch nicht-orthogonalen Einfachstruktur.

Beispiel 3: Die Zerlegung der wahren Varianz in Konsistenz und Spezifität nach Steyer (1987)

Steyer (1987) und Majcen, Steyer & Schwenkmezger (1988) schlagen eine konfirmatorische Faktorenanalyse zweiter Ordnung vor, um "Spezifität" und "Konsistenz" als Anteile der wahren Varianz eines Tests zu unterscheiden. Dazu wird der Test in zwei Hälften geteilt und beide Hälften den Probanden zu mehreren "Meßgelegenheiten", z.B. Zeitpunkten im Abstand von jeweils mehreren Wochen, vorgelegt.

Die Kovarianzmatrix der Daten wird dann nach dem in Abbildung 4.4 dargestellten Modell analysiert. Den beiden Testhälften zum selben Zeitpunkt wird jeweils ein gemeinsamer Faktor erster Ordnung unterstellt. Der Varianzanteil dieses Faktors an der Testvarianz ist die wahre Varianz des Tests. Den Kovarianzen zwischen den Zeitpunkten wird dann ein Generalfaktor als Faktor zweiter Ordnung unterstellt. Den Varianzanteil eines Tests, der durch diesen Generalfaktor erklärt wird, nennt Steyer "Konsistenz", den Anteil an der wahren Varianz, der nicht durch den Generalfaktor erklärt wird, "Spezifität". Den Generalfaktor interpretiert er als "Personfaktor" oder "Trait", die Spezifität als "Situations-" oder "Person-Situations-Interaktionsvarianz". Zum selben Meßzeitpunkt können sich die einzelnen Personen in unterschiedlichen Situationen befinden, z.B. ausgeschlafen oder verkatert sein (Situationsvarianz) und auf diese Situationen personspezifisch reagieren (Person-Situations-Interaktionsvarianz).

Abbildung 4.4: Konfirmatorische Faktorenanalyse zweiter Ordnung zur Unterscheidung von Konsistenz und Spezifität im Sinn von Steyer (1987)



Ein Test, bestehend aus den Hälften A und B, wird zu drei Zeitpunkten vorgegeben. Seine wahre Varianz besteht aus einem Anteil, der auf den Generalfaktor zweiter Ordnung zurückgeht, und einem Anteil, der für den jeweiligen Zeitpunkt spezifisch ist.

Beispiel 4: Die Multitrait-Multimethod-Matrix

Campbell & Fiske (1959) zeigten, wie man anhand einer sogenannten "Multitrait-Multimethod-Matrix" konvergente und diskriminante Validität psychologischer Messungen überprüfen kann: Dazu müssen mehrere Eigenschaften (= traits), z.B. Popularität und Expansivität eines Schülers, mit mehreren Methoden (Selbstauskunft, Rating durch andere, Verhalten in einer Gruppensituation, Rollenspiel) erfaßt worden sein und die Korrelationen aller Messungen (hier: 2 Eigenschaften und 4 Methoden = 8 Messungen) vorliegen. Diese Korrelationsmatrix heißt Multitrait-Multimethod-Matrix und soll eine bestimmte Struktur aufweisen:

Auch wenn man bei psychologischen Maßen immer davon ausgehen muß, daß nur ein Teil der Varianz auf die zu messende Eigenschaft zurückgeht und ein Teil methodenspezifisch ist, so wird man von einem guten Maß doch verlangen, daß der methodenspezifische Anteil gering ist. Dementsprechend sollten die Korrelationen zwischen Maßen, die dieselbe Eigenschaft mit unterschiedlichen Methoden erfassen, deutlich höher ausfallen (konvergente Validität) als die Korrelationen zwischen Maßen, die dieselbe Methode verwenden, aber unterschiedliche Eigenschaften erfassen (niedrige Korrelationen unterschiedlicher Eigenschaften = diskriminante Validität).

Bei mehr als zwei Eigenschaften kann man auch das Muster der Korrelationen der Eigenschaften untereinander betrachten. Wenn die Eigenschaften mit derselben Methode erfaßt wurden (alle durch Selbstauskunft oder alle durch Fremdbeurteilung), sollte sich jeweils dasselbe Korrelationsmuster ergeben, egal um welche Methode es sich handelt.

Die von Campbell & Fiske (1959) angestellten Überlegungen, auf die eine längere Diskussion folgte (dargestellt bei Schmitt et al., 1977) lassen sich gut in eine kon-

firmatorische Faktorenanalyse übertragen. Jeder Eigenschaft und jeder Methode entspricht ein Faktor, wobei jedes Maß hier auf einem Eigenschafts- und einem Methodenfaktor lädt. Diese Ladungen sollen bestimmt werden, alle anderen sind Null. Die Eigenschaftsfaktoren können untereinander korreliert sein, sollen aber von den Methodenfaktoren unabhängig sein. Beispiele solcher Anwendungen findet man u.a. bei Kenny (1976) Schmitt, Coyle & Saari (1977) Schwarzer (1983), Ostendorf et al. (1986). Eines der von Schwarzer (1983) vorgestellten Beispiele wird im folgenden Abschnitt dargestellt und diskutiert.

Grenzen der konfirmatorischen Faktorenanalyse:

Der Haupteinwand gegen die klassische Faktorenanalyse als Mittel zur Prüfung von Theorien über das Zustandekommen von Testleistungen liegt in der Nicht-Falsifizierbarkeit des theoretischen Ansatzes: Das Rechenverfahren führt immer zu einer Faktorlösung - auch dann, wenn die Testleistungen in Wirklichkeit ganz anders zustande kommen.

Die konfirmatorische Faktorenanalyse bringt in diesem Punkt eine Verbesserung: Wenn die Hypothese über das Ladungsmuster sehr restriktiv ist, so muß die Korrelationsmatrix eine ziemlich genau festgelegte Struktur haben, um mit der Hypothese vereinbar zu sein. Hat sie diese Struktur nicht, so wird das Modell verworfen.

Trotzdem bleiben grundlegende Probleme bestehen. Wenn das Modell verworfen wird, weil es nicht auf die Daten paßt, so kann das daran liegen, daß grundlegende Annahmen falsch sind. Es kann aber auch daran liegen, daß die Korrelationen z.B. durch Selektionseffekte (vgl. Kapitel 4.2.2) verzerrt sind. Wenn es z.B. durch eine Selektion nach der Summe aller Faktorwerte zu negativen Korrelationen auch der Residuen kommt, so wird ein ansonsten richtiges Modell mit unabhängigen Residuen verworfen (korrelierende Residuen sind bei konfirmatorischen Faktorenanalysen zwar nicht grundsätzlich unzulässig, führen aber sehr bald zu trivialen oder auch zu nicht identifizierbaren, d.i. nicht schätzbaren Modellen).

Wenn ein Modell nicht paßt, wird es meist nicht pauschal verworfen, sondern man sucht nach Korrekturmöglichkeiten und modifiziert einige weniger wichtige Annahmen, bis der Modelltest keine signifikanten Abweichungen mehr ausweist. Dabei läuft man Gefahr, das Modell im nachhinein an zufällige Eigenschaften der Stichprobe anzupassen. Eine Signifikanzprüfung des modifizierten Modells erfordert in jedem Fall einen neuen, unabhängigen Datensatz. Wenn auf der anderen Seite ein Modell gut an die Daten angepaßt ist, so schließt das nicht aus, daß andere, ebenfalls plausible Modelle ebensogut auf dieselben Daten passen. Auch zu einem hoch restriktiven Modell wie einer Multitrait-Multimethod-Lösung kann es Alternativen geben. Das soll durch eine Reanalyse eines von Schwarzer (1983) vorgestellten Beispiels demonstriert werden. Schwarzer (1983) verwendete die Daten von Winne & Marx (1981, zit. nach Schwarzer, 1983), um eine Multitrait-Multimethod-Analyse zu demonstrieren. Winne & Marx legten 181 Vpn drei Fragebogen zum Selbstbild vor. Jeder der drei Fragebogen (A = Sears' Self-Concept Inventory, B = eigener Fragebogen mit Selbsteinstufungen auf Rating-Skalen, C = eigener Fragebogen mit Vergleichen zu anderen Studenten) enthält eine Skala zu denselben drei Aspekten des Selbstbildes: "Academic", "Physical" und "Social Self-Concept", bezeichnet als Traits 1,2,3. Tabelle 4.3 gibt die Korrelationen zwischen den 3 mal 3 Fragebogenskalen an.

Tabelle 4.3: Korrelationen zwischen 3 mal 3 Fragebogenskalen zum Selbstbild. Daten von Winne & Marx (1981) zitiert nach Schwarzer (1983).

	Al	A2	A3	B1	B2	B3	C1	C2	C3
A1	1.00								
A2	.31	1							
A3	.48	.54	1						
B1	.49	-.03	-.03	1.					
B2	.22	.77	.33	.14	1.				
B3	.11	.35	.37	.06	.54	1.			
C1	.61	-.01	.10	.60	-.02	-.05	1.		
C2	.23	.13	.42	-.02	.70	.39	.14	1.	
C3	.22	.44	.55	-.07	.40	.48	.08	.56	1.

A = Sears' Self-Concept Inventory, B = Selbsteinstufung auf Rating-Skalen, C = Selbsteinstufung im Vergleich zu anderen Studenten.
1 = Academic 2 = Physical 3 = Social Self-Concept

Schwarzer paßte an diese Korrelationsmatrix ein Modell mit 3 Trait-Faktoren und 3 Methoden-Faktoren an. Dabei ließ er Korrelationen zwischen den Trait-Faktoren untereinander und zwischen den Methoden-Faktoren untereinander, nicht aber zwischen Trait- und Methoden-Faktoren zu. Das Ergebnis ist in Tabelle 4.4 angegeben.

Dieses Modell erwies sich als den Daten gut angepaßt (der Chi-Quadrat-Test für die Signifikanz der Abweichungen vom Modell ergab einen Chi-Quadrat-Wert von 10.74 bei 12 Freiheitsgraden und war nicht signifikant).

Unter inhaltlichen Gesichtspunkten fallen vor allem die negativen Korrelationen des Trait-Faktors "Academic Self-Concept" zu den Trait-Faktoren "Physical" und "Social Self-Concept" auf. Sie legen eine Interpretation im Sinne kornpensatorischer Bestrebungen nahe und sind umso bemerkenswerter, als bei jedem der drei Fragebögen die Korrelationen zwischen den drei Aspekten des Selbstbildes positiv sind. Schwarzer (1983, S.226) bemerkt dazu: "Only the structural equation approach reveals that true interrelationships between underlying sources of covariation".

Auf der Suche nach einer Alternativerklärung betrachten wir nochmals die Korrelationen in Tabelle 4.3. Es fällt auf, daß sich die Korrelationsmuster der drei Traits in den verschiedenen Fragebogenarten nahezu mustergültig wiederholen, daß dabei nur Skala Al aus dem Rahmen fällt, indem sie generell zu hoch korreliert. Deshalb wurde eine Modell-Variante konzipiert, in der Al keine reine Academic Self-Concept-Skala ist, sondern die anderen beiden Aspekte mitenthält. Es wurden deshalb Nebenladungen der Variablen Al auch auf den Trait-Faktoren 2 und 3 zugelassen. Als Ausgleich für die zusätzlichen 2 Parameter wurden die zuvor negativen Korrelationen des Trait-Faktors "Academic Self-Concept" zu den beiden anderen Trait-Faktoren auf Null fixiert. Das Ergebnis ist in Tabelle 4.5 angegeben.

Tabelle 4.4: Schwarzers (1983) Multitrait-Multimethod Lösung für die Daten aus Tabelle 4.3

Matrix der Faktorladungen

		Trait-Faktoren			Methoden-Faktoren		
		1	2	3	A	B	C
Frage-	A1	.40	0	0	.82	0	0
bogen-	A2	0	.74	0	.61	0	0
Skalen	A3	0	0	.38	.72	0	0
	B1	.63	0	0	0	.48	0
	B2	0	.70	0	0	.68	0
	B3	0	0	.48	0	.58	0
	C1	.72	0	0	0	0	.58
	C2	0	.63	0	0	0	.64
	C3	0	0	.53	0	0	.64

Korrelationen der Faktoren

	1	2	3	A	B	C
1	1					
2	-.47	1				
3	-.75	.40	1			
A	0	0	0	1		
B	0	0	0	.58	1	
C	0	0	0	.70	.57	1

Tabelle 4.5: Ergebnis einer Reanalyse der Korrelationen aus Tabelle 4.3

Matrix der Faktorladungen

		Trait-Faktoren			Methoden-Faktoren		
		1	2	3	A	B	C
Frage-	A1	.71	.18	.21	41	0	0
bogen-	A2	0	.92	0	27	0	0
Skalen	A3	0	0	.63	71	0	0
	B1	.73	0	0	0	.37	0
	B2	0	.84	0	0	.50	0
	B3	0	0	.61	0	.44	0
	C1	83	0	0	***0***	0	.41
	C2	0	77	0	0	0	.55
	C3	0	0	74	0	0	.38

Korrelationen der Faktoren

	1	2	3	A	B	C
1	1					
2	0	1				
3	0	.63	1			
A	0	0	0	1		
B	0	0	0	.03	1	
C	0	0	0	.24	.31	1

Auch dieses Modell ist gut an die Daten angepaßt. Es weist bei ebenfalls 12 Freiheitsgraden sogar noch einen etwas kleineren Chi-Quadrat-Wert aus.

Gegenüber Schwarzers Lösung fallen die Ladungen in den Trait-Faktoren höher, in den Methodenfaktoren niedriger aus und führen damit zu einem insgesamt günstigeren Urteil über die Validität der Fragebogen. Für die inhaltliche Interpretation wesentlich ist der Wegfall der negativen Korrelationen bei den Trait-Faktoren, so daß kein Anlaß zur Annahme kornpensatorischer Mechanismen (etwa entsprechend dem Klischee von den dummen Schönen und häßlichen Intellektuellen) besteht.

Mit diesem Beispiel sollte deutlich gemacht werden, daß auch mit einer konfirmatorischen Faktorenanalyse die Modellgeltung nicht bewiesen werden kann, sondern Interpretationsmöglichkeiten aufgezeigt werden. Bei hoch restriktiven Modellen wird es allerdings sehr schwer sein, gleichwertige Alternativen zu finden und damit die Interpretation in Frage zu stellen. Damit ist die Beweiskraft einer konfirmatorischen Faktorenanalyse zwar auch begrenzt, aber doch wesentlich besser als die einer klassischen Faktorenanalyse, wo Alternativlösungen routinemäßig hergestellt werden können.

Zusammenfassung

Im klassischen faktorenanalytischen Modell mit mehreren gemeinsamen Faktoren gehen individuelle Unterschiede in den Testwerten auf individuelle Unterschiede in mehreren latenten Dimensionen (= Faktoren, z.B. Fähigkeiten) zurück. Der Testwert wird als gewichtete Summe der gemeinsamen Faktoren plus einem für den jeweiligen Test spezifischen Anteil gedacht. Aufgrund der Korrelationen zwischen den Tests als Ausgangsdaten sollen die gemeinsamen Faktoren und ihr relatives Gewicht für die einzelnen Tests (= die Ladungen) bestimmt werden.

Neben der mathematischen Uneindeutigkeit der Lösung (Rotationsproblem, Kommunalitäten-Schätzproblem) haben vor allem eine Reihe weiterer Kritikpunkte, die Ende der Sechzigerjahre vorgetragen wurden (Unüberprüfbarkeit des theoretischen Ansatzes, Populationsabhängigkeit der Ergebnisse, Artefakte durch Selektion und simultane Überlagerung), dazu geführt, daß der Anspruch, Ergebnisse von Faktorenanalysen könnten als funktional erklärende Theorien über das Zustandekommen der Testwerte interpretiert werden, aufgegeben wurde. Unter Zurücknahme des ursprünglichen Anspruchs, wird die Faktorenanalyse nunmehr als Daten explorierendes, Hypothesen generierendes Verfahren eingesetzt, oder als Methode zur Definition von Beschreibungsdimensionen, oder als bloßes Datenreduktionsverfahren.

Die konfirmatorische Faktorenanalyse unterscheidet sich von der klassischen dadurch, daß der Forscher bereits Hypothesen über die Zahl der Faktoren, das Ladungsmuster, die Korrelationen der Faktoren usw. haben muß. Wenn die Hypothesen restriktiv genug sind, kann ihre Vereinbarkeit mit der empirischen Korrelations- oder Kovarianzmatrix geprüft werden. Dazu wurden vier Beispiele aus dem Bereich der Testtheorie dargestellt. Wie an Beispiel 4 gezeigt, schließt aber auch ein gut angepaßtes, hoch restriktives Modell nicht aus, daß für dieselben Korrelationen plausible Alternativerklärungen gefunden werden.

Einführende Literatur:

Bortz, J. (1989). *Statistik für Sozialwissenschaftler*. (3. Aufl.). Kapitel 15: Faktorenanalyse. Berlin: Springer.

Weiterführende Literatur:

Pawlik, K. (1971). *Dimensionen* des *Verhaltens. (2.* Aufl.). Bern: Huber.
Revenstorf, D. (1980). *Faktorenanalyse.* Stuttgart: Kohlhammer.
McDonald, R.P. (1985). *Factoranalysis and related methods.* Hillsdale: Erlbaum Ass.
Bernstein, I.H. (1987). *Applied multivariate analysis.* Chapter 7: Confirmatory factor analysis (pp. 198-245). New York: Springer.

Ein inhaltliches Beispiel, bei dem die einzelnen Schritte bei der Planung einer Faktorenanalyse detailliert dargestellt sind, findet man bei:

Rost, D.H.(1987). Leseverständnis oder Leseverständnisse? *Zeitschrift für Pädagogische Psychologie,* 1, 175-196.

4.3 Einsatzmöglichkeiten und Grenzen der Clusteranalyse

1. Wozu dienen Clusteranalysen?
2. Welche Ausgangsdaten werden benötigt ?
3. Wie können mit Hilfe von Clusteranalysen Klassifikationen erstellt werden?

Vorstrukturierende Lesehilfe

Ziel von Clusteranalysen ist es, eine Klassifikation von Objekten zu erstellen, wobei Objekte, die in dieselbe Klasse eingeordnet werden, einander möglichst ähnlich, die Klassen untereinander aber möglichst unähnlich sein sollen. Solche Aufgabenstellungen kommen in verschiedensten Wissenschaftsbereichen vor (u.a. Psychologie, Biologie, aber auch z.B. Bibliothekswissenschaften), woraus sich eine Vielfalt sich überschneidender Ansätze und Verfahren entwickelt hat, die sich ihrerseits nicht leicht in Klassen ordnen läßt. Im folgenden wird weder ein vollständiger Überblick angestrebt, noch werden einzelne Verfahren im Detail dargestellt. Es sollen lediglich die Grundgedanken skizziert und Hinweise auf mögliche Anwendungen gegeben werden. Dabei kommen im Zusammenhang mit psychologisch diagnostischen Fragestellungen vor allem zwei Anwendungsbereiche in Betracht:

(a) die Clusteranalyse von Personen als "Objekten", mit dem Ziel, möglichst homogene Personengruppen zu bilden (z.B. um in der Folge zu untersuchen, ob sich diese Gruppen in ihrer Reaktion auf eine Behandlung unterscheiden) und

(b) die Clusteranalyse von Testaufgaben, um homogene Aufgabengruppen zu finden, aus denen sich Testskalen entwickeln lassen.

Ausgangspunkt der Clusteranalyse sind Ähnlichkeitsmaße. Will man z.B. Personen zu Clustern zusammenfassen, so hat man zunächst die Ähnlichkeit (oder Unähnlichkeit, Distanz) von jeder Person zu jeder anderen festzustellen. Dazu kommen direkte Ähnlichkeitsbeurteilungen in Betracht (so könnte z.B. der Lehrer die Ähnlichkeit jedes Schülers zu jedem anderen auf einer Punkteskala beurteilen) oder auch Ähnlichkeitswerte, die aufgrund von Merkmalsausprägungen errechnet werden. Sollen z.B. Schüler nach Ähnlichkeit ihrer Interessen gruppiert werden, so könnte das Ausgangsmaterial ein standardisierter Interessentest mit zehn Unterskalen für zehn verschiedene Interessenrichtungen sein. Die Unähnlichkeit zwischen zwei Schülern könnte dann z.B. als quadrierte *euklidische Distanz* bestimmt werden: Auf jeder Interessenskala wird die Differenz bestimmt, quadriert und über alle Skalen aufaddiert. Euklidische Distanzen haben zwar den Vorteil einer anschaulichen geometrischen Bedeutung, doch kommen andere Distanzmaße oft ebensogut in Betracht: Wenn man z.B. die Differenzen nicht quadriert, sondern einfach dem Betrag nach aufaddiert, so entspricht das dem sog. *City-block-Abstand.* Bezüglich weiterer Distanzmaße und der Definition von Ähnlichkeitsmaßen aufgrund von nur rangskalierten oder nominalskalierten Merkmalen sei auf die am Ende des Kapitels genannten Lehrbücher verwiesen.

Ist die Ähnlichkeit (bzw. Distanz) von jeder Person zu jeder anderen, allgemeiner von jedem Objekt zu jedem anderen, bestimmt, so soll als Nächstes die bestmögliche

Gruppenaufteilung gefunden werden. Dazu stehen verschiedene Verfahren zur Verfügung: Bei hierarchisch agglutinierenden Clusterverfahren werden, ausgehend von der maximalen Anzahl von Clustern (d.h. jede Person wird als Cluster der Größe Eins aufgefaßt), Cluster schrittweise zusammengefaßt. Es werden zunächst die beiden Personen, die zueinander den geringsten Abstand haben, zu einem Cluster der Größe Zwei zusammengefaßt, dann wird erneut gesucht, welche beiden Cluster zueinander den geringsten Abstand haben und diese beiden zusammengefaßt, bis schließlich nur noch zwei Cluster vorhanden sind, die im letzten Schritt in eines zusammengefaßt werden. Bei jedem Schritt dieser Prozedur muß das Distanzkriterium (zulässige Distanz zwischen zwei Clustern, die zusammengefaßt werden sollen) ein Stück gelokkert werden, und man bricht die Prozedur ab (bzw. entscheidet sich im nachhinein für diese Aufteilung), wenn eine weitere Zusammenfassung einen besonders großen Schritt in der Lockerung des Distanzkriteriums erfordern würde.

Dieser Grundgedanke hierarchisch agglutinierender Clusterverfahren ist in einer Vielfalt von Algorithmen realisiert, die sich u.a. darin unterscheiden, wie der Abstand zwischen Clustern gemessen wird. Zunächst ist ja nur der Abstand zwischen Einzelobjekten, z.B. Einzelpersonen, definiert. Der Abstand zwischen zwei Clustern kann z.B. definiert werden

(a) als der kleinste Abstand zwischen einer Person aus Cluster A und einer Person aus Cluster B (Single linkage), oder

(b) als der größte Abstand zwischen zwei Personen aus A und B (complete linkage), oder auch

(c) als der mittlere Abstand (arithmetisches Mittel oder Median) aller Abstände zwischen Personen aus A und B.

Diesen Definitionen ist gemeinsam, daß sie alle aus den Abständen zwischen den Einzelobjekten (hier: Personen) errechnet werden und nicht auf die Merkmalsausprägungen zurückgreifen. Sie sind deshalb auch dann anwendbar, wenn die Ausgangsdaten beispielsweise globale Ähnlichkeitsurteile über Personen sind oder, wie bei der Clusteranalyse von Items, Korrelationen als Ähnlichkeitsmaße verwendet werden.

Wenn die Ähnlichkeit zwischen den Einzelobjekten aus Merkmalsausprägungen errechnet wurde, z.B. aufgrund von Testwerten als euklidische Distanz, so liegt es nahe, ein Cluster durch die durchschnittliche Merkmalsausprägung der darin enthaltenen Objekte zu kennzeichnen (das *Zentroid*) und die Abstände zwischen Clustern als Abstand zwischen den Zentroiden zu bestimmen. Die Heterogenität innerhalb eines Clusters kann auch als Merkmalsvarianz (Summe der Varianzen der einzelnen Merkmale oder multivariate Varianzmaße) definiert werden, und als Kriterium einer guten Clusterlösung kann definiert werden, daß die Varianz innerhalb der Cluster im Vergleich zur Varianz zwischen den Clustern (errechnet aus den Abständen zwischen den Clustermittelwerten) möglichst gering sein soll.

Neben den hierarchisch agglutinierenden Algorithmen kommen auch nicht hierarchische Verfahren zum Einsatz. Dabei wird die Clusterzahl als bekannt vorausgesetzt und, ausgehend von einer groben Näherungslösung, jedes Element probeweise in ein anderes Cluster verschoben, um zu sehen, ob sich eine Verbesserung der Clusterlösung im Sinne eines der oben genannten Kriterien ergibt. Durch das Verschieben einzelner Elemente ergibt sich eine Neudefinition der Cluster, die Abstände werden neu berechnet und es wird mit dem Verschieben fortgefahren, bis sich keine weitere Verbesserung mehr ergibt. Vielfach werden auch beide Typen von Algorithmen miteinander verbunden, indem zunächst mit hierarchisch agglutinierenden Verfahren eine

Ausgangslösung gesucht und die Clusterzahl festgesetzt wird und danach mit nicht hierarchischen Verfahren noch nach Verbesserungsmöglichkeiten gesucht wird.

Für Forschungsvorhaben im Bereich der pädagogisch-psychologischen Diagnostik kommen, wie schon eingangs erwähnt, vor allem zwei Einsatzbereiche für Clusteranalysen in Betracht: die Clusteranalyse von Personen mit dem Ziel, homogene Personengruppen zu definieren, z.B. um sie als Kriteriumsgruppen bei einer Testvalidierung zu verwenden. Mittels Clusteranalyse gefundene Kategorisierungen könnten aber auch als unabhängige Variable in Versuchsplänen herausgezogen werden, bei denen es darum geht, Behandlungseffekte weiter zu analysieren. Als zweiter Einsatzbereich ist die Clusteranalyse von Items zu sehen, mit dem Ziel aus einer großen Menge von Aufgaben Untergruppen zu bilden, aus denen sich möglichst unabhängige Skalen bilden lassen.

Wenn die Clusteranalyse, verglichen mit anderen multivariaten Verfahren, seltener zum Einsatz kommt, so dürften dafür folgende Gründe verantwortlich sein:

(a) Die Durchführung einer Clusteranalyse erfordert sowohl bei der Auswahl des Ähnlichkeitsmaßes als auch bei der Auswahl der Algorithmen und der Festlegung der Clusterzahl eine Reihe von Entscheidungen, die inhaltlich oft schwer zu begründen sind.

(b) Bei der Clusteranalyse von Personen ist immer zu bedenken, daß die untersuchten Personen nur eine Stichprobe aus der Population sind, über die Aussagen gemacht werden soll. Über den Einfluß von Stichprobenfehlern auf Clusterlösungen ist aber bislang nur wenig bekannt.

(c) Die Clusteranalyse von Items dient im wesentlichen denselben Zielen wie die Faktoranalyse. Neben der starken Tradition der Faktoranalyse ergab sich von daher kein besonderer Bedarf nach Clusterverfahren als Alternative - zumal in beiden Fällen die Korrelationen die Ausgangsbasis bilden und somit alle Einwände gegen die Faktorenanalyse, die die Populationsabhängigkeit und mögliche Artefakte bei der Berechnung von Korrelationen betreffen, für die Clusteranalyse genauso zutreffen.

Zusammenfassung

Clusteranalysen haben zum Ziel, Objekte so zu gruppieren, daß Objekte, die in dieselbe Gruppe (= Cluster) fallen, möglichst ähnlich, die Gruppen untereinander möglichst unähnlich sind. Es steht eine Vielzahl von Verfahren zur Verfügung, die sich danach unterscheiden, wie Ähnlichkeit bestimmt wird und nach welchen Algorithmen die Gruppenzusammenfassung erfolgt. In der psychologisch diagnostischen Forschung können Clusteranalysen zur Gruppierung von Personen oder auch im Rahmen der Testkonstruktion zur Gruppierung von Aufgaben zum Einsatz kommen.

Einführende Literatur:

Bortz, J. (1989). *Statistikfür Sozialwissenschaftler* (3. Aufl.). Kapitel 16: Clusteranalyse. Heidelberg: Springer.

Weiterführende Literatur:

Eckes, T. & Roßbach, H. (1980). Clusteranalysen. Stuttgart: Kohlhammer.
Krauth, J. (1983). Typenanalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S. 440-496). Göttingen: Hogrefe.
Oldenbürger, H.A. (1983). Clusteranalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S.390-439). Göttingen: Hogrefe.
Steinhausen, D. & Langer, K. (1977). *Clusteranalyse.* Berlin: de Gruyter.